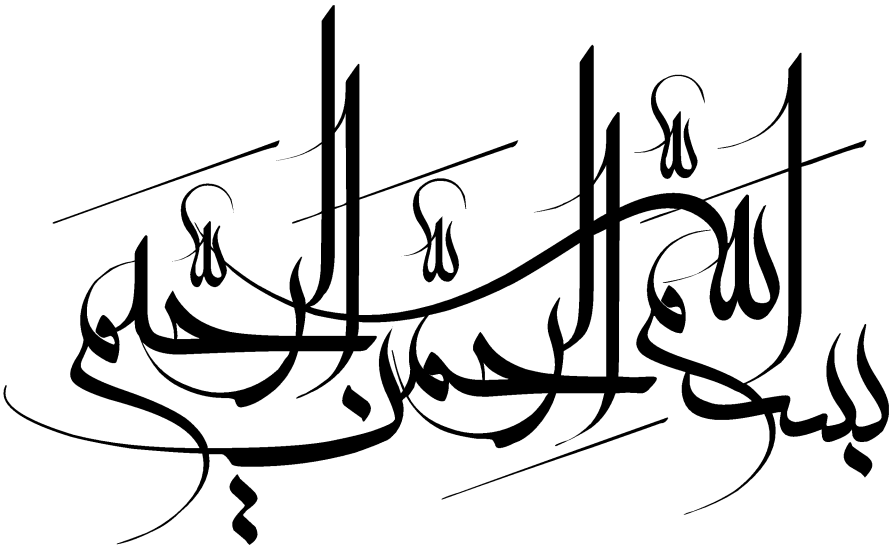

بسم الله الرحمن الرحيم

دانشکده‌ی برق و رباتیک

گروه الکترونیک – گرایش دیجیتال

آنالیز موضوعی و تخصیص مؤلف متون فارسی و عربی با استفاده از اطلاعات ساختاری اجزای متن

علی شهنما

اساتید راهنما:

دکتر علیرضا احمدی‌فرد

دکتر حسین مروی

استاد مشاور:

دکتر مرتضی زاهدی

پایان‌نامه جهت اخذ درجه‌ی کارشناسی ارشد

بهمن ۱۳۹۳

شماره : ۱۲۵۵/آ.ت.ب
تاریخ : ۹۳/۱۱/۲۸
ویرایش : ------

بسمه تعالی

دانشگاه صنعتی شاهرود
مدیریت تحصیلات تکمیلی
فرم شماره (۶)

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای :

علی شهنما رشته : **بـــرق - الکترونیک** گرایش : **دیجیتال**

تحت عنوان : **آنالیز موضوعی و تخصیص مؤلف متون فارسی – عربی، با استفاده از اطلاعات ساختاری اجزای متن**

که در تاریخ **۹۳/۱۱/۲۸** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : عالی) امتیاز (۱۹٫۴۴) ☑	دفاع مجدد ☐	مردود ☐	

۱ـ عالی (۲۰ ـ ۱۹) ۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ ۱۶) ۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضـــاء
۱ـ استـاد راهنمـا	علیرضا احمدی فرد حسین خسروی	دانشیار استادیار	
۲ـ استـاد مشـاور	—	—	—
۳ـ نماینده شورای تحصیلات تکمیلی	ساسان ناصح	استادیار	
۴ـ استـاد ممتحـن	سید محمد شهرام امیری	دانشیار	
۵ ـ استـاد ممتحـن	حسین حسنی	استادیار	

رئیس دانشکده

تشکر و قدردانی

نخست،

حمد و سپاس خداوند رحمان و رحیم را

که تمامی موفقیت‌های این پایان‌نامه در سایه‌ی لطف او رقم خورده است.

دوم،

سپاس و قدردانی ویژه از استاد عزیزم جناب آقای **دکتر علیرضا احمدی‌فرد**

که دلسوزانه مرا در مراحل مختلف انجام این پایان‌نامه یاری کردند و

از راهنمایی‌ها و تجربیاتشان بهره‌های فراوان بردم.

سوم،

تشکر و سپاس از جناب آقای **دکتر حسین مروی**،

که از محضر ایشان نیز بهره‌ها جستم.

و در پایان،

سپاس و قدردانی ویژه از **پدر و مادر مهربانم**،

که در تمامی مراحل زندگی یار و یاور من بوده‌اند.

تقدیم به

پدر و مادر مهربانم

تعهدنامه

اینجانب **علی شـــهنما** دانشـجوی کارشـناسـی ارشـد رشـته **الکترونیک دیجیتال** دانشـکده **برق و رباتیک دانشگاه شاهرود**، نویسنده پایان‌نامه با عنوان:

«آنالیز موضوعی و تخصیص مؤلف متون فارسی و عربی،
با استفاده از اطلاعات ساختاری اجزای متن»،

تحت راهنمایی **دکتر علیرضا احمدی‌فرد** و **دکتر حسین مروی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه تا کنون توســـط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- حقوق معنوی این اثر، به دانشـگاه شـاهرود تعلق دارد و مقالات مسـتخرج با نام «دانشـگاه شاهرود» یا «Shahrood University» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دســـت آوردن نتایج اصـــلی پایان‌نامه تأثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی دسترسی یافته (یا استفاده) شده است، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

علی شهنما

مالکیت نتایج و حق نشر

- تمـام حقوق معنوی این اثر و محصـــولات آن (مقالات مســـتخرج، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

تخصیص نویسنده یکی از زیرشاخه‌های پردازش متن می‌باشد که هدف اصلی آن تعیین هویت نویسنده‌ی یک متن است. به عبارت دیگر هدف اصلی این حوزه، طراحی سیستمی است که بتواند هویت نویسنده‌ی یک متن را از میان چند نویسنده‌ی نامزد تعیین نماید. به منظور طراحی چنین سیستمی می‌بایست تعدادی متن از هر نامزد در اختیار داشته باشیم.

تمامی پژوهش‌های گذشته در حوزه‌ی تخصیص نویسنده‌ی متون فارسی به روش‌های مبتنی بر سیستم‌های پردازش زبان‌های طبیعی (NLP) منحصر می‌شوند، اما هدف اصلی این پایان‌نامه بررسی عملکرد روش‌هایی موسوم به NDP بر روی مسائل تخصیص نویسنده‌ی زبان فارسی است. این روش‌ها بر مبنای تعداد تکرار انگرام‌ها طراحی شده و کاملاً مستقل از سیستم‌های NLP می‌باشند.

در این پایان‌نامه مهم‌ترین روش‌های NDP موجود مطالعه شده و سپس با الهام از آن‌ها، دو روش جدید پیشنهاد شده است. در روش پیشنهادی اول (CNG-WIS) به جای استفاده از فرکانس انگرام‌ها، اندیس آن‌ها جهت حل مسائل به کار رفته است. در دومین روش پیشنهادی (VNG) به جای آنکه انگرام‌های پرتکرار مبنای کار قرار گیرند، از انگرام‌های پراکنده استفاده شده است.

به منظور ارزیابی روش‌های موجود و هم‌چنین مقایسه‌ی روش‌های پیشنهادی با آن‌ها، از چهار مجموعه‌متن (یا پایگاه‌داده) مختلف از دو زبان فارسی و عربی استفاده شده است. یکی از این مجموعه‌متن‌ها (به نام CPPT) توسط نگارنده گردآوری شده و دارای ۱۴۵متن از ۶ نویسنده‌ی معاصر فارسی‌زبان می‌باشد. نتایج بدست‌آمده حاکی از آنست که علاوه بر روش‌های NDP موجود، روش‌های پیشنهادی نیز قدرت بالایی در حل مسائل تخصیص نویسنده‌ی زبان‌های فارسی و عربی دارند.

در انتها، دو مسئله‌ی خاصِ حوزه‌ی ادبیات فارسی بررسی شده‌اند: نظیره‌های گلستان و غزلیات سبک هندی. بدین‌منظور دو مجموعه‌متن دیگر با نام‌های GBP (شامل ۷۵ حکایت از سه نویسنده) و SBH (شامل ۹۰ غزل به سبک هندی از سه شاعر) توسط نگارنده جمع‌آوری شده است. نتایج بدست‌آمده نشان می‌دهد که روش‌های NDP علاوه بر مسائل تخصیص نویسنده‌ی متون نثر فارسی، در حکایات (ترکیبی از نثر و نظم) و اشعار نیز قدرت بالایی دارند.

کلمات کلیدی: تخصیص نویسنده، انگرام، روش‌های مبتنی بر پروفایل، نشانگر سبک، مجموعه‌متن CPPT، مجموعه‌متن GBP و مجموعه‌متن SBH.

مقاله‌ی استخراج‌شده از پایان‌نامه

ع. شهنما و ع. احمدی‌فرد، *«تخصیص نویسنده‌ی مبتنی بر انگرام در داستان‌های کوتاه زبان فارسی»*، ششمین کنفرانس فناوری اطلاعات و دانش، دانشگاه شاهرود، شاهرود، ایران، ۱۳۹۳.

فهرست مطالب

فهرست شکل‌ها:

فهرست جداول:

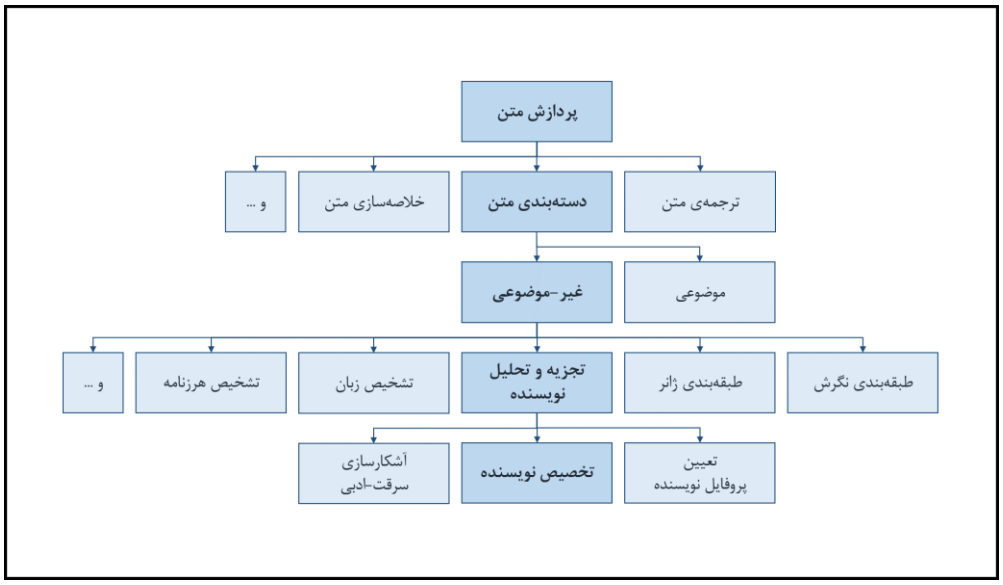
فصل اول

معرفی و
شرح صورت‌مسئله

۱-۱ مقدمه

این پایان‌نامه بر روی یکی از حوزه‌های چالش‌برانگیز پردازش متن، یعنی «تخصیص نویسنده[1]» متمرکز شده است. هدف اصلی این حوزه تعیین نویسنده‌ی یک متن بدون پرچسب (یعنی متنی که نویسنده‌ی آن نامشخص است) بر اساس تعدادی متون برچسب گذاری شده (متونی با نویسنده‌ی مشخص) می‌باشد.

به منظور آشنایی بهتر با حوزه‌ی تخصیص نویسنده، بهترست ابتدا نگاهی گذرا به مسئله‌ی «پردازش متن[2]» و مباحث زیرمجموعه‌ی آن بیندازیم. بدین منظور ابتدا به بررسی اجمالی حوزه‌های مختلف زیرمجموعه‌ی پردازش متن پرداخته (بخش ۱-۲)، سپس شاخه‌ی تخصیص نویسنده را به صورت مجزا شرح می‌دهیم (بخش ۱-۳). زیرشاخه‌های مختلف حوزه‌ی پردازش متن به صورت نمودار درختی در شکل ۱-۱ نمایش داده شده اند.



شکل ۱-۱: زیرشاخه‌های پردازش متن

[1] Authorship Attribution (AATT)
[2] Text Processing

۲-۱ پردازش متن و زیرشاخه‌ها

هرچند تا کنون تعاریفی گوناگون و پیچیده از حوزه‌ی پردازش متن ارائه شده است، اما می‌توان عملکرد این حوزه را در یک جمله خلاصه نمود: «استخراج هرگونه اطلاعات از روی متن الکترونیکی». اصطلاح پردازش متن خود از دو واژه تشکیل شده است:

- **پردازش:** هر نوع فرایندی که بتوان توسط آن اطلاعاتی را از سوژه‌ی مورد-پردازش استخراج نمود.
- **متن:** یک دنباله‌ی بهم پیوسته از نویسه‌های مختلف (حروف و نمادها) می‌باشد. به عبارت دیگر منظور از متن در این حوزه، هر فایل متنی[۱] ذخیره‌شده در رایانه می‌باشد.

پس از بررسی تعریف این حوزه‌ی پردازشی، حال سراغ شاخه‌های آن رفته و آن‌ها را به صورت اجمالی بررسی می‌نماییم. پردازش متن دارای سه شاخه‌ی اصلی می‌باشد:

- **ترجمه‌ی خودکار متن[۲]:** در این مورد یک متن به طور خودکار از زبان مبدأ به زبان مقصد ترجمه می‌شود. امروزه سرویس‌های ترجمه‌ی برخط (همچون سرویس ترجمه‌ی مایکروسافت[۳] و سرویس ترجمه‌ی گوگل[۴]) این نوع ترجمه را به ازای طیف وسیعی از زبان‌های مبدأ و مقصد (شامل زبان فارسی) انجام می‌دهند.
- **خلاصه‌سازی خودکار متن[۵]:** در این شاخه یک متن طولانی به طور خودکار به متنی کوتاه‌تر خلاصه می‌شود؛ به طوریکه نکات اصلی متن طولانی در متن کوتاه‌تر نیز بیان شده باشند.
- **دسته‌بندی متن[۶]:** در این حالت متون به صورت خودکار به دسته‌های مختلفی تقسیم‌بندی می‌شوند، که این تقسیم‌بندی مبناهای متفاوتی (از جمله موضوع، نویسنده و ...) دارد.

از آنجا که هدف اصلی پایان‌نامه بررسی حوزه‌ی تخصیص نویسنده بوده و این حوزه در قالب یک مسئله‌ی دسته‌بندی متن ارائه می‌شود، لذا از میان سه شاخه‌ی اصلی پردازش متن، بر روی شاخه‌ی دسته‌بندی متن متمرکز شده و آن را بیش‌تر شرح می‌دهیم.

---

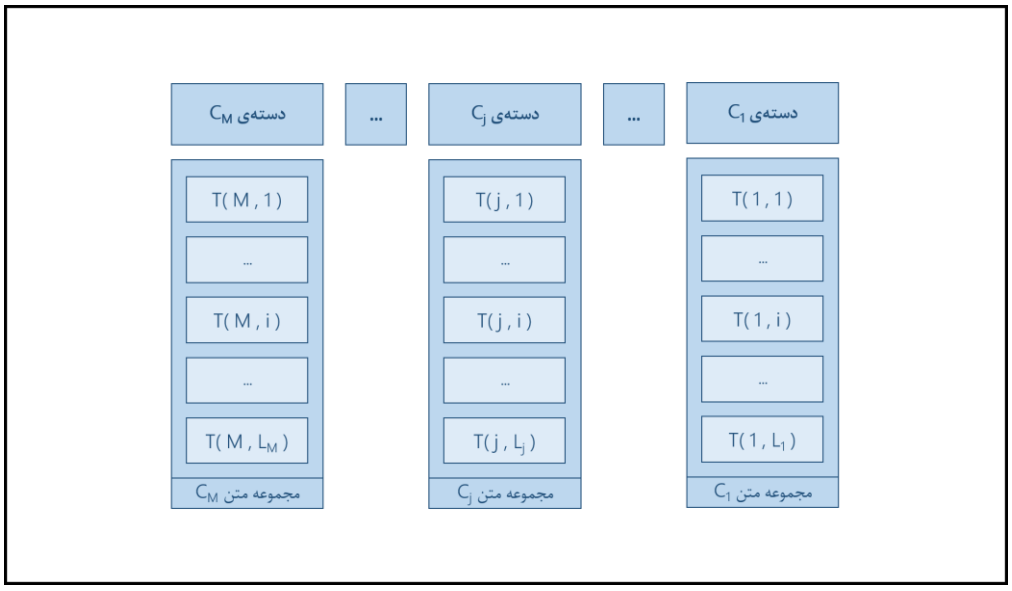
[1] Text-File (*.txt)
[2] Automatic Text Translation
[3] Bing Translator: http://www.bing.com/translator/
[4] Google-Translate: https://translate.google.com/
[5] Automatic Text Abstracting
[6] Text Classification (TC)

۱-۲-۱ دسته‌بندی متن

در مقالات و کتب مربوط به حوزه‌ی دسته‌بندی متون، تعاریف متعدد و بعضاً متفاوتی از این حوزه بیان شده است. این تعاریف به دلیل استفاده از عبارات ریاضی، شاید در ابتدا گمراه‌کننده باشند؛ لذا بهترست در ابتدا تعریفی ساده و روان از این حوزه ارائه داده و سپس تعاریف علمی‌تر را ارائه نماییم.

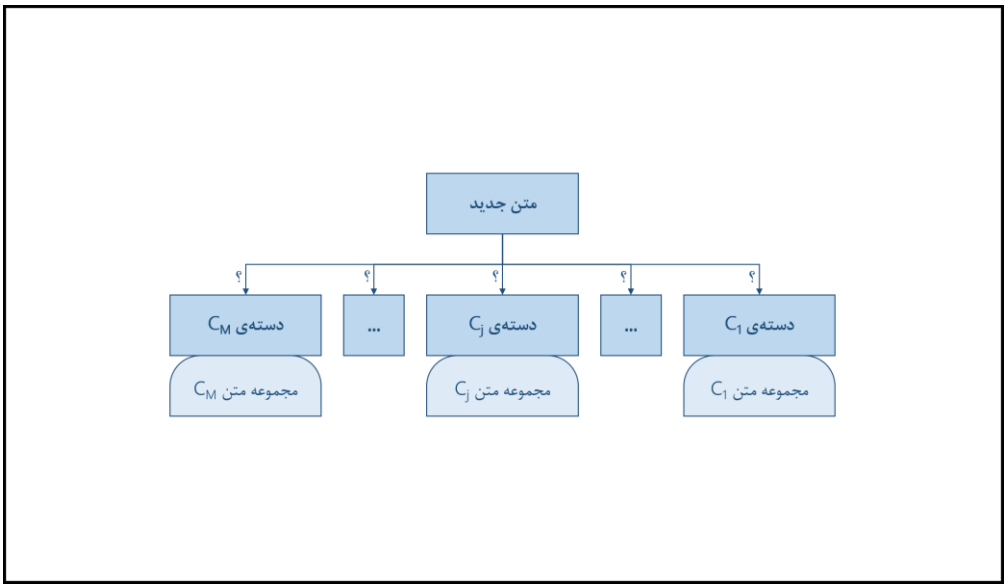
فرض کنیم تعدادی متن دسته‌بندی شده (یا برچسب گذاری شده) در اختیار داریم. در ادبیات دسته‌بندی متن، چنین مجموعه‌ای را اصطلاحاً «مجموعه‌متن[1]» می‌نامند. یک نمای کلی از مجموعه‌متن در شکل ۱-۲ نمایش داده شده است. در این شکل منظور از T(i,j) عبارتست از: متن نمونه‌ی i–ام از نویسنده‌ی j–ام.

حال یک متن جدید را در نظر بگیرید که در مجموعه متن وجود نداشته است. هدف از دسته‌بندی متن، تخصیص دادن این متن جدید به یک یا چند دسته از دسته‌های موجود در مجموعه‌متن می‌باشد (شکل ۱-۳). لازم بذکرست که این تخصیص دادن می‌بایست بر مبنای متن‌های موجود در مجموعه‌متن و دسته‌بندی آن‌ها انجام پذیرد.



شکل ۱-۲: نمای کلی از یک مجموعه‌متن مناسب برای مسئله‌ی دسته‌بندی متن

[1] Corpus or Corpora



شکل ۱-۳: نمای کلی از یک مسئله دسته‌بندی متن

با توجه به توضیحات فوق می‌توان عملکرد و هدف حوزه‌ی دسته‌بندی متن را خلاصه نموده و تعریف کوتاهی از آن ارائه داد: «دسته‌بندی متن عبارتست از تخصیص یک متن به یک یا چند دسته‌ی از پیش تعیین شده، بر مبنای مجموعه‌ای از متون دسته‌بندی شده.» البته تعریف دقیق‌تری از این حوزه در [۱] آمده است: «دسته‌بندی متن یک شیوه‌ی بانظارت است که از داده‌های آموزشی برچسب زده شده برای آموزش دادن سیستم طبقه‌بند استفاده نموده و سپس باقیمانده‌ی متون را توسط طبقه‌بند آموزش داده شده به صورت خودکار دسته‌بندی می‌نماید.»

همانگونه که بیان شد تخصیص متن جدید به یک یا چند دسته، با توجه مجموعه‌متن صورت می‌پذیرد. به عنوان مثال اگر متون موجود در مجموعه‌متن بر اساس موضوع به سه دسته‌ی دینی، سیاسی و اقتصادی تقسیم‌بندی شده باشند، متن جدید نیز توسط سیستم طبقه‌بند به یک یا چند دسته از موضوعات مذکور تخصیص داده می‌شود. حال سؤال این‌جاست که آیا مبناهای دیگری جز موضوع برای دسته‌بندی متون وجود دارد؟ پاسخ مثبت است و این مبناها اجمالاً در بخش ۱-۲-۲ بررسی می‌شوند.

۱-۲-۲ مبناهای رایج دسته‌بندی متن

همانگونه که در شکل ۱-۱ آمده است، حوزه‌ی دسته‌بندی متون بر اساس مبنای دسته‌بندی به دو گروه تقسیم می‌شود [۲]:

- **تقسیم‌بندی موضوعی**[1]: در این حالت متون مجموعه‌متن بر مبنای موضوع تقسیم‌بندی شده‌اند و لذا هدف از دسته‌بندی، تعیین موضوع کلی متن (به عنوان مثال دینی، سیاسی، اقتصادی و ...) می‌باشد. گروه‌بندی متن اخبار روزنامه‌ها یکی از رایج‌ترین نمونه‌های این زیرشاخه می‌باشد. در این مورد خبرها به دسته‌هایی از قبیل خبرهای سیاسی، اجتماعی، دینی، ورزشی، اقتصادی، هنری و ... تقسیم‌بندی می‌شوند. تحقیقات ارائه شده در [۳] و [۴] دو نمونه از پژوهش‌های صورت گرفته در این شاخه می‌باشند.
- **تقسیم‌بندی غیر-موضوعی**[2]: در زیرشاخه‌ی دوم، متون مجموعه‌متن بر مبنایی غیر از موضوع تقسیم‌بندی شده‌اند. هرچند پژوهش‌های آغازین دسته‌بندی متن بر روی تقسیم‌بندی موضوعی متمرکز بوده، اما در دو دهه‌ی اخیر فعالیت‌ها و پژوهش‌های زیاد انجام گرفته در حوزه‌ی تقسیم‌بندی غیرموضوعی، موجب پیشرفت و توسعه‌ی این حوزه گردیده است [۲]. مهم‌ترین و رایج‌ترین زیرشاخه‌های این حوزه عبارتند از:
  - **تجزیه و تحلیل نویسنده‌ی متن**[3]: هدف این زیرشاخه، دسته‌بندی متون بر اساس یکی از ویژگی‌های خاص نویسنده (مثلاً سن نویسنده) است. از آنجا که تخصیص نویسنده (که هدف اصلی این پایان‌نامه می‌باشد)، در این زیرشاخه طبقه‌بندی می‌شود، لذا این زیرشاخه در بخش ۱-۲-۳ به صورت جداگانه شرح داده خواهد شد.
  - **طبقه‌بندی ژانر متن**[4]: در این حالت، متون به ژانرهای مختلف تقسیم‌بندی می‌شوند. به عنوان مثال تقسیم‌بندی مطالب یک پایگاه اینترنتی به ژانرهایی همچون: مقالات تحقیقاتی[5]، مقالات مروری[6]، پرسش-پاسخ[7]. از [۴ تا ۸] می‌توان به عنوان نمونه‌ای از پژوهش‌های این حوزه نام برد.
  - **طبقه‌بندی نگرش متن**[8]: در این حالت نحوه‌ی نگرش ارائه‌شده در متن مورد توجه قرار می‌گیرد. به عنوان مثال نقدهای نوشته‌شده در مورد آثار سینمایی، به سه گروه زیر تقسیم‌بندی می‌شوند: نقد مثبت (نویسنده در کل نظر مثبتی به فیلم دارد)، نقد منفی

---

[1] Topic-Based TC
[2] Non-Topic-Based TC
[3] Author Analysis (AA)
[4] Genre Classification
[5] Research Articles
[6] Reviews
[7] Question and Answer (Q&A)
[8] Sentiment Classification

(نظر کلی نویسنده نسبت به فیلم منفی است)، نقد خنثی[1] (نویسنده نقاط ضعف و قوت فیلم را در یک سطح می‌داند). پژوهش‌های ارائه‌شده در [۹ تا ۱۲] در این دسته جای می‌گیرند.

- **شناسایی هرزنامه‌ی الکترونیکی[2]:** در این شاخه که به شدت مورد علاقه‌ی ارائه‌دهندگان سرویس پست الکترونیکی (همچون مایکروسافت، گوگل، یاهو، چاپار و ...) می‌باشد، هدف اصلی تشخیص و جداسازی هرزنامه از سایر نامه‌های الکترونیکی است. هرزنامه‌ها معمولاً در پوشه‌ای خاص و جدا از سایر پست‌های الکترونیکی ذخیره می‌شوند. پژوهش [۱۳] از جمله‌ی این پژوهش‌ها به شمار می‌آید.
- **تشخیص زبان متن[3]:** در این شاخه زبان متن مورد بررسی قرار گرفته و متون به دسته‌هایی همچون انگلیسی، فرانسوی، یونانی تقسیم‌بندی می‌شوند. یکی از کاربردهای این حوزه در سرویس‌های ترجمه‌ی برخط (همچون سرویس ترجمه‌ی مایکروسافت و سرویس ترجمه‌ی گوگل)، جهت تعیین خودکار زبان متن مبدأ (متنی که می‌بایست به زبان مقصد ترجمه شود) می‌باشد. یکی دیگر از مثال‌های رایج این شاخه، تعیین زبان یک برنامه‌ی کامپیوتری[4] می‌باشد. در این مورد متون (که در این حالت، همان برنامه‌های کامپیوتری می‌باشند) به دسته‌هایی همچون VB.NET، C++، C#، Java و ... تقسیم‌بندی می‌شوند. پژوهش گزارش شده در [۳] بر روی زبان طبیعی و پژوهش‌های صورت گرفته در [۱۴] بر روی زبان برنامه‌نویسی انجام پذیرفته است.

۳-۲-۱ تجزیه و تحلیل نویسنده‌ی متن

در این زیرشاخه، متون مجموعه‌ی آموزش بر مبنای ویژگی‌های مبتنی بر نویسنده تقسیم‌بندی می‌شود. این زیرشاخه خود به سه قسمت تقسیم می‌شود:

---

[1] Neutral
[2] Spam Identification
[3] Language Detection
[4] Source-Code Language Detection

- **تخصیص نویسنده:** در این حالت، متون بر اساس هویت نویسنده تقسیم‌بندی شده و هدف اصلی تعیین نویسنده‌ی متن جدید است. این حالت که در واقع حوزه‌ی اصلی مورد مطالعه در این پایان‌نامه است، مفصلاً در بخش ۱-۳ شرح داده خواهد شد.
- **تعیین پروفایل نویسنده**[1]: در این مورد، هدف تشخیص ویژگی خاصی از نویسنده‌ی متن می‌باشد. به عنوان مثال: متون مجموعه‌ی آموزش بر مبنای جنسیت نویسنده تقسیم‌بندی شده و هدف، تشخیص مرد یا زن بودن نویسنده‌ی متن جدید می‌باشد. رایج‌ترین ویژگی‌های مورد استفاده در این حوزه عبارتند از: سن نویسنده، جنسیت نویسنده و ویژگی‌های اخلاقی و رفتاری نویسنده. از جمله تحقیقات صورت گرفته در این حوزه می‌توان به [۱۵ تا ۱۷] اشاره نمود.
- **آشکارسازی سرقت ادبی**[2]: در این مورد، هدف تشخیص و آشکارسازی وقوع سرقت ادبی در یک متن (متن مشکوک[3]) می‌باشد. طبق [۱۸] این شاخه دارای دو زیرمجموعه است:
  - **بازیابی منبع**[4]: در این حالت هدف جستجو برای یافتن منابع احتمالی (که متن مشکوک از روی آنها نوشته شده)، می‌باشد.
  - **چیدمان متن**[5]: در این مورد، هدف تعیین و تطبیق قطعات دوباره-استفاده-شده[6] بین دو متن می‌باشد.

  از تحقیقات صورت پذیرفته در این حوزه می‌توان به پژوهش‌های [۱۹ تا ۲۱] اشاره نمود.

۱-۳ تخصیص نویسنده‌ی متن

پس از آشنایی اجمالی با پردازش متن و زیرشاخه‌های آن، حال به سراغ حوزه‌ی اصلی مورد مطالعه در این پایان‌نامه (یعنی تخصیص نویسنده‌ی متن) رفته و آن را شرح می‌دهیم.

در ابتدا تعریفی ساده و روان از این حوزه ارائه می‌دهیم: «تخصیص نویسنده عبارتست از مسئله‌ی شناسایی هویت نویسنده‌ی یک متن فاقد نویسنده‌ی مشخص یا متنی که در مورد نویسنده‌ی آن شک وجود دارد.

---

[1] Author Profiling
[2] Plagiarism Detection
[3] Suspicious Text
[4] Source Retrieval
[5] Text Alignment
[6] Passages of reused text

[۲]» تعریف دقیق‌تری از این حوزه در [۲۲] آمده است: «تخصیص نویسنده، تشخیص هویت نویسنده‌ی واقعی یک متن است، با در اختیار داشتن نمونه‌های از متون نوشته‌شده توسط مجموعه‌ای از نویسنده‌های کاندید.»

پس تخصیص نویسنده، همان دسته‌بندی متن است؛ زمانی‌که دسته‌بندی متون بر مبنای نویسنده‌ی آن‌ها انجام شده باشد و هر دسته بیانگر یک نویسنده‌ی خاص باشد. لازم بذکرست بر خلاف حوزه‌ی دسته‌بندی متن که در آن، متن جدید به یک یا چند دسته تخصیص داده می‌شد، غالباً در تخصیص نویسنده هر متن جدید فقط به یک دسته (معادل یک نویسنده) تخصیص داده می‌شود.

به منظور شرح بیش‌تر این حوزه، مراحل تعریف صورت مسئله را به صورت گام به گام بیان می‌کنیم:

۱) مجموعه‌ای از متون در اختیار داریم که نویسنده‌ی هر کدام از آن‌ها تعیین شده می‌باشد. به متنی که نویسنده‌ی آن مشخص است، اصطلاحاً متن برچسب‌خورده[۱] یا متن خارج از مناقشه[۲] گفته می‌شود.
۲) حال متن جدیدی را در نظر بگیریم که در مجموعه وجود ندارد و نویسنده‌ی آن برای ما مشخص نیست. چنین متنی را اصطلاحاً متن بدون برچسب[۳] یا متن مورد مناقشه[۴] می‌نامند.
۳) حال صورت مسئله‌ی تخصیص نویسنده را می‌توان این‌گونه تعریف نمود: چگونه بر مبنای مجموعه‌ی (۱)، نویسنده‌ی متن (۲) را تعیین نماییم؟

به منظور آسان شدن روند توضیح مسائل و جلوگیری از سردرگمی خواننده، تمامی تعاریف فوق را بدین‌صورت خلاصه می‌نماییم:

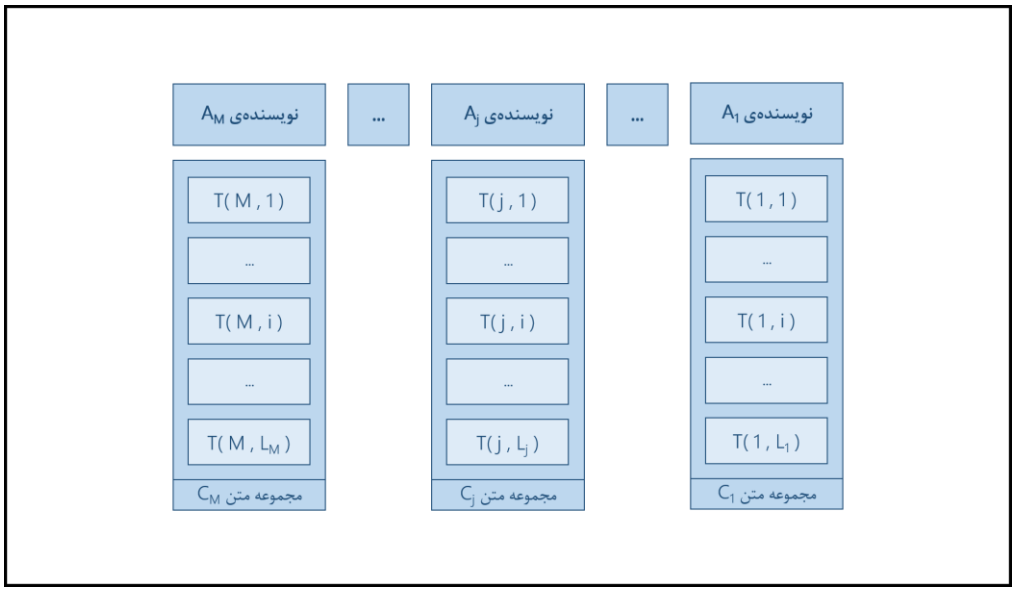
۱) مجموعه‌متن[۵]: مجموعه‌ای از متون دارای نویسنده‌ی مشخص (شکل ۱-۴)
۲) متن دیده‌نشده[۶]: متنی جدید و خارج از مجموعه‌متن، دارای نویسنده‌ی نامشخص (شکل ۱-۵)
۳) سیستم تخصیص نویسنده[۷] (AAS): سیستمی که بتواند بر مبنای متون موجود در مجموعه‌متن، نویسنده‌ی متن دیده‌نشده را تعیین نماید. (شکل ۱-۵)

---

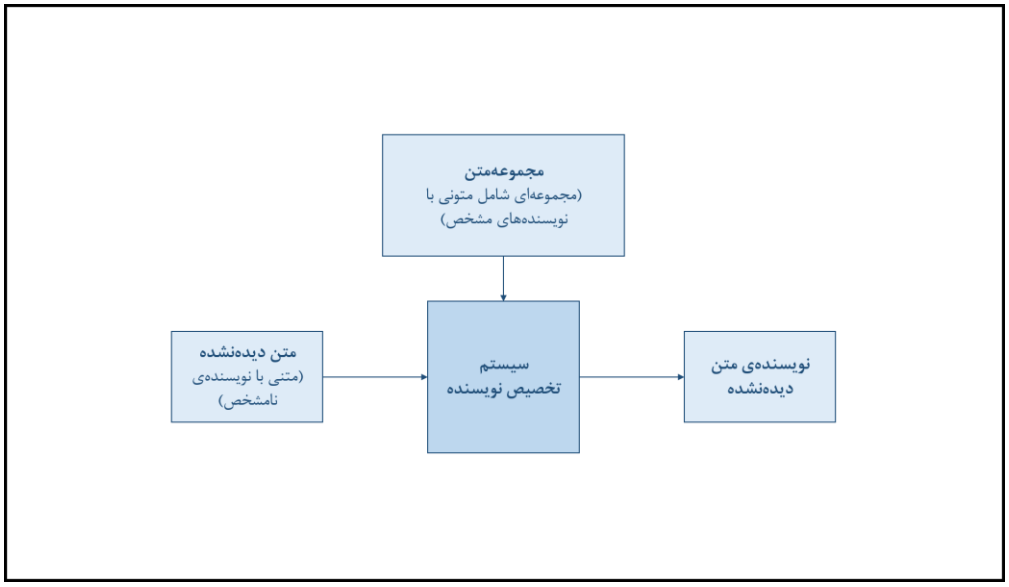
[1] Labeled Text
[2] Undisputed Text
[3] Unlabeled Text
[4] Disputed Text
[5] Corpus
[6] Unseen Text
[7] Author Attribution System (AAS)

بر مبنای سه تعریف فوق، صورت مسئله‌ی حوزه‌ی تخصیص نویسنده به دو صورت قابل بیان است:

- تشخیص نویسنده‌ی متن دیده‌نشده، بر مبنای متون موجود در مجموعه‌متن، یا
- طراحی سیستم AAS



شکل ۱-۴: نمای کلی از یک مجموعه‌متن در مسئله‌ی تخصیص نویسنده



شکل ۱-۵: نمای کلی از یک سیستم تخصیص نویسنده

اطلاعات تکمیلی در مورد حوزه‌ی تخصیص نویسنده در پیوست ۱ آمده است. در این پیوست دو قالب اصلی (مجموعه‌باز و مجموعه‌بسته) و دو مسئله‌ی بنیادین این حوزه بررسی شده است.

۱-۴ کابردهای تخصیص نویسنده

پس از معرفی حوزه‌ی تخصیص نویسنده و آشنایی کلی با صورت مسائل آن، در این بخش به بیان برخی از کاربردهای این حوزه می‌پردازیم. بر مبنای مقالات [۲۲ و ۲۵ تا ۲۹] این کاربردها را می‌توان به سه شاخه‌ی اصلی تقسیم‌بندی نمود. در این‌جا علاوه بر بیان این سه شاخه‌ی اصلی، به بیان مثال‌هایی از هر کدام می‌پردازیم. این مثال‌ها بعضاً خود دارای زیرشاخه‌هایی بوده که از بیان آن‌ها صرف نظر می‌کنیم.

- «حق طبع و نشر[۱]» و «مالکیت معنوی[۲]»:
  - تشخیص سرقت ادبی در مقالات
  - تشخیص کپی برداری در تمارین برنامه‌نویسی دانش آموزان
  - بررسی ادعای مالکیت نرم‌افزارهای تجاری
  - تشخیص هویت برنامه‌نویس اصلی یک برنامه‌ی مخرب یا سرقتی
  - حل اختلافات میان کارشناسان در تعیین نویسنده‌ی اصلی برخی آثار ادبی

- جرایم سایبری[۳]:
  - تشخیص نویسنده‌ی رایانامه‌ها و پیام‌های اینترنتی مجرمانه
  - تشخیص هویت اصلی نویسنده‌ی وبلاگ‌ها و سایت‌های دارای مصادیق مجرمانه
  - تشخیص هویت اصلی افراد در شبکه‌های اجتماعی
  - پیگیری تهدیدات اینترنتی[۴]
  - پیگیری حملات سایبری از جمله ویروس‌ها، تروجان‌ها و ...

- امور اطلاعاتی و امنیتی[۵]:
  - پیگیری اعلامیه‌های تروریستی[۶]
  - تعیین هویت کاربر استفاده کننده از یک سیستم

---

[1] Copyright
[2] Intellectual Property
[3] Cyber-Crimes
[4] Cyber-Bullying
[5] Intelligence and Security
[6] Terrorist Proclamations

۱-۵ ساختار پایان‌نامه

در پایان فصل اول، نگاهی گذرا به سایر فصول این پایان‌نامه می‌اندازیم. ادامه‌ی پایان‌نامه به صورت زیر سازماندهی شده است:

- **فصل دوم: مروری بر پژوهش‌های گذشته**

در این فصل راه‌حل‌ها و روش‌های گوناگونی که تا کنون برای حل مسائل تخصیص نویسنده ارائه شده‌اند، تقسیم‌بندی شده و مراجع مربوطه بیان می‌شوند.

- **فصل سوم: مباحث نظری**

در این فصل هشت مورد از مهم‌ترین روش‌های NDP که تا کنون ارائه شده‌اند، شرح داده می‌شود. در انتهای فصل نیز یک تقسیم‌بندی کلی از این هشت روش ارائه می‌گردد.

- **فصل چهارم: معرفی روش‌ها پیشنهادی و مجموعه‌متن‌های مورداستفاده**

در این فصل ابتدا دو روش جدید به منظور حل مسائل تخصیص نویسنده پیشنهاد می‌گردد. سپس چهار مجموعه‌متنی که در این پایان‌نامه از آن‌ها استفاده شده است، معرفی شده و با هم مقایسه می‌شوند.

- **فصل پنجم: آزمایش‌ها، نتایج و تحلیل آن‌ها**

در این فصل ده روش ارائه‌شده در پایان‌نامه (هشت موردِ بیان‌شده در فصل سوم به همراه دو روش پیشنهادی) بر روی چهار مجموعه‌متن معرفی‌شده در فصل چهارم اعمال شده و نتایج بدست‌آمده تحلیل می‌شود.

- **فصل ششم: مطالعات موردی در حکایات و غزلیات**

در این فصل دو مطالعه‌ی موردی در حوزه‌ی ادبیات شرح داده می‌شوند. مطالعه‌ی اول بر روی گلستان سعدی بوده و مطالعه‌ی دوم مربوط به غزلیات با سبک هندی می‌باشد.

- **فصل هفتم: نتیجه‌گیری و پیشنهادات**

در این فصل نتایج بدست‌آمده از پایان‌نامه مختصراً شرح داده می‌شوند. هم‌چنین پیشنهاداتی برای پژوهش‌های آینده مطرح می‌شود.

فصل دوم

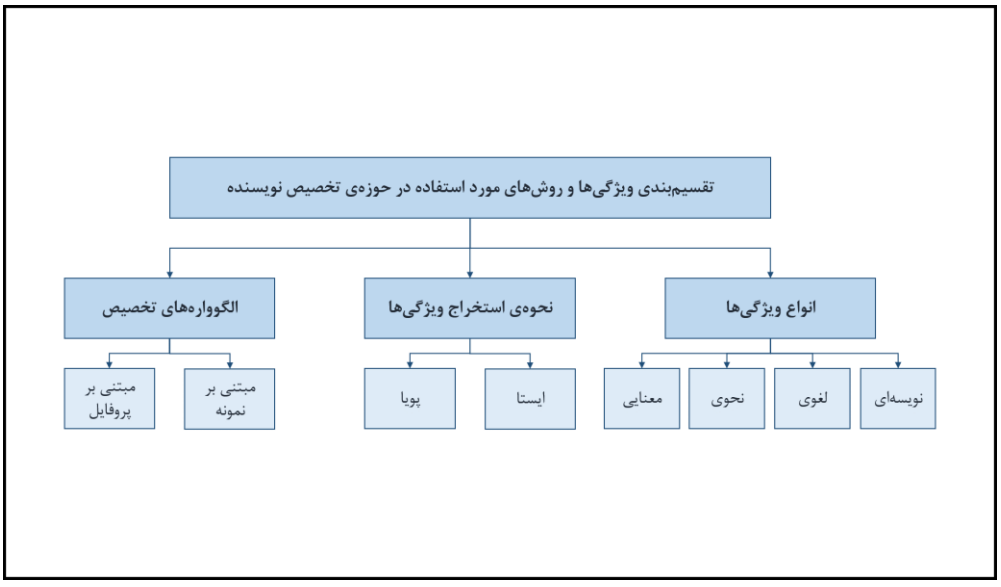
مروری بر پژوهش‌های گذشته

۲-۱ مقدمه

در فصل اول با صورت‌مسئله‌های مختلف موجود در حوزه‌ی تخصیص متن آشنا شدیم. در فصل دوم مروری گذرا خواهیم داشت به روش‌هایی که تا کنون جهت حل مسائل تخصیص متن ارائه شده اند. در این فصل ویژگی‌ها و روش‌های مختلف ارائه‌شده در حوزه‌ی تخصیص نویسنده از سه دیدگاه بررسی و تقسیم‌بندی می‌شوند (شکل ۲-۱):

۱) بررسی ویژگی‌های مختلف ارائه‌شده در حوزه‌ی تخصیص متن (بخش ۲-۲)
۲) بررسی روش‌های استخراج ویژگی (بخش ۲-۳)
۳) بررسی الگوواره[1]های تخصیص (بخش ۲-۴)

سپس نگاهی به پژوهش‌های گذشته در زبان فارسی انداخته (پیوست ۲) و در انتهای فصل نیز با مقایسه‌ی ویژگی‌ها و روش‌های شرح‌داده‌شده، حوزه‌ی خاصی از ویژگی‌ها و روش‌ها را انتخاب نموده (پیوست ۳) و در ادامه‌ی پایان‌نامه بر روی آن متمرکز می‌شویم.

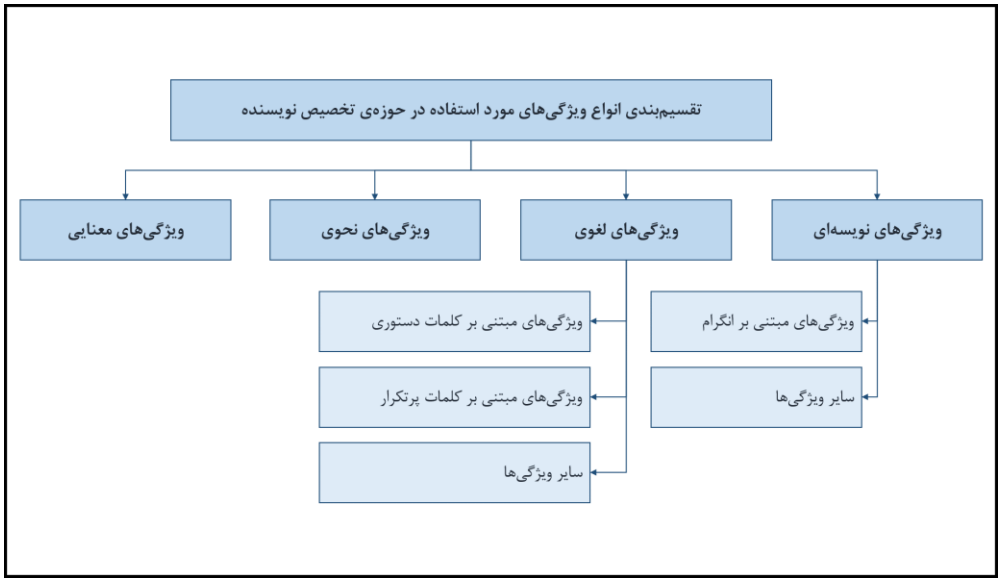


شکل ۲-۱: دیدگاه‌های مختلف جهت بررسی ویژگی‌ها و روش‌های مختلف ارائه‌شده در حوزه‌ی تخصیص نویسنده

[1] Paradigm

۲-۲ انواع مختلف ویژگی‌های مورد استفاده در تخصیص نویسنده

از آن‌جا که هدف اصلی حوزه‌ی تخصیص نویسنده، شناسایی نویسنده‌ی متن می‌باشد؛ لذا ویژگی‌های استخراج‌شده از یک متن می‌بایست نمایانگر سبک نویسندگی نویسنده‌ی متن باشند. از طرفی باید ویژگی‌های استخراج‌شده به صورت کمی بوده تا بتوان آن‌ها را در یک فرایند هوشمند استفاده نمود. پس به منظور طراحی سیستم AAS می‌بایست بتوانیم ویژگی‌هایی عددی از متن استخراج کنیم که نمایانگر سبک نویسندگی نویسنده‌ی آن بوده و میان نویسندگان مختلف متمایز باشند. چنین ویژگی‌هایی را اصطلاحاً «نشانگر سبْک[1]» می‌نامند [۳۱]. بر مبنای پژوهش‌های [۲۸ و ۳۱] نشانگرهایی که تا کنون ارائه شده‌اند را می‌توان به چهار دسته‌ی اصلی تقسیم نمود (شکل ۲-۲):



شکل ۲-۲: تقسیم‌بندی انواع ویژگی‌های مورد استفاده در حوزه‌ی تخصیص نویسنده

[1] Style Marker

۲-۲-۱ ویژگی‌های نویسه‌ای

در این دیدگاه، متن به عنوان دنباله‌ای از نویسه‌ها در نظر گرفته می‌شود. بر مبنای چنین دیدگاهی می‌توان نشانگرهای سبک مختلفی را از یک متن استخراج نمود، که در دو دسته‌ی کلی قابل تقسیم‌بندی می‌باشند:

الف) ویژگی‌های مبتنی بر انگرام[1]

این نوع ویژگی‌ها عملکرد بسیار موفقی در حوزه‌ی تخصیص نویسنده از خود نشان داده در دهه‌ی اخیر بسیار موردتوجه پژوهش‌گران این حوزه قرار گرفته‌اند. [۳۱]. دلیل اصلی رواج استفاده از این ویژگی، سادگی پیاده‌سازی در عین دقت بالای آن می‌باشد. از جمله پژوهش‌هایی که در آن‌ها ویژگی‌های مبتنی بر انگرام مورد استفاده قرار گرفته است، می‌توان به موارد زیر اشاره نمود: کجل در سال ۱۹۹۴ میلادی [۷۲]، فرسیس و هلمز در سال ۱۹۹۶ میلادی [۷۳]، پنگ و همکاران در سال ۲۰۰۳ میلادی [۷۴]، کسلج و همکاران در سال ۲۰۰۳ میلادی [۷۵] و استاماتاتوس در سال ۲۰۰۶ [۷۶].

همانگونه که در بخش ۲-۶ بیان خواهد شد، این نوع از ویژگی‌ها مبنای اصلی تحقیقات این پایان‌نامه را شکل می‌دهند، لذا به طور مجزا و کامل در فصل سوم شرح داده خواهند شد. هم‌چنین پژوهش‌های نوین صورت گرفته در سال‌های اخیر (سال ۲۰۱۰ میلادی به بعد) نیز در فصل سوم بیان می‌شوند.

ب) سایر ویژگی‌ها

این نوع از ویژگی‌ها هیچ ارتباطی به انگرام نداشته و مستقیماً با توجه به نویسه‌ها استخراج می‌شوند. از جمله‌ی این ویژگی‌ها می‌توان به موارد زیر اشاره نمود: تعداد نویسه‌های الفبایی[2]، تعداد نویسه‌های رقمی[3]، تعداد نشانه‌های نوشتاری[4]، تعداد نویسه‌های فاصله و تعداد نویسه‌های بزرگ و کوچک[5] (در زبان‌های لاتین)

---

[1] N-gram-Based Features
[2] Alphabetic Character Count
[3] Digit Character Count
[4] Punctuation-Marks Count
[5] Uppercase and Lowercase characters

۲-۲-۲ ویژگی‌های لغوی

در این دیدگاه، متن به عنوان دنباله‌ای از «توکن[1]»ها در نظر گرفته می‌شود که در کنار هم یک جمله را تشکیل می‌دهند. هر توکن شامل یکی از این موارد است: یک کلمه[2]، یک رقم[3] یا یک نشانه‌ی نوشتاری[4]. این نوع دیدگاه به متن (که قدیمی‌ترین دیدگاه‌ها می‌باشد) در گذشته بسیار مورد توجه پژوهش‌گران قرار گرفته بوده، اما پس از ارائه‌ی ویژگی‌های مبتنی بر انگرام و موفقیت آن در حوزه‌های مختلف دسته‌بندی متن (از جمله تخصیص نویسنده)، از شدت گرایش به این نوع ویژگی‌ها کاسته شد [۳۱].

ویژگی‌هایی که با استفاده از این دیدگاه از یک متن استخراج می‌شوند را می‌توان به سه دسته تقسیم‌بندی کرد:

الف) ویژگی‌های مبتنی بر کلمات دستوری[5]

همانگونه که از نام این دسته از ویژگی‌ها بر می‌آید، مبنای اصلی آن‌ها کلمات دستوری می‌باشد. لذا برای شناخت بهتر این نوع از ویژگی‌های لغوی بهترست ابتدا تعریفی از کلمات دستوری ارائه دهیم. در فرهنگ لغت لانگمن[6] کلمه‌ی دستوری بدین‌صورت تعریف شده است: «کلمه‌ایست که به خودی خود معنایی نداشته؛ اما رابطه‌ی بین سایر کلمات جمله را نشان می‌دهد.»

حال سؤالی که ممکن است پیش آید آنست که این کلمات چگونه تبدیل به ویژگی‌های عددی می‌شوند؟ پاسخ آنست که ابتدا با توجه به زبان متون موجود در مجموعه‌متن، مجموعه‌ای از کلمات دستوری مشخص انتخاب می‌شود. سپس تعداد هر یک از آنها در یک متن، به عنوان یک ویژگی عددی برای آن متن در نظر گرفته می‌شود. به عنوان مثال اگر مجموعه‌ای شامل ۱۰۰ کلمه‌ی دستوری انتخاب نماییم، آن‌گاه می‌توانیم از یک متن ۱۰۰ ویژگی عددی استخراج کنیم که هر کدام از این ۱۰۰ ویژگی، تعداد تکرار یکی از کلمات دستوری در متن می‌باشند.

---

[1] Token
[2] Word
[3] Digit (or Number)
[4] Punctuation Mark
[5] Function Words
[6] Longman Dictionary

تا کنون مجموعه‌های زیادی از کلمات دستوری برای زبان انگلیسی ارائه شده، اما اطلاعات کمی در مورد نحوه‌ی انتخاب آن‌ها توسط ارائه‌کنندگان بیان شده است [۳۱]. از جمله‌ی این مجموعه‌ها می‌توان به موارد زیر اشاره نمود:

- آرگامون و همکاران در سال ۲۰۰۳ میلادی: مجموعه از ۳۰۳ کلمه‌ی دستوری [۳۶]
- کوپل و اسکلر در سال ۲۰۰۳ میلادی: مجموعه از ۴۸۰ کلمه‌ی دستوری [۳۷]
- عباسی و چن در سال ۲۰۰۵ میلادی: مجموعه‌ای از ۱۵۰ کلمه‌ی دستوری [۳۸]
- ژائو و زوبل در سال ۲۰۰۵ میلادی: مجموعه‌ای از ۳۶۵ کلمه‌ی دستوری [۳۹]
- آرگامون و همکاران در سال ۲۰۰۷ میلادی: مجموعه‌ای از ۶۷۵ کلمه‌ی دستوری [۴۰]

هم‌چنین داورپناه و همکاران در سال ۲۰۰۹ میلادی، مجموعه‌ای از ۹۲۲ کلمه‌ی کلیدی در زبان فارسی را ارائه دادند[۵۲].

البته می‌توان به جای استفاده از مجموعه‌های فوق، مجموعه‌ی کلمات دستوری را مستقیماً از مجموعه‌متن استخراج نمود، بدین‌صورت که تعدادی از پر تکرارترین کلمات دستوری متن را در یک مجموعه جمع‌آوری کرده و از آن جهت استخراج ویژگی استفاده نماییم. (همانگونه که در بخش ۲-۳ بیان خواهد شد، این روش را استخراج پویا می‌نامند.)

دلیل اصلی عملکرد موفق کلمات دستوری در مسئله‌ی تخصیص نویسنده آنست که این کلمات [۵۰]:

- مستقل از موضوع می‌باشند. لذا موضوع متن هر چه که باشد، تأثیری در وجود یا عدم وجود این کلمات ندارد.
- با اینکه از نظر معنایی کم‌ارزش و یا بی‌معنی هستند، اما نقش مهمی در دستور زبان دارند و لذا بسیار غیر محتمل است که تحت کنترل آگاهانه‌ی نویسنده قرار گیرند.

ب) ویژگی‌های مبتنی بر کلمات پر تکرار[1]

این دسته از ویژگی‌ها بر مبنای کلمات پر تکرار تعریف و محاسبه می‌شوند. روند کلی کار مشابه ویژگی‌های مبتنی بر کلمات دستوری می‌باشد. این روند را می‌توان در دو گام بیان نمود:

[1] Most-Frequent Words

- ابتدا تعدادی از پر تکرارترین کلمات موجود در متون مجموعه‌متن استخراج می‌شوند (مثلاً ۱۰۰ کلمه).
- سپس برای تبدیل یک متن به کمیت‌های عددی، تعداد تکرار تک تک کلمات موجود در مجموعه‌ی انتخاب‌شده را محاسبه می‌کنیم. هر یک از این تعداد تکرارها یک ویژگی به حساب می‌آیند. یعنی مثلاً اگر در مرحله‌ی اول مجموعه‌ای شامل ۱۰۰ کلمه انتخاب نماییم، می‌توان از هر متن ۱۰۰ ویژگی استخراج نمود.

پژوهش‌گران مختلفی از این نوع ویژگی‌ها جهت تخصیص نویسنده‌ی متن استفاده نموده‌اند، که از جمله‌ی آن‌ها می‌توان به موارد زیر اشاره نمود:

- باروز در سال ۱۹۹۲ میلادی: مجموعه‌ای از ۱۰۰ کلمه با بیش‌ترین تکرار در مجموعه‌متن [۴۱]
- کوپل و همکاران در سال ۲۰۰۷ میلادی: مجموعه‌ای از ۱۵۰ کلمه با بیش‌ترین تکرار در مجموعه‌متن [۴۲]
- استاماتاتوس در سال ۲۰۰۶ میلادی: مجموعه‌ای از ۱۰۰۰ کلمه با بیش‌ترین تکرار در مجموعه‌متن [۴۳]
- مادیگان و همکاران در سال ۲۰۰۵ میلادی: مجموعه‌ای از کلماتی که حداقل دوبار در متن تکرار شده‌اند [۴۴]

ج) سایر ویژگی‌ها

این نوع از ویژگی‌ها ارتباطی به کلمات دستوری یا کلمات پر تکرار نداشته و مستقیماً با توجه به کلمات متن استخراج می‌شوند. از جمله‌ی آن‌ها می‌توان به موارد زیر اشاره نمود: طول متوسط لغات متن، تعداد متوسط کلمات در جمله، تعداد هاپاکس لگومنا[1] (کلماتی که فقط یک‌بار در متن استفاده شده اند.)، تعداد هاپاکس دیسلگومنا[2] (کلماتی که دقیقاً دوبار در متن استفاده شده اند.)، نسبت تعداد کلمات کوتاه (کلماتی با طول کمتر از ۴ نویسه) به کل کلمات.

[1] Hapax Legomena
[2] Hapax Dislegomena

۳-۲-۲ ویژگی‌های نحوی[1]

یکی از روش‌های دیگر جهت استخراج نشانگرهای سبک، استفاده از اطلاعات نحوی می‌باشد. الگوهای به کار رفته توسط نویسنده برای تشکیل متن را ویژگی‌های نحوی می‌گویند [۵۰]. ایده‌ی اصلی استفاده از چنین اطلاعاتی از آن‌جا الهام گرفته شده است که برخی معتقدند: «نویسنگان به صورت ناخودآگاه تمایل دارند از الگوهای نحوی مشابهی در تمامی متون خود استفاده نمایند [۳۱].» هم‌چنین موفقیت روش استفاده از کلمات دستوری در مسئله‌ی تخصیص نویسنده، خود دلیلی بر مناسب بودن ویژگی‌های نحوی در استخراج نشانگرهای سبک می‌باشد؛ زیرا معمولاً کلمات دستوری در ساختارهای نحوی مشخصی به کار می‌روند [۳۱].

پس از آشنایی کلی با تعریف ویژگی‌های نحوی، در ادامه یکی از مهم‌ترین روش‌های استخراج نشانگرهای سبک بر مبنای اطلاعات نحوی را مختصراً شرح می‌دهیم. این روش ابتدا در سال ۲۰۰ میلادی توسط استاماتاتوس و همکاران ارائه شده [۳۲] و یک سال بعد (یعنی سال ۲۰۰۱ میلادی) توسط همان گروه تغییراتی در آن به وجود آمد [۳۳]. روند کلی روش ارائه‌شده بدین‌صورت است:

۱) ابتدا کل متن به مجموعه‌ای «جملات[2]» تقسیم‌بندی می‌شود.
۲) سپس هر جمله به مجموعه‌ای از «عبارات[3]» تجزیه می‌گردد.
۳) هر عبارت یافته‌شده در گام قبل به یکی از پنج دسته‌ی زیر تخصیص داده می‌شود: عبارت اسمی[4] (NP)، عبارت اضافی[5] (PP)، عبارت فعلی[6] (VP)، عبارت قیدی[7] (AP) و دنباله‌ای از حروف ربط[8] (CON)
۴) درنهایت مواردی از قبیل کمیت‌های زیر را به عنوان ویژگی عددی متن اولیه (نشانگرهای سبک) محاسبه و استخراج می‌نماییم

---

[1] Syntactic Features
[2] Sentences
[3] Phrases
[4] Noun Phrase (NP)
[5] Prepositional Phrase (PP)
[6] Verb Phrase (VB)
[7] Adverbial Phrase (AP)
[8] Sequence of conjunctions (CON)

- نسبت تعداد عبارات: نسبت (تعداد عباراتی که در دسته‌ی عبارات اسمی طبقه‌بندی شده‌اند) به (تعداد کل عبارات متن)، نسبت (تعداد عبارات اضافی) به (کل عبارات)، نسبت (تعداد عبارات فعلی) به (کل عبارات) و ...
- نسبت تعداد کلمات موجود در عبارات: نسبت (تعداد کلمات موجود در عبارات اسمی) به (تعداد عبارات اسمی)، نسبت (تعداد کلمات موجود در عبارات اضافی) به (تعداد عبارات اضافی)، نسبت (تعداد کلمات موجود در عبارات فعلی) به (تعداد عبارات قیدی) و ...

لازم به ذکرست که جهت پیاده‌سازی عملیات فوق نیاز به برخی سیستم‌های جانبی می‌باشد، از جمله: سیستمی که بتواند یک متن را به جملات تجزیه کند، سیستمی که بتواند یک جمله را به عبارات تشکیل‌دهنده‌ی آن تجزیه کند و سیستمی که بتواند یک عبارت را از نظر دستور زبان به یکی از پنج گروه اسمی، اضافی، فعلی، قیدی و ربطی تخصیص دهد.

سیستم‌هایی نظیر موارد فوق را اصطلاحاً ابزارهای NLP می‌نامند، زیرا نحوه‌ی طراحی چنین سیستم‌هایی در شاخه‌ی «پردازش زبان‌های طبیعی[1]» مورد مطالعه قرار می‌گیرد. این ابزارها عموماً وابسته به زبان متن می‌باشند. به عنوان مثال سیستمی که نوع دستوری یک عبارت را در زبان انگلیسی تعیین می‌کند، کاربردی در زبان فارسی ندارد. از طرفی معمولاً طراحی و پیاده‌سازی چنین سیستم‌هایی (مخصوصاً زمانی که دقت بالا مطلوب است) دشوار می‌باشد، زیرا نیاز به دانش ادبیاتی در مورد زبان متن دارند. این نیاز به ابزارهای NLP، به عنوان یک نقطه‌ضعف برای ویژگی‌های نحوی در مقایسه‌ی با ویژگی‌های لغوی و نویسه‌ای به حساب می‌آید.

برخی دیگر از پژوهش‌هایی که بر روی ویژگی‌های نحوی صورت پذیرفته عبارتند از: باین و همکاران در سال ۱۹۹۶ میلادی [۷۷]، کوپل و اسکلر در سال ۲۰۰۳ میلادی [۷۸]، گامون در سال ۲۰۰۴ میلادی [۷۹]، هرست و فیگوینا در سال ۲۰۰۷ میلادی [۸۰] و ون هالترن در سال ۲۰۰۷ میلادی [۸۱]

۴-۲-۲ ویژگی‌های معنایی

ویژگی‌های معنایی نوع دیگر از ویژگی‌ها هستند که بر مبنای مفهوم عبارات متن (نه ساختارهای نحوی آن‌ها) بنا نهاده شده اند. همانطور که در ادامه خواهیم دید این ویژگی‌ها اگرچه نسبت به ویژگی‌های لغوی

[1] Natural Language Processing (NLP)

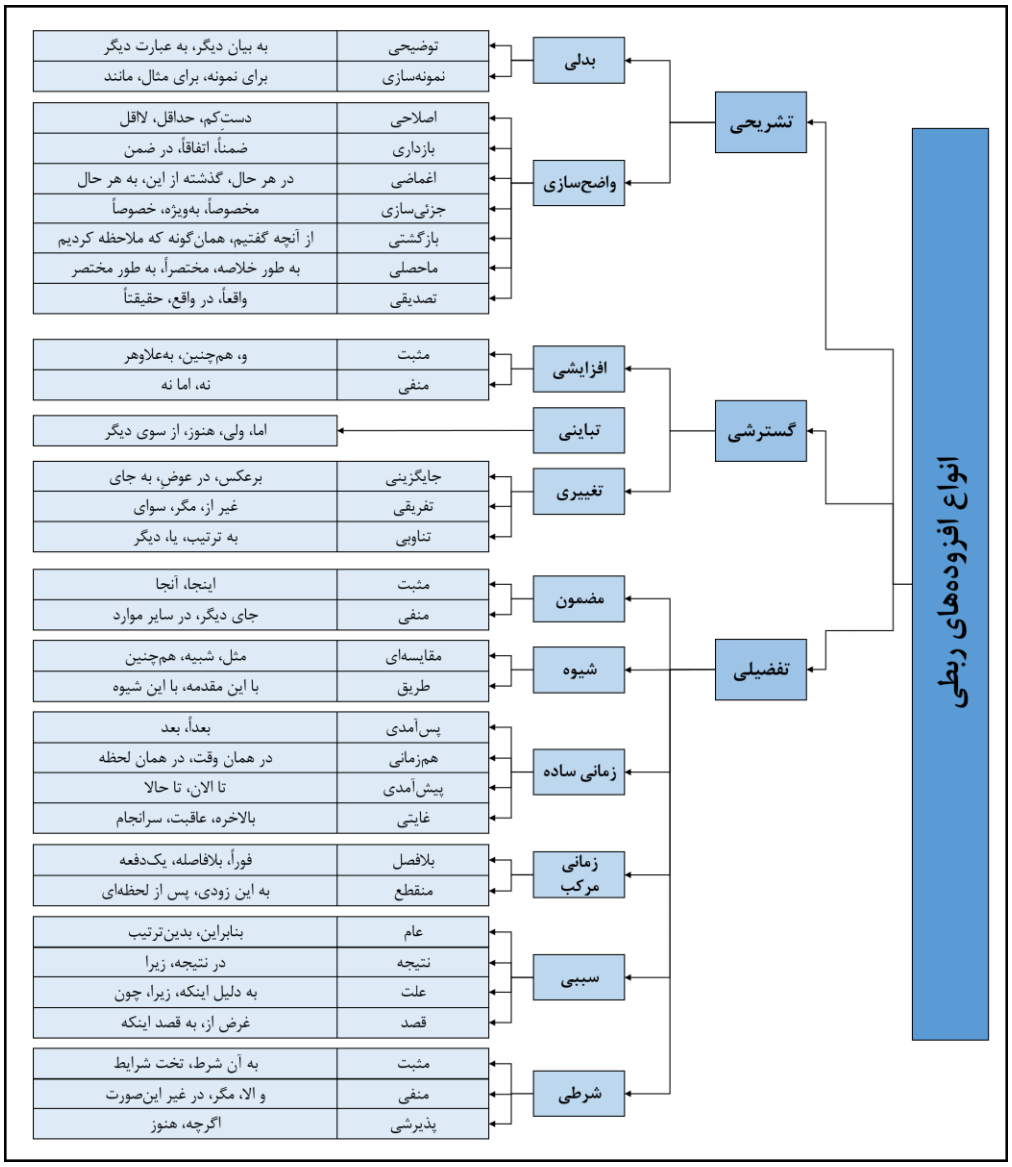
و نویسه‌ای، منطق محکم‌تری برای اثبات مؤثر بودن در مسئله‌ی تخصیص نویسنده دارند، اما همانند ویژگی‌های نحوی برای پیاده‌سازی نیاز به برخی سیستم‌های جانبی دارند. تا کنون روش‌های زیادی برای استخراج نشانگرهای سبک مبتنی بر اطلاعات معنایی متن ارائه شده است که در ادامه یکی از مهم‌ترین آن‌ها را شرح می‌دهیم. این روش که توسط آرگامون و همکاران در سال ۲۰۰۷ ارائه شده است [۳۴]، بر مبنای یک تئوری در ادبیات انگلیسی بنا نهاده شده است. این تئوری «دستور نقش‌گرای نظام‌مند[1]» نام داشته و اولین‌بار توسط هالیدی در سال ۱۹۹۴ ارائه شده است [۳۵]. مراحل کار در روش ارائه‌شده به صورت زیر است:

۱) ابتدا متن به مجموعه‌ای از بند[2]ها تجزیه می‌شود.
۲) سپس کلمات و عباراتی که بندها را بهم متصل می‌کنند، شناسایی می‌شوند (کلماتی همچون ’while‘ و عباراتی همانند ’in other words‘). این کلمات و عبارات را اصطلاحاً «افزوده‌های ربطی[3]» می‌نامیم.
۳) در گام سوم هر افزوده‌ی ربطی شناسایی‌شده در گام قبل را به یکی از دسته‌های نمایش‌داده‌شده در شکل ۲-۳ تخصیص می‌دهیم (مبنای این تخصیص آنست که افزوده‌ی ربطی موردنظر چگونه دو بند قبل و بعد از خود را بهم متصل نموده است. مثلاً افزوده‌ی «برای نمونه»، در واقع بیان‌گر آنست که هدف بند جدید تشریح بیش‌تر بند قبلی از طریق ذکر مثال می‌باشد، لذا در دسته‌ی افزوده‌های «تشریحی»، شاخه‌ی «نمونه‌سازی» قرار می‌گیرد.)
۴) درنهایت مواردی از قبیل کمیت‌های زیر را به عنوان نشانگرهای سبک متن اولیه محاسبه و استخراج می‌نماییم:

- نسبت تعداد افزوده‌های هر دسته: نسبت (تعداد افزوده‌هایی که در دسته‌ی تشریحی طبقه‌بندی شده‌اند) به (تعداد کل افزوده‌های ربطی متن)، نسبت (تعداد افزوده‌هایی که در دسته‌ی گسترشی طبقه‌بندی شده‌اند) به (تعداد کل افزوده‌های ربطی متن) و ...
- نسبت تعداد هر کدام از زیردسته‌ها: نسبت (تعداد افزوده‌های واضح‌سازی) به (تعداد کل افزوده‌های تشریحی)، نسبت (تعداد افزوده‌های تباینی) به (تعداد کل افزوده‌های گسترشی) و ...

---

[1] Systemic Functional Grammar (SFG)
[2] Clause
[3] Conjunctives



شکل ۲-۳: انواع افزوده‌های ربطی

واضح است که برای پیاده‌سازی روش فوق نیاز به دو ابزار NLP داریم:

- سیستمی که بتواند افزوده‌های ربطی را در یک متن تشخیص دهد.
- سیستمی که بتواند هر افزوده‌ی ربطی را به یکی از زیرشاخه‌های موجود در شکل ۲-۳ تخصیص دهد.

در مجموع می‌توان گفت که استفاده از ویژگی‌های معنایی نیز همانند ویژگی‌های نحوی، مستلزم در اختیار داشتن ابزارهای NLP خاص می‌باشد. ابزارهایی که علاوه بر وابستگی شدید به زبان متن، عموماً فرایند طراحی دشواری نیز دارند.

علاوه بر روش فوق، پژوهش‌های دیگری نیز در حوزه‌ی ویژگی‌های معنایی صورت پذیرفته است، از جمله: دیروستر و همکاران در سال ۱۹۹۰ میلادی [۸۲]، گامون در سال ۲۰۰۴ میلادی [۷۹] و مک‌کارتی و همکاران در سال ۲۰۰۶ میلادی [۸۳]

۲-۳ روش‌های استخراج ویژگی

در بخش ۲-۲ با انواع ویژگی‌های مناسب برای استخراج نشانگرهای سبک آشنا شدیم. حال فرض کنیم یکی از آن‌ها را انتخاب نمودیم، به عنوان مثال تصمیم گرفتیم از کلمات دستوری استفاده نماییم. در نحوه‌ی استخراج این کلمات دو حالت پیش می‌آید: یا این کلمات دستوری را بر اساس دانش قبلی و بدون توجه به مجموعه‌متن انتخاب کرده (مثلاً مجموعه‌ی ۹۲۲ کلمه‌ای داورپناه [۵۲] را انتخاب می‌کنیم)، و یا آن‌ها را بر مبنای مجموعه‌متنی که در اختیار داریم بر می‌گزینیم. علاوه بر ویژگی‌های مبتنی بر کلمات دستوری، برخی دیگر از ویژگی‌های ارائه‌شده در بخش قبلی نیز همین حالت دارند و آن‌ها را به دو صورت می‌توان استخراج نمود. به عبارت دیگر در حوزه‌ی تخصیص نویسنده، دو نحوه‌ی استخراج ویژگی وجود دارد: ایستا[1] و پویا[2]. در ادامه هر یک را جداگانه شرح می‌دهیم.

۲-۳-۱ استخراج ایستا

در این حالت ویژگی‌ها قبل از شروع فرایند آموزش و بدون توجه به متون موجود در مجموعه‌متن انتخاب می‌شوند.

به منظور شرح بیش‌تر این حالت، همان مثال ابتدای بخش را در نظر می‌گیریم. می‌خواهیم از کلمات دستوری به عنوان ویژگی‌هایی جهت تخصیص نویسنده استفاده نماییم. در حالت استخراج ایستا این کلمات

[1] Static
[2] Dynamic

بدون توجه به مجموعه‌متن انتخاب می‌شوند. یعنی با توجه به زبان متون موجود در مجموعه‌متن، اما مستقل از کلمات تشکیل‌دهنده‌ی آن‌ها، مجموعه‌ای از کلمات دستوری را جهت استخراج ویژگی برمی‌گزینیم. مثلاً اگر زبان متون انگلیسی بود، مجموعه‌ی ۶۷۵ کلمه‌ی آرگامون [۴۰] را انتخاب می‌کنیم، حال متون داخل مجموعه‌متن هرچه باشند مهم نیست.

۲-۳-۲ استخراج پویا

در این حالت ویژگی‌ها در حین فرایند آموزش و با توجه به متون موجود در مجموعه‌متن انتخاب می‌شوند.

مجدداً به سراغ مثال ابتدای فصل می‌رویم (می‌خواهیم از کلمات دستوری به عنوان ویژگی‌هایی جهت تخصیص نویسنده استفاده نماییم). در حالت استخراج ایستا این کلمات با توجه به مجموعه‌متن انتخاب می‌شوند. به عنوان مثال ۵۰۰ تا از پر تکرارترین کلمات دستوری موجود در مجموعه‌متن را به عنوان مجموعه‌ی کلمات دستوری بر می‌گزینیم و سپس بر مبنای این مجموعه استخراج ویژگی می‌نماییم.

معمولاً در هر دسته از انواع ویژگی‌های تخصیص نویسنده، یکی از این دو روش استخراج رایج‌تر می‌باشد، به عنوان مثال:

- در ویژگی‌های مبتنی بر کلمات دستوری، استخراج ایستا رایج‌تر است.
- در ویژگی‌های مبتنی بر کلمات پرتکرار، استخراج پویا رایج‌تر است.
- در ویژگی‌های مبتنی بر انگرام، استخراج پویا رایج‌تر است.

در مجموع می‌توان گفت: هر چند در گذشته استخراج ایستا در نزد پژوهشگران تخصیص نویسنده بسیار رایج بوده است، اما پس از ارائه‌ی روش استخراج پویا و موفقیت بالاتر آن در حل مسائل این حوزه، امروزه استخراج پویا بیش‌تر از استخراج ایستا مورد استفاده‌ی پژوهش‌گران قرار می‌گیرد [۲۸].

۲-۴ الگوواره‌های تخصیص نویسنده

پس از آشنایی با ویژگی‌های مورداستفاده در حوزه‌ی تخصیص نویسنده و روش‌های استخراج آن، حال به سراغ بررسی دو الگوواره‌ی اساسی در این حوزه می‌رویم. همانطور که در ادامه خواهیم دید این الگوواره‌ها در واقع نحوه‌ی رفتار ما با متون موجود در مجموعه‌متن را تعیین می‌کنند، این‌که تک تک متون را جداگانه بررسی نموده و یا این‌که تمامی متون یک نویسنده را در کنار هم تحلیل نماییم. لازم بذکرست که جهت

حل یک مسئله‌ی تخصیص متن ابتدا می‌بایست ویژگی مورداستفاده را انتخاب نمود، سپس ایستا یا پویا بودن استخراج آن را تعیین کرد و در نهایت یکی از دو الگوواره‌ی اساسی این حوزه را مورد استفاده قرار داد.

۲-۴-۱ الگوواره‌ی مبتنی بر نمونه[1]

در این الگوواره هر متن موجود در مجموعه‌متن به صورت مستقل تحلیل شده و نمایش مختص به خود را دارد [۲۸]. این الگوواره خود به سه دسته از مدل‌ها تقسیم‌بندی می‌شود: مدل‌های فضای برداری[2]، مدل‌های مبتنی بر شباهت[3] و مدل‌های فرا-آموزشی[4].

ما در ادامه فقط مدل فضای برداری را شرح می‌دهیم، زیرا اکثر پژوهش‌های صورت گرفته در مورد الگوواره‌ی مبتنی بر نمونه، بر روی آن متمرکز شده است، تا جایی که پوسا و استاماتاتوس در سال ۲۰۱۴ میلادی در [۲۲]، هنگام شرح این الگوواره، فقط مدل فضای برداری را شرح داده و حتی نامی از دو مدل (مبتنی بر شباهت و فرا-آموزشی) دیگر نبرده‌اند.

یک نمای کلی از روند کاری مدل فضای برداری در شکل ۲-۴ نمایش داده شده است. مراحل اصلی این روند به شرح زیر است:

۱) ابتدا هر متن موجود در مجموعه‌متن با استفاده از ویژگی مشخص‌شده و نحوه‌ی استخراج انتخاب‌شده، تبدیل به برداری از ویژگی‌ها می‌شود.
۲) پس از آن، تمامی بردارهای استخراج‌شده را در مجموعه‌ای به نام مجموعه‌ی آموزش[5] جمع‌آوری می‌نماییم. این مجموعه در واقع شامل نمونه‌های برچسب‌خورده از تمامی نویسندگان موجود در مجموعه‌متن می‌باشد.
۳) سپس با استفاده از مجموعه‌ی آموزش جمع‌آوری‌شده، یک طبقه‌بند آموزش داده می‌شود.

---

[1] Instance-Based Paradigm
[2] Vector-Space models
[3] Similarity-Based models
[4] Meta-Learning models
[5] Learning Set or Learning Corpus

۴) در نهایت، متن دیده‌نشده همانند سایر متون مجموعه‌متن تبدیل به بردار می‌شود (با همان ویژگی‌های مورداستفاده در گام نخست) و سپس به طبقه‌بند اعمال می‌شود. طبقه‌بند نیز نویسنده‌ی متن دیده‌نشده را از میان طبقه‌های از پیش تعریف‌شده تعیین می‌نماید.

از آن‌جا طبقه‌بند نقش اساسی را در این مدل ایفا می‌کند، گاهی اوقات این مدل را «مبتنی بر طبقه‌بندی[۱]» و یا «مبتنی بر طبقه‌بند[۲]» می‌نامند [۵۳]. البته الزامی وجود ندارد که طبقه‌بند مورد استفاده در این مدل دارای فاز آموزشی باشد. به عنوان مثال می‌توان از الگوریتم k-نزدیک‌ترین همسایه[۳] جهت طبقه‌بندی متن دیده‌نشده استفاده نمود.

پژوهش‌های مختلفی تا کنون جهت پیاده‌سازی این مدل مورد استفاده قرار گرفته‌اند، از جمله: با استفاده از تحلیل مرز جداساز[۴]: پژوهش‌های [۳۲، ۵۴ و ۵۵]، با استفاده از ماشین بردار پشتیبان[۵]: پژوهش‌های [۵۶ تا ۶۰]، با استفاده از درخت تصمیم[۶]: پژوهش‌های [۶۱ تا ۶۳]، با استفاده از شبکه‌های عصبی مصنوعی[۷]: پژوهش‌های [۶۵ تا ۶۹] و با استفاده از الگوریتم ژنتیک[۸]؛ پژوهش [۷۰]

۲-۴-۲ الگوواره‌ی مبتنی بر پروفایل[۹]

در این الگوواره تمامی متون مربوط به یک نویسنده با هم تحلیل شده و نمایش واحدی ارائه می‌دهند. به عبارت دیگر بر خلاف الگوواره‌ی پیشین که در آن هر متن یک نمایش خاص مربوط به خود را داشت، در این الگوواره هر نویسنده نمایش مختص خود را داراست و دیگر تک تک متون یک نویسنده نمایش مستقلی ندارند.

---

[1] Classification-Based
[2] Classifier-Based
[3] k-Nearest Neighbor (kNN)
[4] Discriminant Analysis
[5] Support Vector Machine (SVM)
[6] Decision Tree
[7] Artificial Neural Network (ANN)
[8] Genetic Algorithm (GA)
[9] Profile-Based Paradigm

همانند الگوواره‌ی قبل، این الگوواره نیز به سه دسته از مدل‌ها تقسیم‌بندی می‌شود: مدل‌های مبتنی بر انگرام[۱]، مدل‌های آماری[۲] و مدل‌های فشرده‌سازی[۳]

از میان مدل‌های فوق، مدل مبتنی بر انگرام چنان رواج پیدا کرده است که در تحقیقات جدید دیگر کمتر نامی از دو مدل دیگر دیده می‌شود. همانطور که در بخش ۲-۵ شرح داده خواهد شد، در این پایان‌نامه نیز مدل مبتنی بر انگرام مورد استفاده قرار خواهد گرفت. لذا از میان سه مدل فوق، در ادامه فقط مدل‌های مبتنی بر انگرام را شرح می‌دهیم.

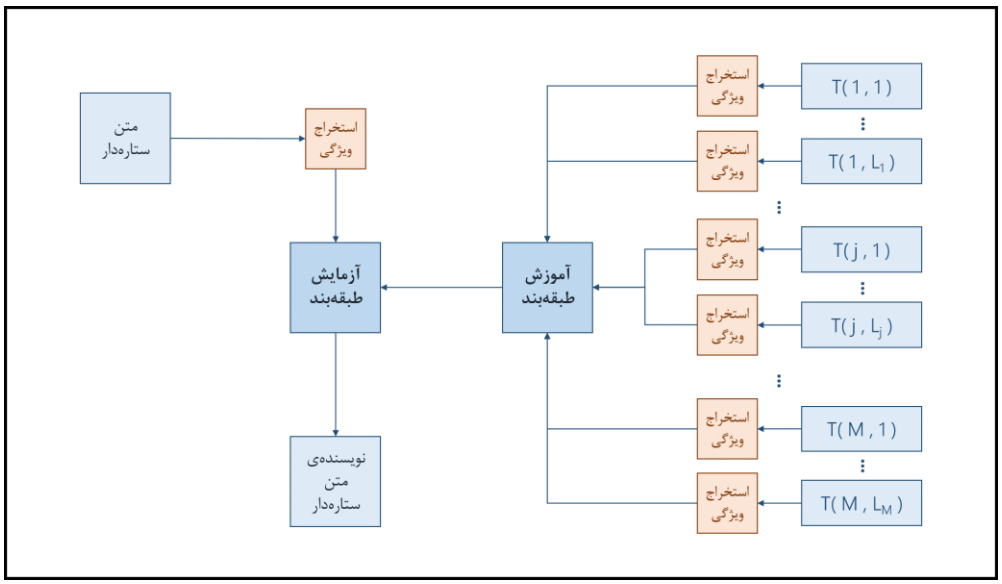
در شکل ۲-۵ نمای کلی از روند کاری الگوواره‌ی مبتنی بر پروفایل نمایش داده شده است. این الگوواره برای تعیین نویسنده‌ی متن دیده‌نشده مراحل زیر را ارائه می‌کند:

۱) ابتدا تمامی متون هر نویسنده را بهم الحاق کرده تا یک متن طولانی برای هر نویسنده ایجاد شود. چنین متنی را «اَبرمتن[۴]» می‌نامیم. در پایان این مرحله، برای هر نویسنده، یک ابرمتن در اختیار خواهیم داشت.
۲) سپس بر مبنای ویژگی‌های انتخاب شده و ایستا یا پویا بودن آن‌ها، از هر ابرمتن استخراج ویژگی کرده و آن‌ها را به یک بردار عددی تبدیل می‌نماییم. در پایان این مرحله، به ازای هر ابرمتن، یک بردار عددی متناظر با آن در اختیار داریم. این بردار عددی متناظر با نویسنده را اصطلاحاً «پروفایل[۵]» آن نویسنده می‌نامند.
۳) متن دیده‌نشده نیز بر مبنای ویژگی‌های انتخاب شده و ایستا یا پویا بودن آن‌ها، تبدیل به یک بردار می‌گردد. این بردار را «پروفایل دیده‌نشده[۶]» می‌نامیم.
۴) در نهایت پروفایل‌های بدست آمده در دو مرحله‌ی قبل (نویسندگان و دیده‌نشده) به سیستم AAS اعمال شده و این سیستم، متن دیده‌نشده را به یکی از نویسندگان اختصاص می‌دهد (معمولاً روند اصلی کار این سیستم بر مبنای محاسبه‌ی فاصله‌ی بین پروفایل دیده‌نشده و پروفایل‌های تک تک نویسندگان می‌باشد. نویسنده‌ای که پروفایل آن دارای کمترین فاصله با پروفایل دیده‌نشده باشد به عنوان پاسخ سیستم اعلام می‌شود.)

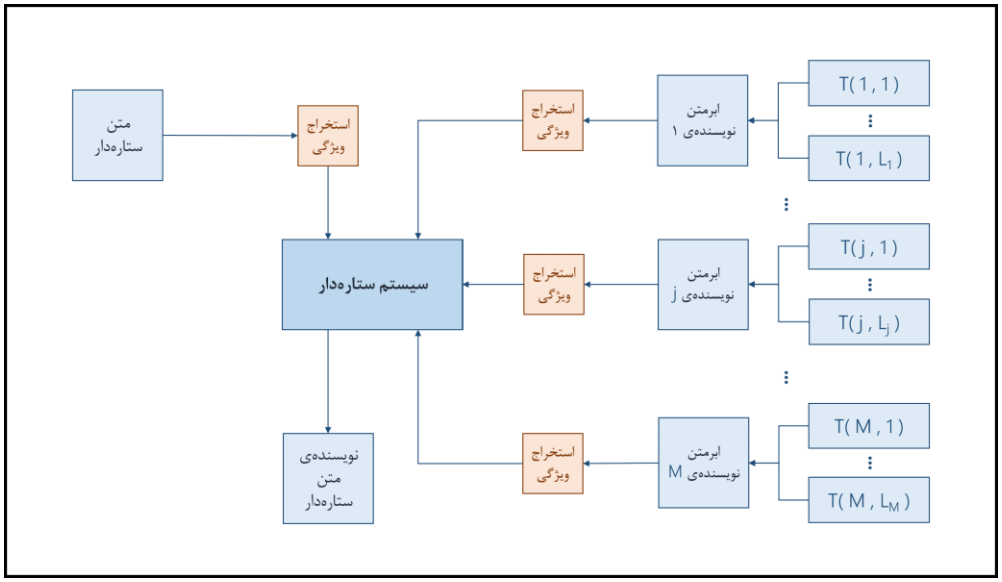
---

[1] N-gram-Based Models
[2] Probabilistic Models
[3] Compression Models
[4] Super-Text
[5] Profile
[6] Stared-Profile

پژوهش‌های زیادی در حوزه‌ی تخصیص نویسنده با استفاده از الگووارهی مبتنی بر پروفایل صورت پذیرفته است، از جمله: کسلج و همکاران در سال ۲۰۰۳ میلادی [۲]، پنگ و همکاران در سال ۲۰۰۴ میلادی [۸۵]، مارتون و همکاران در سال ۲۰۰۵ میلادی [۸۴] و فرانتزسکو و همکاران در سال ۲۰۰۷ میلادی [۲۵].



شکل ۲-۴: نمای کلی از روند الگووارهی مبتنی بر نمونه



شکل ۲-۵: نمای کلی از روند الگووارهی مبتنی بر پروفایل

فصل سوم

مباحث نظری

۳-۱ مقدمه

در مقدمه‌ی این فصل ابتدا مروری کوتاه بر مطالب فصل دوم می‌اندازیم. در فصل قبل ابتدا با انواع ویژگی‌ها و روش‌های ارائه‌شده جهت حل مسائل تخصیص نویسنده آشنا شده، و سپس از میان آن‌ها روش‌هایی که دارای سه ویژگی زیر بودند را جهت مطالعه در این پایان‌نامه، انتخاب نمودیم:

۱) ویژگی‌های مورداستفاده مبتنی بر انگرام در سطح نویسه باشند.
۲) ویژگی‌ها به صورت پویا استخراج شوند.
۳) الگووارهٔ مبتنی بر پروفایل در دستور کار قرار گیرد.

این روش‌ها را اصطلاحاً روش‌های NDP نامیدیم. همانگونه که گفته شد روش‌های زیادی را می‌توان در دسته‌ی روش‌های NDP جای داد. حال در این فصل می‌خواهیم مهم‌ترین آن‌ها را بررسی نماییم. مطالب ارائه‌شده در فصل سوم بدین شرح می‌باشند:

- ابتدا در بخش ۳-۲ برخی از مبانی مقدماتی لازم جهت تشریح روش‌های NDP بیان می‌گردد.
- سپس در بخش ۳-۳ لیستی از روش‌های مطالعه‌شده در این پایان‌نامه، به همراه نام ارائه‌دهندگان آن‌ها ارائه می‌شود.
- پس از آن اصلی‌ترین روش NDP موجود (یعنی روش CNG) را به تفصیل شرح می‌دهیم (بخش ۳-۴).
- در پیوست ۴، هفت مورد از روش‌هایی را که با ایجاد برخی اصلاحات بر روی روش CNG ابداع شده‌اند، مختصراً معرفی می‌کنیم.
- و در نهایت، روش‌های معرفی‌شده را با هم مقایسه می‌کنیم. این مقایسه در انتهای پیوست ۴ آمده است.

۳-۲ مبانی مقدماتی

در این بخش برخی مفاهیم اولیه‌ی موردنیاز جهت تشریح روش‌های NDP ارائه می‌شوند. ترتیب ارائه‌ی این مفاهیم به شرح زیر است:

- ابتدا بحث مختصری در مورد تفاوت نویسه و بایت را ارائه می‌دهیم (قسمت ۳-۲-۱).

- سپس به سراغ انگرام رفته، آن را تعریف نموده و نحوه‌ی استخراج آن را شرح می‌دهیم (قسمت ۳-۲-۲).

لازم بذکرست که انگرام یکی از مفاهیم اساسی مورداستفاده در روش‌های NDP بوده و لذا درک صحیح آن قبل از تشریح این نوع روش‌ها الزامی است.

- در قسمت سوم یک قالب استاندار برای پروفایل متن ارائه می‌شود. (قسمت ۳-۲-۳)
- در چهارمین قسمت، یک روند کلی برای روش‌های NDP ارائه می‌دهیم که توسط تمامی روش‌های ارائه‌شده در این پایان‌نامه (حتی روش‌های پیشنهادی تشریح‌شده در فصل چهارم) مورداستفاده قرار می‌گیرند. (قسمت ۳-۲-۴)

نکته‌ی جالب آنست که تقریباً تمامی روش‌های NDP موجود، بر مبنای یک روش قدیمی طراحی شده‌اند. این روش قدیمی توسط یک فیزیکدان آمریکایی به نام «ویلیام رالف بِنِت[1]» در سال ۱۹۷۶ میلادی و در کتابی با عنوان «حل مسئله‌ی علمی و مهندسی با استفاده از کامپیوتر[2]» ارائه گردید.

- در نهایت دو سیستم اساسی مورداستفاده در یک روش NDP را شرح می‌دهیم (قسمت ۳-۲-۵). این قسمت در واقع خلاصه‌ای از مباحث مطرح‌شده در قسمت قبل (۳-۲-۴) می‌باشد.

۳-۲-۱ نویسه و بایت

در نخستین تحقیقات حوزه‌ی تخصیص نویسنده‌ی متون انگلیسی، گاهی به جای نویسه از بایت[3] استفاده می‌شد. به عبارت دیگر برخی پژوهش‌گران به جای آنکه یک متن را رشته‌ای از نویسه‌های متوالی در نظر بگیرند، آن را به صورت سلسله‌ای از بایت‌ها فرض می‌کردند که پشت سر یکدیگر قرار گرفته اند. سؤالی که در این‌جا مطرح می‌شود آنست که آیا نویسه و بایت تفاوتی با یکدیگر دارند؟ پاسخ این سؤال بستگی به رویه‌ی کدگذاری نویسه‌ها[4]ی متن دارد. دو رویه‌ی رایج برای کدگذاری نویسه‌های متن عبارتند از: استاندارد اَسکی توسعه‌یافته[5] و استاندارد یونیکُد[6].

---

[1] William Ralph Bennett Jr. (January 30, 1930 – June 29, 2008)
[2] Scientific and engineering problem-solving with the computer
[3] Byte (8-bits)
[4] Character-Encoding Scheme
[5] Extended ASCII (American Standard Code for Information Interchange)
[6] Unicode (Universal Code)

دو تفاوت اصلی این رویه‌ها را می‌توان بدین صورت بیان نمود:

- در استاندارد اسکی توسعه‌یافته برای ذخیره‌سازی هر نویسه به یک بایت نیاز است. لذا در این حالت نویسه و بایت را می‌توان معادل یکدیگر در نظر گرفت. اما در استاندارد یونیکد چنین تناظری وجود ندارد.
- استاندارد اسکی فقط شامل نویسه‌های لاتین و برخی نویسه‌های خاص دیگر می‌باشد، اما یونیکد سعی دارد نویسه‌های تمامی زبان‌های دنیا را پشتیبانی نماید.

از آن‌جا که تحقیقات اولیه‌ی تخصیص نویسنده بر روی متون انگلیسی بود، لذا در ابتدا استاندارد اسکی مناسب به نظر می‌رسید و از انگرام‌های در سطح بایت جهت حل مسائل تخصیص نویسنده استفاده می‌شد؛ اما پس از گسترش این حوزه به سایر زبان‌ها، دیگر استاندارد اسکی توسعه‌یافته انتخاب مناسبی به حساب نمی‌آید، زیرا این استاندارد از نویسه‌های سایر زبان‌ها پشتیبانی نمی‌کند. در نتیجه به‌تدریج استفاده از استاندارد اسکی توسعه‌یافته کاهش یافته و یونیکد جایگزین آن گردید.

امروزه معمولاً بحثی در مورد بایت و نویسه صورت نمی‌پذیرد و تقریباً همه‌ی پژوهشگران متن را به صورت رشته‌ای از نویسه‌ها در نظر گرفته و از استاندارد یونیکد استفاده می‌نمایند. به همین جهت در تمامی آزمایش‌های این پایان‌نامه کدگذاری یونیکد مورد استفاده قرار گرفته و دیگر بحثی در مورد استفاده از نویسه یا بایت صورت نمی‌پذیرد.

۳-۲-۲ تعریف انگرام و نحوه‌ی استخراج آن

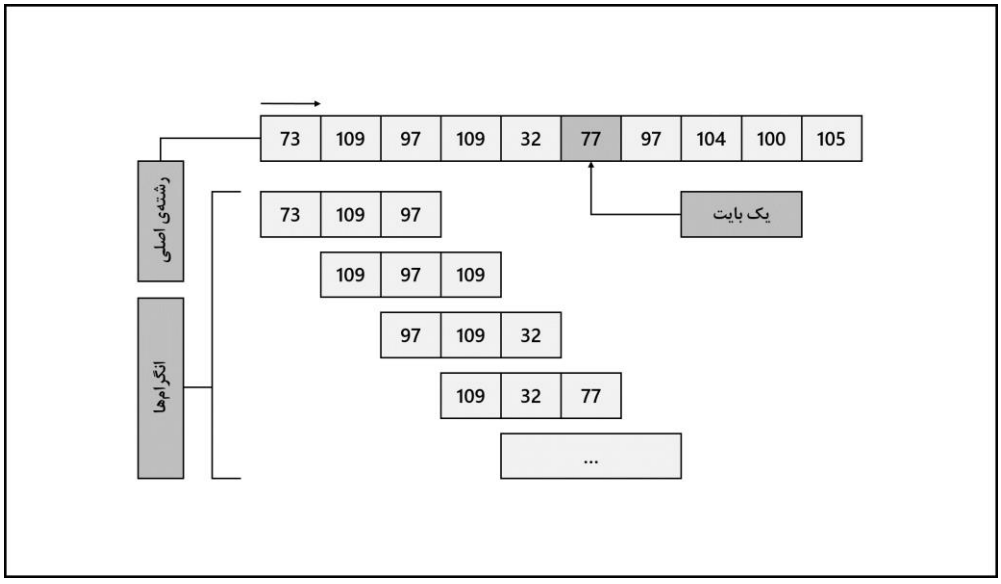
همانطور که در مقدمه‌ی فصل بیان شد، اولین خصیصه‌ی روش‌های NDP آنست که از ویژگی‌هایی مبتنی بر انگرام استفاده می‌نمایند. لذا قبل از تشریح این روش‌ها، باید ابتدا به تعریف انگرام و شرح نحوه‌ی استخراج آن بپردازیم.

الف) تعریف انگرام

در حوزه‌ی تخصیص نویسنده، انگرام در واقع دنباله‌ای پیوسته شامل n جز بوده که از یک متن استخراج می‌گردد. با توجه به نحوه‌ی نگرش به اجزای متن، انگرام در سه سطح قابل تعریف است:

۱) انگرام در سطح بایت[1]:

در این حالت متن به عنوان دنباله‌ای از بایت‌ها در نظر گرفته می‌شود. در نتیجه یک انگرام در سطح بایت عبارتست از دنباله‌ای بهم پیوسته شامل n بایت پشت سر هم. مثالی از این نوع انگرام‌ها به ازای n=3 در شکل ۳-۱ نمایش داده شده است.



شکل ۳-۱: مثالی از انگرام در سطح بایت به ازای n=3

۲) انگرام در سطح نویسه[2]:

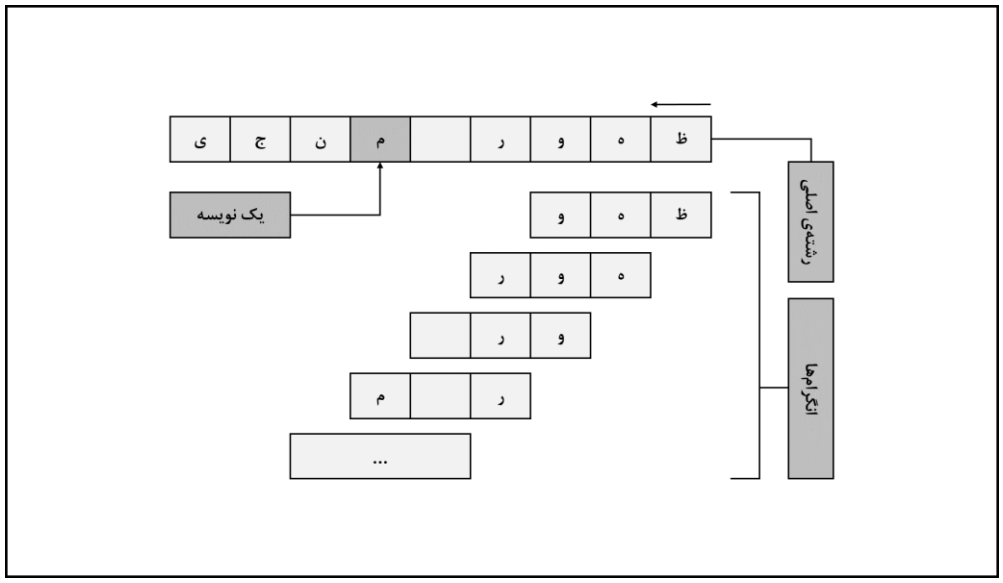
در این حالت متن به عنوان دنباله‌ای از نویسه‌ها در نظر گرفته شده و لذا یک انگرام در سطح نویسه این‌گونه تعریف می‌گردد: دنباله‌ای بهم پیوسته شامل n نویسه‌ی پشت سر هم. شکل ۳-۲ نمایانگر مثالی ساده از این نوع انگرام‌هاست.

۳) انگرام در سطح کلمه[3]:

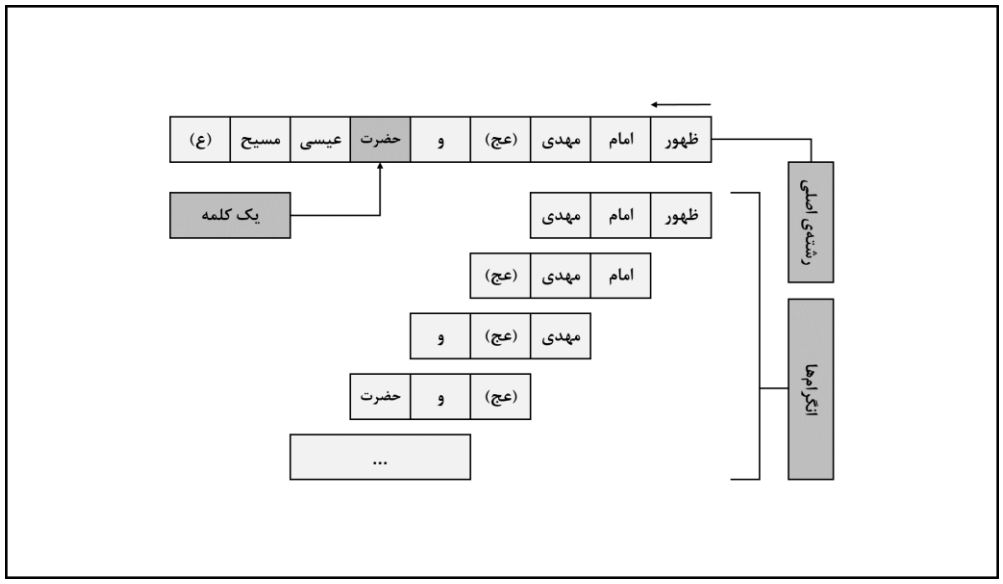
در این حالت متن به عنوان دنباله‌ای از کلمه‌ها فرض می‌شود. به همین جهت یک انگرام در سطح کلمه را به صورت دنباله‌ای بهم پیوسته شامل n کلمه‌ی پشت سر هم تعریف می‌نمایند. البته نحوه‌ی تعریف کلمه در این حالت محل بحث است، زیرا در زبان‌های مختلف، تعاریف متفاوتی از «کلمه[4]» ارائه می‌شود. یکی از رایج‌ترین این تعاریف (که در حوزه‌ی تخصیص نویسنده به طور گسترده مورد

[1] Byte-Level N-gram
[2] Character-Level N-gram
[3] Word-Level N-gram
[4] Word

استفاده قرار می‌گیرد) بدین شرح است: «هر رشته‌ی مابین دو نویسه‌ی فاصله (یا همان Space) را یک کلمه فرض کنیم.» نمونه‌ای از انگرام‌های در سطح کلمه به ازای n=3 در شکل ۳-۳ نمایش داده شده است.



شکل ۳-۲: مثالی از انگرام در سطح نویسه به ازای n=3



شکل ۳-۳: مثالی از انگرام در سطح کلمه به ازای n=3

همان‌طور که در مطالب گذشته نیز اشاره شد، روش‌های NDP بر روی ویژگی‌های مبتنی بر انگرام در سطح نویسه متمرکز شده‌اند. لذا در ادامه‌ی این پایان‌نامه هرگاه انگرام بدون ذکر جزئیات به کار رود، منظور انگرام در سطح نویسه است. در ادامه‌ی بخش نحوه‌ی استخراج این نوع از انگرام‌ها را بیش‌تر توضیح می‌دهیم.

ب) فرایند استخراج انگرام‌های یک متن

فرض کنیم یک متن به طول Q نویسه در اختیار داریم و می‌خواهیم انگرام‌های آن را به ازای یک مقدار خاص از n استخراج نماییم. بدین منظور می‌بایست مراحل زیر طی شود:

۱) ابتدا متن را به صورت رشته‌ای از نویسه‌های متوالی در نظر می‌گیریم. در این حالت متن اصلی (که شامل Q نویسه است) را با نماد T و نویسه‌ی k-ام موجود در آن را با نماد $ch(T,k)$ نمایش می‌دهیم. (شکل ۳-۴، الف)

۲) سپس می‌بایست تعداد (n-1) نویسه‌ی فاصله (Space) به انتهای رشته‌ی اصلی افزوده شود. رشته‌ی حاصل را که دارای طولی معادل با $Q+(n-1)$ نویسه است، «متن پَدشده[1]» نامیده و با نماد T_P نمایش می‌دهیم. (شکل ۳-۴، ب)

۳) طبق تعریف انگرام، هر کدام از نویسه‌های رشته‌ی اصلی، یک انگرام مخصوص به خود تولید می‌کنند. لذا می‌بایست به ازای هر نویسه‌ی موجود در متن اصلی، یک انگرام استخراج نماییم. انگرام مخصوص به هر نویسه عبارتست از دنباله‌ای شامل آن نویسه و (n-1) نویسه‌ی بعد از آن. به عبارت دیگر، انگرام متناظر با $ch(T,k)$ عبارتست از دنباله‌ای شامل نویسه‌های $ch(T_P,k)$ تا $ch(T_p,k+n-1)$. این موضوع در شکل ۳-۴ (ج) نمایش داده شده است.

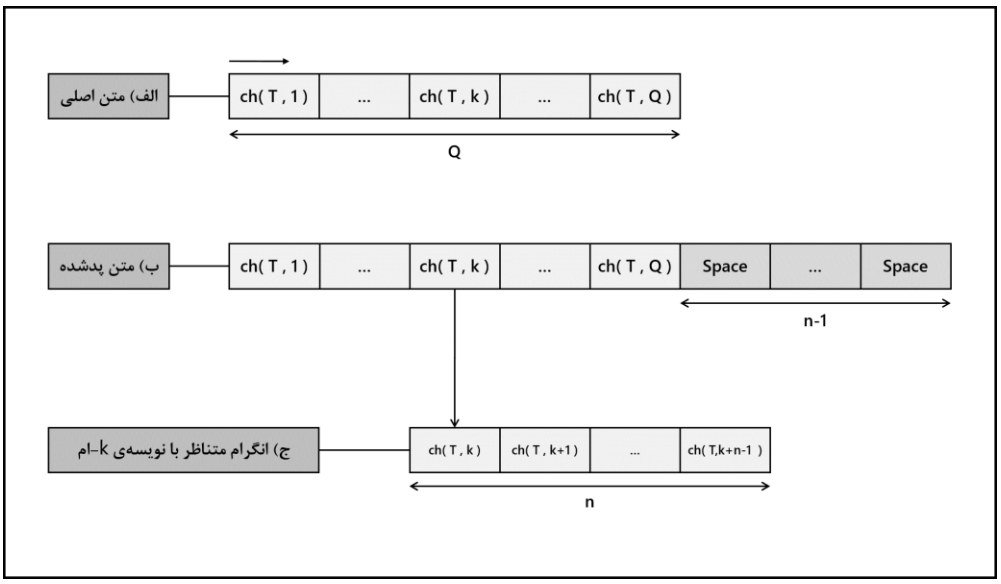
در انتهای این قسمت ذکر دو نکته ضروری است:

- تعداد انگرام‌های تولید شده توسط یک متن (مستقل از مقدار n)، برابرست با تعداد نویسه‌های موجود در آن متن. یعنی یک متن با طول Q نویسه، Q انگرام تولید می‌نماید؛ حال مقدار n هر چقدر که باشد مهم نیست.
- تعداد نویسه‌های موجود در یک انگرام (مستقل از نویسه‌ی متناظر با آن)، برابر n می‌باشد.

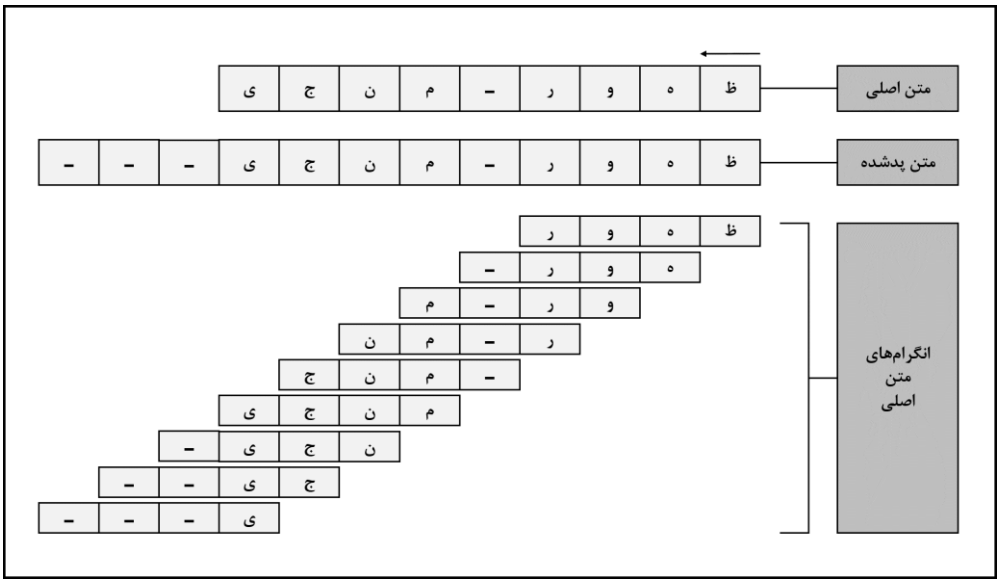
---

[1] Padded Text

به منظور روشن شدن روند استخراج انگرام از یک متن، تمامی انگرام‌های استخراج‌شده از رشته‌ی «ظهور منجی»، به ازای n=3 در شکل ۳-۵ آمده است. جهت وضوح بیش‌تر، نویسه‌های فاصله با استفاده از نماد «/» نمایش داده شده اند.



شکل ۳-۴: روند استخراج انگرام از یک متن



شکل ۳-۵: روند استخراج انگرام‌های رشته‌ی «ظهور منجی»

۳-۲-۳ قالب استاندارد پروفایل

از فصل دوم به یاد داریم که «پروفایل» یک متن عبارتست از بردار، جدول یا هر کمیت عددی دیگری که نمایانگرِ سبک نوشتاری نویسنده‌ی متن باشد. در روش‌های NDP معمولاً پروفایل یک متن دلخواه (مثل T)، با استفاده از دو مجموعه بیان می‌شود:

$$PR_T^{METHOD} = \left\{ G_T^{METHOD}, F_T^{METHOD} \right\} \qquad (۱-۳)$$

که در آن METHOD در واقع بیان‌گر نام روش مورداستفاده جهت استخراج پروفایل متن T است.
هم‌چنین:

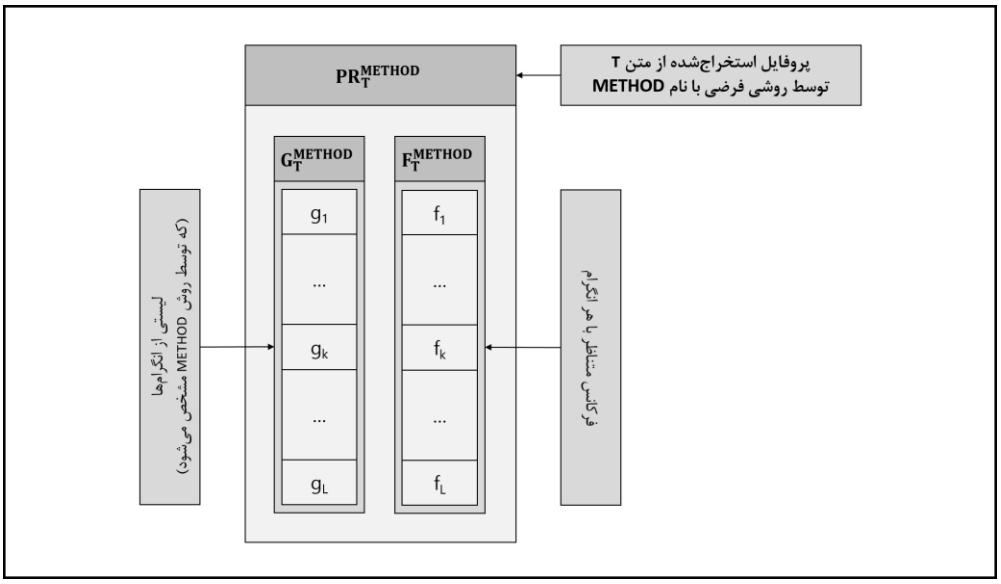
- مجموعه‌ی G_T^{METHOD} در واقع یک لیست از انگرام‌ها می‌باشد. هر روش NDP می‌بایست تعیین کند که انگرام‌های موجود در این لیست چگونه انتخاب می‌شوند.
- مجموعه‌ی F_T^{METHOD} در واقع لیستی از فرکانس‌های متناظر با هر کدام از انگرام‌های موجود در مجموعه‌ی G_T^{METHOD} می‌باشد. هر روش NDP می‌بایست نحوه‌ی محاسبه‌ی این فرکانس‌ها را نیز شرح دهد. معمولاً فرکانس متناظر با هر انگرام، بر مبنای تعداد تکرار آن انگرام در متن T محاسبه می‌شود.

لازم به ذکر است که در ادامه‌ی این فصل، هنگام معرفی روند استخراج پروفایل در هر روش NDP، نحوه‌ی تشکیل مجموعه‌ی G_T^{METHOD} و نحوه‌ی محاسبه‌ی فرکانس متناظر با هر انگرام موجود در آن را شرح خواهیم داد.

در مجموع پروفایل یک متن را می‌توان به صورت جدولی شامل دو ستون نمایش داد، که در آن:

- ستون اول: شامل انگرام‌های موجود در مجموعه‌ی G_T^{METHOD} است.
- ستون دوم: شامل فرکانس‌های متناظر با انگرام‌های موجود در ستون اول می‌باشد.

نمای کلی چنین جدولی در شکل ۳-۶ نمایش داده شده است.



شکل ۳-۶: نمای کلی از یک پروفایل استخراج‌شده توسط یک روش فرضی به نام METHOD

۳-۲-۴ روند کلی روش‌های NDP

هدف اصلی این فصل بررسی مهم‌ترین روش‌های NDP ارائه‌شده توسط پژوهش‌گران حوزه‌ی تخصیص نویسنده می‌باشد. نکته‌ی قابل توجه آنست که تقریباً تمامی روش‌های NDP موجود، از روندی مشابه یکدیگر استفاده می‌کنند. لذا ابتدا در این بخش به شرح این روند کلی پرداخته و سپس در بخش‌های بعد، مهم‌ترین این روش‌ها را بررسی می‌نماییم.

فرض کنیم مجموعه‌متنی متشکل از متون M نویسنده‌ی مختلف در اختیار داریم. می‌خواهیم یک متن جدید را که خارج از این مجموعه‌متن است (متن دیده‌نشده[1])، به یکی از این M نویسنده تخصیص بدهیم. همان‌گونه که در شکل‌های ۳-۷ و ۳-۸ آمده است، روند کلی یک روش NDP شامل پنج گام اصلی است:

گام اول: تشکیل ابرمتن هر نویسنده

ابتدا ابرمتن متناظر با هر نویسنده را تشکیل می‌دهیم. از فصل دوم به خاطر داریم که ابرمتن یک نویسنده از کنارهم قرار دادن تمامی متون آن نویسنده حاصل می‌شود. در پایان این مرحله، به ازای هر نویسنده یک

[1] Unseen Text

ابرمتن متناظر با آن در اختیار داریم. ابرمتن متناظر با نویسنده‌ی j-ام را با نماد $ST(A_j)$ نمایش می‌دهیم. این گام به صورت نمادین در شکل ۳-۷ (الف) نمایش داده شده است.

گام دوم: استخراج پروفایل هر نویسنده

در این گام پروفایل متناظر با هر یک ابرمتن‌های تشکیل‌شده در گام نخست را استخراج می‌نماییم. لازم بذکرست که این استخراج می‌بایست به صورت پویا انجام شده و با استفاده از ویژگی‌های مبتنی بر انگرام باشد. سؤالی که در این گام پیش می‌آید آنست که «چگونه می‌توان پروفایل یک ابرمتن را استخراج نمود؟» پس اولین سؤالی که یک روش NDP باید به آن پاسخ دهد نحوه‌ی استخراج یک ابرمتن می‌باشد. در پایان این مرحله، به ازای هر نویسنده یک پروفایل متناظر با آن در اختیار داریم. پروفایل متناظر با نویسنده‌ی j-ام را که توسط یک روش فرضی با نام METHOD استخراج شده است را با نماد $PR_{A_j}^{METHOD}$ نمایش می‌دهیم (شکل ۳-۷ ب).

گام سوم: تشکیل پروفایل متن دیده‌نشده

در گام سوم پروفایل متن دیده‌نشده استخراج می‌شود. این استخراج نیز می‌بایست به صورت پویا انجام شده و با استفاده از ویژگی‌های مبتنی بر انگرام باشد. دومین سؤالی که یک روش NDP باید به آن پاسخ دهد، در این گام ایجاد می‌شود: «چگونه می‌توان پروفایل متن دیده‌نشده را استخراج نمود؟» در حالت کلی الزامی بر یکسان بودن شیوه‌ی استخراج پروفایل از ابرمتن‌ها و استخراج پروفایل از متن دیده‌نشده وجود ندارد، اما در اکثر روش‌های NDP این دو استخراج به یک صورت انجام می‌پذیرند. در پایان این مرحله، پروفایل متن دیده‌نشده را (که با نماد PR_U^{METHOD} نمایش داده می‌شود) در اختیار خواهیم داشت (شکل ۳-۷ ج).

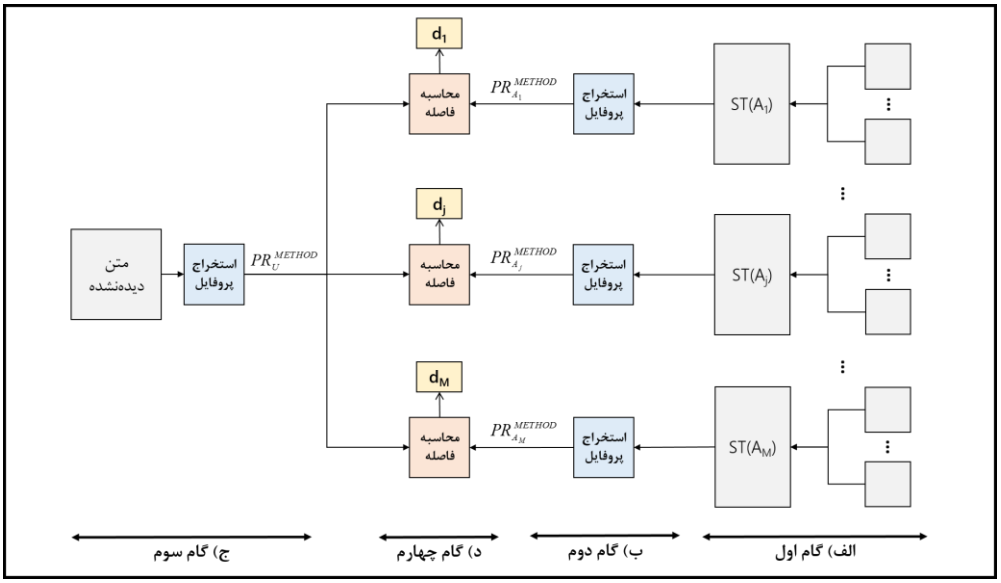
گام چهارم: محاسبه‌ی فاصله‌ها

در این گام فاصله‌ی بین پروفایل متن دیده‌نشده با پروفایل‌های تک تک نویسنده‌ها محاسبه می‌شود (شکل ۳-۷ د). فاصله‌ی بین پروفایل نویسنده‌ی j-ام و پروفایل متن دیده‌نشده را با نماد $Dist^{METHOD}(PR_U^{METHOD}, PR_{A_j}^{METHOD})$ نمایش می‌دهیم. در پایان این مرحله، به ازای هر نویسنده یک

مقدار عددی در اختیار داریم که بیان‌گر فاصله‌ی پروفایل متن دیده‌نشده تا پروفایل آن نویسنده است. به عبارت دیگر حاصل این گام، برداری‌ست شامل فواصل محاسبه شده:

$$D(j) = Dist^{METHOD}(PR_U^{METHOD}, PR_{A_j}^{METHOD}) \quad , \quad j = 1, \ldots, M \qquad (۲-۳)$$

پس حاصل گام چهارم عبارتست از برداری به نام D که شامل M المان است.



شکل ۳-۷: روند کلی روش‌های NDP - گام‌های اول تا چهارم

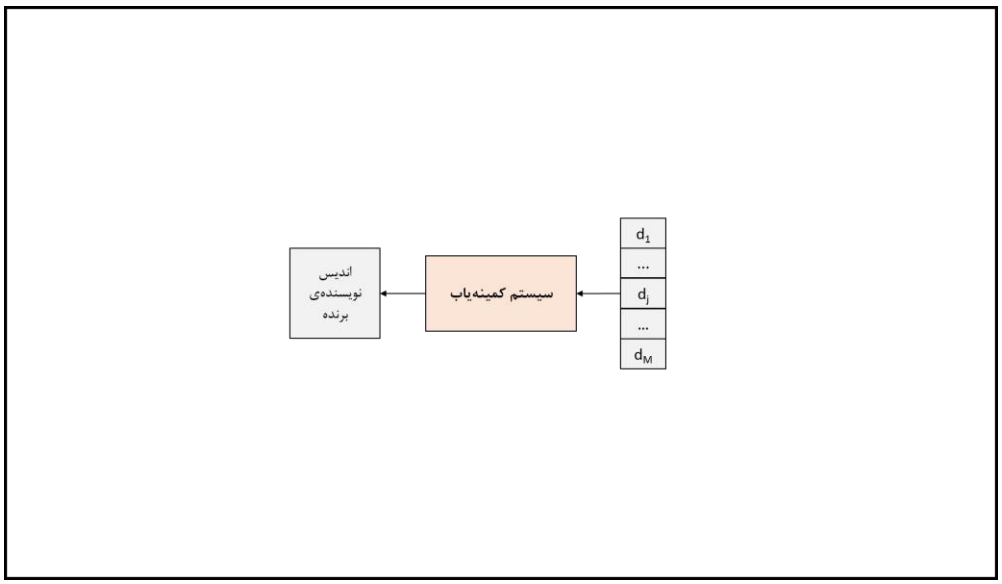
گام پنجم: تخصیص نویسنده

در گام نهایی نویسنده‌ای را که پروفایل وی دارای کمترین فاصله با پروفایل متن دیده‌نشده است، به عنوان «نویسنده‌ی برنده[1]» انتخاب کرده و متن دیده‌نشده را به وی تخصیص می‌دهیم. از آن‌جا که فاصله‌ی میان پروفایل تک تک نویسنده‌ها و پروفایل متن دیده‌نشده در گام قبلی محاسبه شده و در بردار D ذخیره شده‌اند، در این گام فقط می‌بایست اندیس المانی از بردار D را که دارای مقدار کمینه است بدست آورد. به عبارت دیگر اگر اندیس نویسنده‌ی برنده را با J^* نمایش دهیم، داریم:

[1] Winner Writer

$$J^* = \arg\min_j \left[D(j) \right] \quad , \quad j = 1, ..., M \qquad (۳-۳)$$

یعنی اندیس المانی از بردار D را که دارای کمترین مقدار است، به عنوان اندیس نویسنده‌ی برنده در نظر می‌گیریم (شکل ۳-۸).



شکل ۳-۸: روند کلی روش‌های NDP - گام پنجم

سؤالی که ممکن است پس از مطالعه‌ی گام پنجم در ذهن خواننده ایجاد شده باشد، آنست که: اگر چندین المان از D همزمان حاوی مقدار کمینه باشند، نویسنده‌ی برنده چگونه انتخاب می‌شود؟

به عنوان مثال حالتی را در نظر بگیرید که بردار D به صورت زیر باشد:

$$D = [\, 0.75, 0.62, 0.6, 0.85, 0.73, 0.6, 0.94 \,]$$

در این حالت نویسنده‌های شماری ۳ و ۶ همزمان دارای کمترین فاصله (0.6) با متن دیده‌نشده می‌باشند. حال سؤال این‌جاست که کدام نویسنده برنده است؟ نویسنده‌ی شماره ۳ یا نویسنده‌ی شماره ۶؟

جهت پاسخ دادن به این سؤال، چهار روند را می‌توان در نظر گرفت:

- برای تشخیص نویسنده‌ی برنده، سیستم دیگری را این بار فقط بر روی نویسنده‌ی ۳ و ۶ آزمایش نماییم.

- هر دو نویسنده را به عنوان برنده اعلام نماییم.
- هیچ نویسنده‌ای را به عنوان برنده اعلام ننماییم. (حالت Reject)
- یکی از دو نویسنده را به صورت تصادفی به عنوان برنده انتخاب نماییم.

از آن‌جا که در این پایان‌نامه مسائل را به صورت مجموعه-بسته در نظر گرفته و حل می‌نماییم، لذا می‌بایست در هر صورت یک نویسنده را به عنوان برنده اعلام نماییم. در نتیجه راهکارهای دوم و سوم مناسب نمی‌باشند. از طرفی روش‌های NDP معمولاً هیچ سیستم اضافی برای چنین حالتی (که چند نویسنده به طور همزمان دارای مقدار کمینه فاصله می‌باشند) ارائه نمی‌دهند، پس تنها راهکار پیش‌رو همان انتخاب نویسنده‌ی برنده به صورت تصادفی از بین نویسنده‌ی ۳ و ۶ است.

به طور خلاصه می‌توان گفت: اگر چند نویسنده به طور همزمان دارای کمترین فاصله با متن دیده‌نشده باشند، می‌بایست یکی از آن‌ها را به صورت تصادفی به عنوان برنده انتخاب نماییم. (اما بهترست هنگام محاسبه‌ی دقت عملکرد سیستم طراحی‌شده، چنین حالاتی را به عنوان حالات پاسخ صحیح در نظر نگیریم. این نکته در فصل پنجم بیش‌تر تشریح خواهد شد.)

۳-۲-۵ دو سیستم اساسی مورداستفاده در یک روش NDP

با توجه به پنج گام ذکرشده در قسمت قبل (۳-۲-۴) می‌توان چنین نتیجه گرفت که در حالت کلی، یک روش NDP می‌بایست به سه سؤال اساسی پاسخ دهد:

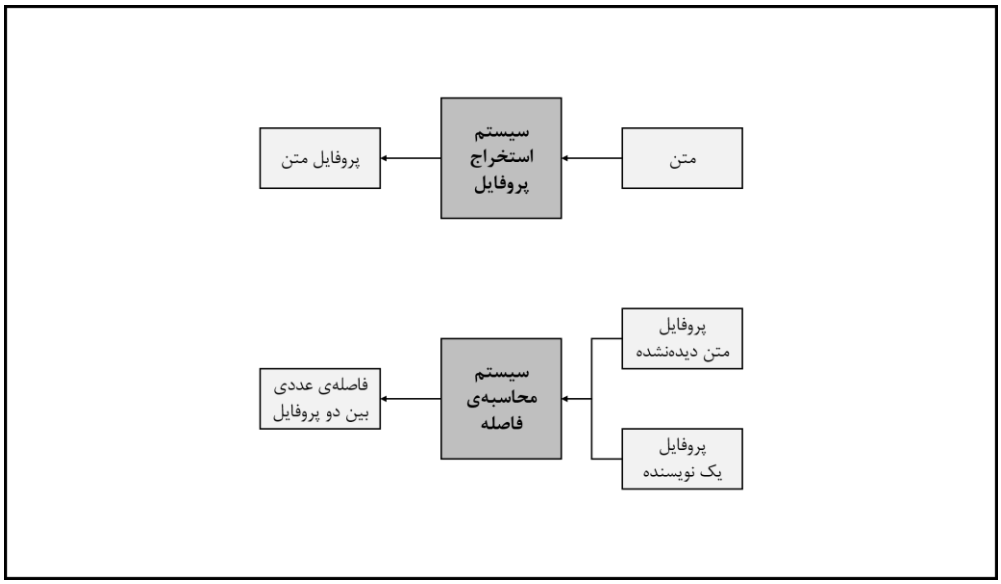
- پروفایل یک نویسنده به چه صورت از ابرمتن وی استخراج می‌شود؟
- نحوه‌ی استخراج پروفایل متن دیده‌نشده چگونه است؟
- فاصله‌ی میان پروفایل متن دیده‌نشده و پروفایل یک نویسنده چگونه محاسبه می‌شود؟

همانطور که گفته شد اکثر روش‌های NDP، روند یکسانی جهت استخراج پروفایل یک نویسنده و استخراج پروفایل متن دیده‌نشده ارائه می‌دهند. به عبارت دیگر معمولاً یک روش NDP، سیستمی جهت استخراج پروفایل یک متن دلخواه ارائه می‌دهد. حال اگر متن دیده‌نشده را به عنوان ورودی به این سیستم اعمال کنیم، خروجی سیستم همان پروفایل متن دیده‌نشده است؛ اما اگر ابرمتن یک نویسنده را در ورودی سیستم قرار دهیم، خروجی در واقع پروفایل متناظر با آن نویسنده خواهد بود.

با توجه به این توضیحات می‌توان گفت که یک روش NDP می‌بایست نحوه‌ی عملکرد دو سیستم را شرح دهد:

۱) **سیستم استخراج پروفایل:** سیستمی که بتواند پروفایل یک متن دلخواه را استخراج نماید. این متن می‌تواند ابرمتن مربوط به یک نویسنده باشد، و یا متن دیده‌نشده.
۲) **سیستم محاسبه‌ی فاصله:** سیستمی که بتواند فاصله‌ی میان دو پروفایل را محاسبه نماید. یکی از این پروفایل‌ها مربوط به متن دیده‌نشده و دیگری متناظر با ابرمتن یک نویسنده است.

پس در مجموع برای مطالعه‌ی یک روش NDP می‌بایست نحوه‌ی طراحی دو سیستم در آن روش را بررسی نماییم: «سیستم استخراج پروفایل» و «سیستم محاسبه‌ی فاصله». یک نمای کلی از این دو سیستم در شکل ۳-۹ نمایش داده شده است. در ادامه‌ی این پایان‌نامه هنگام بررسی یک روش NDP فقط نحوه‌ی طراحی دو سیستم فوق را شرح می‌دهیم.



شکل ۳-۹: دو سیستم اصلی مورداستفاده در یک روش NDP

۳-۳ مهم‌ترین روش‌های NDP موجود

در تحقیقات این پایان‌نامه هشت مورد از مهم‌ترین روش‌های NDP موجود مورد مطالعه و بررسی قرار گرفته است. این روش‌ها به ترتیب سال انتشار در جدول ۳-۱ گردآوری شده‌اند.

جدول ۳-۱: هشت مورد از مهم‌ترین روش‌های NDP

ردیف	سال انتشار	نام روش	ارائه‌دهندگان	مرجع
۱	۲۰۰۳	انگرام‌های متداول[1] CNG	Vlado Keselj Fuchun Peng Nick Cercone Calvin Thomas	۲
۲	۲۰۰۷	پروفایل نویسنده‌ی کد مبدأ[2] SCAP یا اشتراک پروفایل‌های ساده‌شده[3] SPI	Georgia Frantzeskou Efstathios Stamatatos Stefanos Gritzalis	۲۹
۳ و ۴	۲۰۰۷	انگرام‌های متداول با دو معیار D1 و D2 CNG-D1 و CNG-D2	Efstathios Stamatatos	۲۶
۵	۲۰۱۱	انگرام‌های متداول با اشتراک وزن‌دار پروفایل‌ها[4] CNG-WPI	Hugo Jair Escalante Manuel Montes-y-Gomez Thamar Solorio	۲۷
۶	۲۰۱۱	پروفایل محلی باز-مرکزی‌شده[5] RLP	Robert Layton Paul Watters Richard Dazeley	۲۸
۷	۲۰۱۲	پروفایل محلی باز-مرکزی‌شده با فرکانس معکوس نویسنده[6] RLP-IAF	Robert Layton Stephen McCombie Paul Watters	۴۵
۸	۲۰۱۴	اشتراک پروفایل‌های ساده‌شده‌ی حذف‌شده[7] O-SPI	Matthew F. Tennyson Francisco J. Mitropoulos	۲۹

[1] Common N-Grams (CNG)
[2] Source-Code Author Profile (SCAP)
[3] Simplified Profile Intersection (SPI)
[4] Common N-Grams, with Weighted Profile Intersection (CNG-WPI)
[5] Re-centered Local Profile (RLP)
[6] Inverse Author Frequency (IAF)
[7] Omitted Simplified Profile Intersection (O-SPI)

می‌بایست توجه داشت که روش CNG در واقع پایه و اساس تمامی هفت روش دیگر می‌باشد (جالب آن‌که این روش خود برمبنای روش قدیمی بنت {۸۶} ارائه گردیده است). به همین جهت در ادامه‌ی فصل روش CNG را به طور کامل و با ذکر جزئیات شرح داده، اما سایر روش‌ها را (که عموماً از ایجاد برخی اصلاحات بر روی روش CNG ارائه شده‌اند)، مختصراً توضیح می‌دهیم.

هنگام معرفی یک روش، تلاش می‌کنیم تا موارد زیر را ارائه دهیم:

- ابتدا فلسفه‌ی ارائه‌ی روش را بیان می‌کنیم.
- سپس نحوه طراحی سیستم استخراج پروفایل در روش مربوطه را شرح می دهیم.
- بعد از آن سیستم محاسبه‌ی فاصله‌ی معرفی‌شده توسط آن روش را معرفی می‌کنیم.
- و درنهایت نتایج ارائه‌شده در مقاله‌ی مرجع هر روش (مقاله‌ای که روش موردنظر برای اولین‌بار در آن معرفی شده است) را مختصراً شرح می‌دهیم.

۳-۴ روش اول: انگرام‌های متداول (CNG)

نسخه‌ی اولیه‌ی این روش در سال ۲۰۰۳ میلادی توسط کسلج و همکاران ارائه شد {۲}. اما در سال‌های بعد نسخه‌های تکامل‌یافته‌ی آن توسط سایر پژوهش‌گران ارائه گردید. هرچند هم‌چنان زمانی که یک مجموعه‌متن با متون کافی در اختیار باشد، معمولاً روش CNG بهتر از نسخه‌های تکامل‌یافته‌ی آن عمل می‌کند.

۳-۴-۱ فلسفه‌ی ارائه‌ی روش

ایده‌ی اصلی روش CNG آنست که «پرتکرارترین انگرام‌های یک متن، در واقع مهم‌ترین انگرام‌های آن می‌باشند.» به عبارت دیگر هنگام استخراج پروفایل یک متن می‌بایست انگرام‌هایی که بیش‌تر از سایرین تکرار شده‌اند را مورد توجه قرار دهیم. لازم به ذکرست که استفاده از تعداد تکرار انگرام‌ها جهت حل مسائل تخصیص نویسنده اولین‌بار توسط بنت {۸۶} ارائه گردیده است.

از بخش ۳-۲ به یاد داریم که یک روش NDP می‌بایست نحوه‌ی طراحی دو سیستم را بیان نماید: سیستم استخراج پروفایل و سیستم محاسبه‌ی فاصله. در ادامه روند طراحی این سیستم‌ها در روش CNG را شرح می‌دهیم.

۳-۴-۲ سیستم استخراج پروفایل

در روش CNG پروفایل یک متن بدین صورت تعریف می‌گردد: «پروفایل یک متن عبارتست از مجموعه‌ای شامل L جفت از پرتکرارترین انگرام‌های آن متن، به همراه فرکانس نرمال‌شده‌ی آن‌ها.» از آن‌جا که در تعریف پروفایل، پرتکرارترین (یا در واقع متداول‌ترین) انگرام‌ها مبنا قرار گرفته‌اند، این مدل را اصطلاحاً «انگرام‌های متداول» می‌نامند. احتمالاً در ذهن خواننده این سؤال ایجاد شده است که منظور از «فرکانس نرمال‌شده‌ی یک انگرام» چیست؟ این سؤال در ادامه پاسخ داده خواهد شد.

در این بخش می‌خواهیم فرایندی گام به گام جهت استخراج پروفایل یک متن دلخواه (T) با استفاده از روش CNG ارائه دهیم. به منظور روشن‌تر شدن این فرایند، تمامی گام‌های آن بر روی یک متن کوتاه پیاده‌سازی شده و در شکل‌های ۳-۹ تا ۳-۱۱ نمایش داده شده‌اند.

گام صفرم: تعیین پارامترها

پیش از آغاز فرایند اصلی استخراج پروفایل یک متن، می‌بایست ابتدا پارامترهای پروفایل را تعیین نماییم. دو پارامتر اساسی برای هر پروفایل عبارتند از:

- طول انگرام: بیانگر تعداد نویسه‌ی موجود در هر انگرام است. این پارامتر را با N نمایش می‌دهیم.
- طول پروفایل: بیانگر حداکثر تعداد انگرام موجود در پروفایل است. این پارامتر را با L نمایش می‌دهیم. (مفهوم طول پروفایل در گام سوم بیان می‌شود.)

از آن‌جا پیاده‌سازی این گام مستلزم عملیات خاصی نبوده و به نوعی مقداردهی اولیه‌ی پارامترها به حساب می‌آیند، آن را به عنوان گام «صفرم» در نظر می‌گیریم تا در شماره‌گذاری گام‌های اصلی ظاهر نشود.

گام اول: استخراج لیست انگرام‌ها

در این گام با توجه به طول انتخاب شده برای هر انگرام (یعنی همان پارامتر N)، تمامی انگرام‌های متن T را استخراج می‌نماییم. بدین منظور می‌بایست مراحل بیان‌شده در بخش ۳-۲-۲ طی شود. اگر فرض کنیم متن T دارای Q نویسه است، در پایان این گام، لیستی شامل Q انگرام استخراج‌شده خواهیم داشت که به آن «لیست انگرام‌های خام» می‌گوییم. شکل ۳-۱۰ مثالی از استخراج این لیست را نمایش می‌دهد.

شکل ۳-۱۰: استخراج لیست انگرام‌های خام از یک متن نمونه (N=3)

گام دوم: سرشماری انگرام‌ها

در این گام می‌بایست دو مرحله‌ی کوتاه طی شود:

- ابتدا لیست «انگرام‌های ویژه» تشکیل می‌شود. این لیست در واقع همان لیست انگرام‌های خام است، با این تفاوت که موارد تکراری از آن حذف شده‌اند. تعداد انگرام‌های ویژه را (که معمولاً به مراتب کمتر از تعداد انگرام‌های موجود در لیست خام است) با نماد Q’ نمایش می‌دهیم.
- سپس می‌بایست به ازای هر کدام از انگرام‌های موجود در لیست ویژه، تعداد تکرار آن انگرام در لیست خام را محاسبه و ثبت نماییم.

پس از اتمام این گام جدولی با Q’ سطر خواهیم داشت که در ستون اول آن، انگرام‌های ویژه و در ستون دوم آن، تعداد تکرار هر انگرام ویژه ذخیره شده است. این جدول را جدول انگرام‌های ویژه می‌نامیم. مثالی از لیست انگرام‌های ویژه و جدول انگرام‌های ویژه به ترتیب در شکل‌های ۳-۱۱ و ۳-۱۲ (الف) آمده است.

شکل ۳-۱۱: استخراج لیست انگرام‌های ویژه از یک متن نمونه (N=3)

گام سوم: جداسازی انگرام‌های دارای بالاترین تعداد تکرار

گام سوم نیز خود به دو مرحله‌ی کوچک‌تر تقسیم می‌شود:

- ابتدا می‌بایست جدول انگرام‌های ویژه را بر اساس تعداد تکرار به صورت نزولی مرتب نماییم (شکل ۳-۱۲ ب)، طوری که انگرام‌های دارای تعداد تکرار بیش‌تر (انگرام‌های مهم‌تر) در ردیف‌های ابتدایی جدول قرار بگیرند. پس از این مرتب‌سازی، انگرام ویژه‌ی i-ام را با نماد $ngram(T,i)$ و تعداد تکرار متناظر با آن را با نماد $RC(T,i)$ نمایش می‌دهیم.
- سپس L سطر اول جدول را حفظ و مابقی را حذف می‌نماییم. جدول کوتاه‌شده‌ی حاصل، نسخه‌ی اولیه از پروفایل متن T است که در گام بعد تغییرات کوچکی بر روی آن ایجاد می‌شود (شکل ۳-۱۲ ج). در این مرحله مفهوم طول پروفایل روشن می‌گردد.

گام چهارم: محاسبه‌ی فرکانس نرمال‌شده

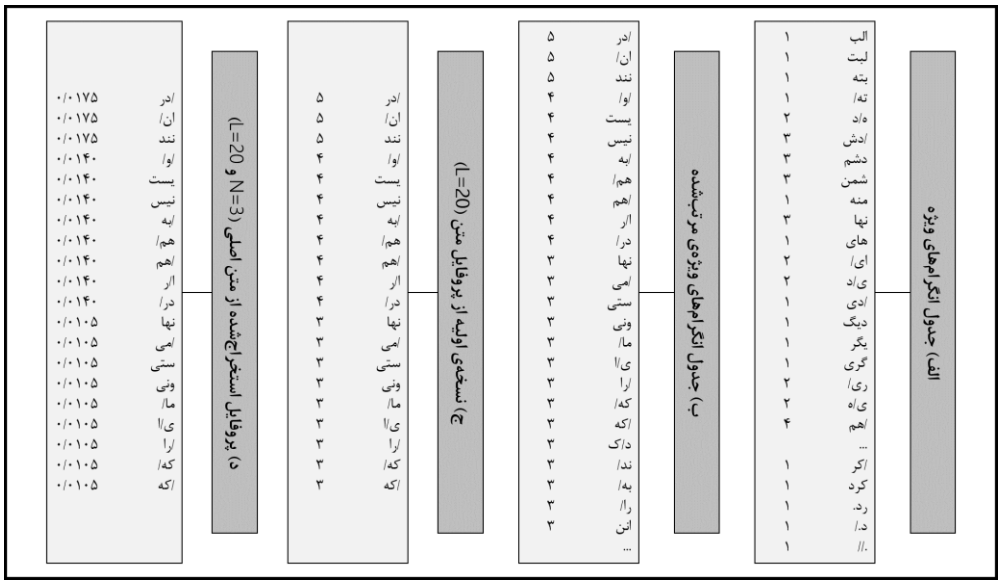
در گام نهایی، فرکانس نرمال‌شده‌ی تمامی انگرام‌های موجود در جدول حاصل از گام قبل را محاسبه می‌نماییم. فرکانس نرمال‌شده‌ی انگرام ویژه‌ی $ngram(T,i)$ (که آن را با $f(T,i)$ نمایش می‌دهیم) عبارتست از تعداد تکرار آن انگرام ویژه تقسیم بر تعداد کل انگرام‌های متن. یعنی:

$$f(T,i)=\frac{RC(T,i)}{Q} \quad , \qquad i=1,\ldots,\mathrm{L} \tag{۴-۳}$$

به عنوان مثال اگر در متنی با طول ۱۰۰ نویسه، یک انگرام ویژه ۷ بار تکرار شده باشد، آن‌گاه فرکانس نرمال‌شده‌ی آن انگرام خاص برابرست با:

$$\frac{7}{100} = 0.07$$

جدول حاصل پس از محاسبه‌ی فرکانس نرمال‌شده‌ی هر انگرام، همان پروفایل متن اصلی خواهد بود (شکل ۳-۱۲ د).



شکل ۳-۱۲: استخراج پروفایل از یک متن نمونه (N=3 و L=20)

با جمع‌بندی گام‌های فوق می‌توان گفت در روش CNG، پروفایل یک متن دلخواه (T) در واقع جدولی شامل L سطر و دو ستون می‌باشد که:

- در ستون اول آن L تا از پرتکرارترین انگرام‌های ویژه‌ی متن قرار دارند.
- در ستون دوم فرکانس نرمال‌شده‌ی انگرام‌های ستون اول ذخیره شده‌اند.

برای رسیدن به قابل استاندار پروفایل (که در بخش ۳-۲-۲ معرفی شد)، می‌بایست هر کدام از این ستون‌ها را به صورت یک مجموعه فرض نماییم، در این صورت پروفایلی که مدل CNG برای یک متن دلخواه (T) ارائه می‌دهد شامل دو مجموعه‌ی G و F خواهد بود:

$$PR_T^{CNG} = \left\{ G_T^{CNG}, F_T^{CNG} \right\} \qquad (۵-۳)$$

که این مجموعه‌ها عبارتند از:

- مجموعه‌ی G_T^{CNG} شامل L تا از پرتکرارترین انگرام‌های ویژه‌ی استخراج‌شده از متن T می‌باشد.
- مجموعه‌ی F_T^{CNG} شامل فرکانس‌های نرمال‌شده‌ی متناظر با هر انگرام موجود در مجموعه‌ی G_T^{CNG} می‌باشد.

چنین پروفایلی را «پروفایل کلاسیک[1]» نامیده و با نماد CP نمایش می‌دهیم. پس به طور خلاصه می‌توان گفت: روش CNG از پروفایل CP استفاده می‌نماید.

چند نکته

در پایان ذکر سه نکته، ضروری به نظر می‌رسد:

نکته‌ی اول: حالت خاصِ Q'<L

سؤالی که ممکن است در ذهن ایجاد شده باشد، آنست که اگر حالتی رخ دهد که در آن Q'<L باشد، آن‌گاه استخراج پروفایل چگونه صورت می‌گیرد؟ به عبارت دیگر فرض کنیم پارامتر انتخابی برای طول پروفایل L=1000 باشد، اما متن T در کل دارای ۵۰۰ انگرام ویژه باشد (یعنی Q'=500)، آن‌گاه پروفایل متن T چگونه تعریف می‌شود؟ پاسخ آنست که در این حالت تمامی انگرام‌های ویژه را به همراه فرکانس‌های نرمال‌شده‌ی آن‌ها در پروفایل قرار می‌دهیم و لذا پروفایل در این حالت Q' عضو دارد، نه L عضو. به همین جهت پارامتر L را به صورت «حداکثرِ تعداد انگرام موجود در پروفایل» تعریف کردیم، چون ممکن است در برخی حالات تعداد انگرام‌های موجود در پروفایل کمتر از L شود (همانند حالتی که Q'<L باشد)؛ اما تعداد انگرام‌های موجود در یک پروفایل هیچ‌گاه از مقدار L بالاتر نخواهد رفت.

نکته‌ی دوم: پروفایل کامل

حالتی را فرض کنیم که بخواهیم پروفایل استخراج‌شده شامل تمامی انگرام‌های ویژه بوده و هیچ انگرامی را حذف ننماید. در این حالت اصطلاحاً می‌گوییم «پروفایل را با طول بی‌نهایت استخراج می‌نماییم، یعنی:

---

[1] Classic Profile (CP)

L=infinite». این نوع پروفایل را معمولاً «پروفایل با طول بی‌نهایت[1]» یا «پروفایل کامل[2]» می‌نامند. تمامی انگرام‌های ویژه‌ی یک متن در پروفایلِ کامل حضور دارند و لذا تعداد اعضای آن برابر با 'Q می‌باشد.

نکته‌ی سوم: تعریف فرکانس انگرام در یک مدل

از بخش ۳-۲-۳ به یاد داریم که هر روش NDP روندی خاص برای محاسبه‌ی پارامتر $f_T^{METHOD}(g)$ ارائه می‌دهد و قرار بر آن شد که هنگام معرفی هر روش، نحوه‌ی محاسبه‌ی این پارامتر در روش مذکور را نیز شرح دهیم. در روش CNG فرکانس یک انگرام دلخواه (g) در یک متن دلخواه (T) بدین صورت محاسبه می‌شود:

$$f_T^{CNG}(g)=\begin{cases} F_T^{CNG}[k]\ ,\ k=\text{index of } g \text{ in} G_T^{CNG} & g\in G_T^{CNG} \\ 0 & g\notin G_T^{CNG} \end{cases} \qquad (۶-۳)$$

که در آن $F_T^{CNG}[k]$ عبارتست از المان k-ام مجموعه‌ی F_T^{CNG}. به عبارت دیگر در مدل CNG هنگام محاسبه‌ی فرکانس انگرامِ g دو حالت ممکن است رخ دهد:

- اگر g عضوی از پروفایل استخراج‌شده از متن T باشد، آن‌گاه ابتدا اندیس متناظر با انگرام g در مجموعه‌ی G_T^{CNG} را یافته (k) و سپس فرکانس را از المان k-ام مجموعه‌ی F_T^{CNG} می‌خوانیم.
- اگر g عضوی از پروفایل استخراج‌شده از متن T نباشد، آن‌گاه فرکانس آن برابر صفر فرض می‌شود.

لازم به ذکرست که مقدار $f_T^{CNG}(g)$ به پارامترهای N و L نیز وابسته است، زیرا این پارامترها در تعیین مجموعه‌های G_T^{CNG} و F_T^{CNG} مؤثر می‌باشند. به عنوان مثال فرض کنیم یک انگرام خاص (مثلاً gx) در جدول انگرام‌های ویژه‌ی مرتب‌شده دارای اندیس ۱۲۰۰ باشد. حال اگر L=1000 انتخاب شود، آن‌گاه انگرام gx در مجموعه‌ی G_T^{CNG} قرار نگرفته و لذا فرکانس آن برابر صفر در نظر گرفته می‌شود. اما اگر L=1500 باشد، این انگرام در مجموعه‌ی G_T^{CNG} جای گرفته و فرکانس آن از مجموعه‌ی F_T^{CNG} خوانده می‌شود.

[1] Infinite-Length Profile
[2] Full-Profile

۳-۴-۳ سیستم محاسبه‌ی فاصله

حال که با نحوه‌ی استخراج پروفایل یک متن در روش CNG آشنا شدیم، به سراغ دومین سیستم اساسی موجود در یک روش NDP می‌رویم: سیستم محاسبه‌ی فاصله. از بخش ۵-۲-۳ به یاد داریم که هدف این سیستم، محاسبه‌ی فاصله‌ی میان پروفایل متن دیده‌نشده و پروفایل یک نویسنده می‌باشد. در ادامه نحوه‌ی محاسبه‌ی این فاصله را شرح می‌دهیم:

فرض کنیم ابرمتن متناظر با نویسنده‌ی موردنظر (A) و متن دیده‌نشده (U) را در اختیار داشته و پروفایل هر یک را توسط روش CNG استخراج نموده‌ایم. حال می‌خواهیم فاصله‌ی میان پروفایل U و پروفایل A (یا بیان ساده‌تر فاصله‌ی بین U و A) را محاسبه نماییم. روش CNG رابطه‌ی زیرا را جهت محاسبه‌ی این فاصله پیشنهاد می‌دهد:

$$Dist^{CNG}(\mathrm{U},A)=\sum_{g\in G}4\left[\frac{f_U^{CNG}(g)-f_A^{CNG}(g)}{f_U^{CNG}(g)+f_A^{CNG}(g)}\right]^2 \qquad (۷-۳)$$

که در آن، مجموعه‌ی G عبارتست از:

$$G=G_U^{CNG}\cup G_A^{CNG} \qquad (۸-۳)$$

این معیار فاصله را «فاصله‌ی نسبی[1]» نامیده و با عنوان RD از آن یاد می‌کنیم. به طور خلاصه می‌توان گفت: روش CNG از معیار فاصله‌ی RD استفاده می‌نماید.

طبق معیار RD، فاصله‌ی میان دو پروفایل، حاصل مجموع فاصله‌ی متناظر با هر عضو مجموعه‌ی G است. به عبارت دیگر هر عضو موجود در مجموعه‌ی G (که با g نمایش داده می‌شود)، مقدار فاصله‌ی مخصوص به خود را به فاصله‌ی کل می‌افزاید. از طرفی می‌توان ضریب ۴ را فاکتور گرفت، لذا رابطه‌ی فوق بدین صورت بازنویسی می‌شود:

$$Dist^{CNG}(U,A)=4\times\sum_{g\in G}d^{CNG}(g) \qquad (۹-۳)$$

که در آن:

[1] Relative Distance (RD)

$$d^{CNG}(g)=\left[\frac{f_U^{CNG}(g)-f_A^{CNG}(g)}{f_U^{CNG}(g)+f_A^{CNG}(g)}\right]^2 \qquad (۱۰-۳)$$

حال به گام‌بندی فرایند پیاده‌سازی سیستم محاسبه‌ی فاصله در مدل CNG می‌پردازیم. به منظور محاسبه‌ی فاصله‌ی میان پروفایل متن دیده‌نشده و پروفایل یک نویسنده با استفاده از مدل CNG، می‌بایست مراحل زیر را طی نماییم:

گام اول: تشکیل مجموعه‌ی G

در نخستین گام می‌بایست مجموعه‌ی G را تشکیل دهیم. همان‌طور که در رابطه‌ی ۳-۸ آمده است، این مجموعه در واقع اجتماع دو مجموعه‌ی G_U^{CNG} و G_A^{CNG} می‌باشد.

گام دوم: محاسبه‌ی فاصله‌ی متناظر با هر عضو G

به ازای هر انگرام موجود در مجموعه‌ی G، می‌بایست یک فاصله بر مبنای معیار RD محاسبه نماییم. بدین‌منظور می‌بایست ابتدا مقداری عددی دو فرکانس $f_U^{CNG}(g)$ و $f_A^{CNG}(g)$ را تعیین کرده (طبق مطالب بیان‌شده در انتهای بخش ۳-۴-۲) و سپس از رابطه‌ی ۳-۱۰ استفاده نماییم.

نکته‌ی جالب‌توجه آنست که اگر انگرام g فقط در یکی از پروفایل‌ها وجود داشته باشد، آن‌گاه فاصله‌ی متناظر با g، مستقل از مقدار عددی فرکانس‌ها برابر ۱ می‌شود. به عنوان مثال فرض کنیم انگرام g فقط در پروفایل A وجود داشته باشد، آن‌گاه:

$$g \notin G_U^{CNG} \Rightarrow f_U^{CNG}(g)=0 \Rightarrow d^{CNG}(g)=\left[\frac{0-f_A^{CNG}(g)}{0+f_A^{CNG}(g)}\right]^2=1 \qquad (۱۱-۳)$$

گام سوم: تجمیع فواصل متناظر هر عضو G

پس از محاسبه‌ی فاصله‌ی متناظر با هر عضو مجموعه‌ی G، طبق رابطه‌ی ۳-۹ می‌بایست تمامی این فواصل را با هم جمع کنیم تا فاصله‌ی کل بین دو پروفایل بدست آید. با توجه به آن‌که در تعیین نویسنده‌ی برنده، نویسنده‌ای انتخاب می‌شود که فاصله‌ی پروفایل آن تا پروفایل U کمینه باشد، لذا می‌توان در محاسبات ضریب ۴ موجود در رابطه‌ی ۳-۹ را حذف نمود، زیرا این ضریب در تک تک فاصله‌های محاسبه‌شده ضرب

می‌شود؛ پس وجود یا عدم وجود ضریب ۴ تأثیری در تعیین نویسنده‌ی برنده نداشته و با حذف آن می‌توان بار محاسباتی سیستم محاسبه‌ی فاصله را کمی کاهش داد.

۳-۴-۴ نتایج ارائه‌شده برای روش CNG

ارائه‌دهندگان روش CNG در مقاله‌ی مرجع {۲}، آزمایش‌هایی بر روی مجموعه‌متون‌هایی از سه زبان مختلف انجام داده‌اند که بهترین نتایج بدست آمده طی این آزمایش‌ها در جدول ۳-۲ درج شده است. در این جدول برای هر مجموعه‌متن علاوه بر بالاترین دقت بدست‌آمده، پارامترهایی که منجر به این دقت شده‌اند نیز درج شده است. همانطور که ملاحظه می‌شود در این پژوهش برای مجموعه‌متنِ زبان انگلیسی دقت ۱۰۰ درصد به ازای گستره‌ی نسبتاً وسیعی از پارامترهای N و L بدست آمده، در صورتی که برای سایر زبان‌ها هیچ‌گاه دقت ۱۰۰ درصد حاصل نشده است. البته نمی‌توان حکم کلی ارائه داد که دقت روش CNG در زبان انگلیسی بالاتر از زبان‌های یونانی و چینی است، زیرا این پژوهش فقط بر روی چند مجموعه‌متن محدود صورت پذیرفته است و نتایج آن را نمی‌توان به کل متون زبان انگلیسی تعمیم داد.

جدول ۳-۲: بهترین نتایج ارائه‌شده برای روش CNG در مقاله‌ی [۲]

مجموعه‌متن	تعداد نویسنده	بالاترین دقت (درصد)	پارامترهایی که منجر به بالاترین دقت شدند	
			L	N
انگلیسی	8	100	$1500 \leq L \leq 2000$	$4 \leq N \leq 8$
یونانی (الف)	10	85	4000	3
یونانی (ب)	10	97	5000	4
چینی	8	89	5000	4
			$4000 \leq L \leq 5000$	5

فصل چهارم

روش‌های پیشنهادی و مجموعه‌متن‌های استفاده‌شده

۴-۱ مقدمه

در ابتدای این فصل بهتر است مروری کوتاه بر مطالب ارائه‌شده از ابتدای پایان‌نامه تا کنون داشته باشیم. در فصل نخست با مسائل مختلف حوزه‌ی پردازش متن آشنا شده و مسئله‌ی تخصیص نویسنده را به عنوان حوزه‌ی اصلی پژوهش انتخاب نمودیم. سپس در فصل دوم روش‌های مختلف حل یک مسئله‌ی تخصیص نویسنده را بررسی کرده و از میان آن‌ها روش‌هایی موسوم به NDP را برای مطالعه‌ی بیش‌تر برگزیدیم. بعد از آن، در فصل سوم هشت مورد از مهم‌ترین روش‌های NDP را شرح دادیم.

حال می‌خواهیم دو پیشنهاد جدید برای حل مسائل تخصیص نویسنده ارائه دهیم. با توجه به اینکه این دو مورد نیز در دسته‌ی روش‌های NDP جای می‌گیرند، لذا با احتساب روش‌های ارائه شده در فصل سوم، در مجموع ۱۰ روش NDP خواهیم داشت. بدیهی است که به منظور بررسی و مقایسه‌ی این روش‌ها، می‌بایست آن‌ها را بر روی چند مجموعه‌متن مختلف آزمایش نماییم. در نتیجه نیاز به مجموعه‌متنی از زبان‌های فارسی و عربی داریم. در این پایان‌نامه از چهار مجموعه‌متن متفاوت استفاده شده است که یک مورد آن توسط نگارنده‌ی پایان‌نامه و مابقی توسط سایر پژوهش‌گران این حوزه گردآوری شده‌اند.

با توجه به موارد فوق، در ادامه‌ی پایان‌نامه می‌بایست سه موضوع را مورد بررسی قرار دهیم:

- شرح روش‌های پیشنهادی
- معرفی مجموعه‌متن‌های مورد استفاده در این پایان‌نامه
- ارائه و تحلیل نتایج حاصل از اعمال روش‌های NDP تشریح‌شده بر روی مجموعه‌متن‌های معرفی‌شده

از سه مورد فوق، دو مورد اول را در همین فصل بیان نموده و مورد سوم را به فصل بعدی وا می‌گذاریم. ادامه‌ی این فصل به صورت زیر سازماندهی شده است:

- در بخش‌های ۴-۲ و ۴-۳ دو روش پیشنهادی ارائه می‌شوند.
- مجموعه‌متن‌های مورداستفاده در بخش ۴-۴ معرفی شده و پیوست ۵ مقایسه می‌شوند.

در انتهای مقدمه نکته‌ای را از فصل اول یادآور می‌شویم: در حوزه‌ی تخصیص نویسنده، معمولاً برای اشاره به مجموعه‌ی داده‌های جمع‌آوری شده، به جای استفاده از واژه‌ی «پایگاه داده[1]»، از واژه‌ی «مجموعه‌متن[2]»

[1] Database
[2] Corpus

استفاده می‌شود. در این فصل نیز همانند سایر فصول قبل، عرف رایج حوزه‌ی تخصیص نویسنده رعایت شده و واژه‌ی مجموعه‌متن مورد استفاده قرار گرفته است. لازم به ذکر است که این کلمه با استفاده از «ها» جمع بسته شده و لذا عبارت «مجموعه‌متن‌ها» به معنای «چند مجموعه‌متن» می‌باشد.

۴-۲ روش تفاضل-اندیس وزن‌دار

در این بخش می‌خواهیم یکی از دو روش پیشنهادی را شرح دهیم. به منظور شرح این روش همان روند مورد استفاده در فصل سوم را پی می‌گیریم، بدین صورت که ابتدا فلسفه‌ی ارائه‌ی روش را شرح داده، سپس «سیستم استخراج پروفایل» را بیان نموده و در نهایت نحوه‌ی عملکرد «سیستم محاسبه‌ی فاصله» را بررسی می‌نماییم.

۴-۲-۱ فلسفه‌ی ارائه‌ی روش

اولین روش پیشنهادی در واقع حاصل ایجاد تغییراتی در روش CNG می‌باشد. از فصل سوم به یاد داریم که در روش CNG، انگرام‌های پرتکرار هر متن در پروفایل آن متن جای گرفته و هنگام محاسبه‌ی فاصله، فرکانس انگرام‌ها مبنای کار قرار می‌گیرد. ایده‌ی بنیادین این روش پیشنهادی، استفاده از اندیس انگرام‌های موجود در پروفایل می‌باشد. به عبارت دیگر به جای مقدار فرکانس یک انگرام موجود در پروفایل، مکان قرار گرفتن آن در پروفایل (اندیس انگرام) برای ما مهم است. این روش پیشنهادی را «تفاضل اندیس وزن‌دار[1]» نامیده و مختصراً آن را با WIS نمایش می‌دهیم. دلیل این نام‌گذاری در ادامه روشن خواهد شد. این روش نیز همانند سایر روش‌های NDP دارای دو پارامتر اساسی N (طول انگرام) و L (طول پروفایل) است.

۴-۲-۲ سیستم استخراج پروفایل

همانند روش CNG-SPI، در روش WIS نیز پروفایل یک متن دلخواه (T) تنها دارای یک مجموعه است::

[1] Weighted Index-Subtraction (WIS)

$$PR_T^{WIS} = \left\{ G_T^{WIS} \right\} \qquad (۱-۴)$$

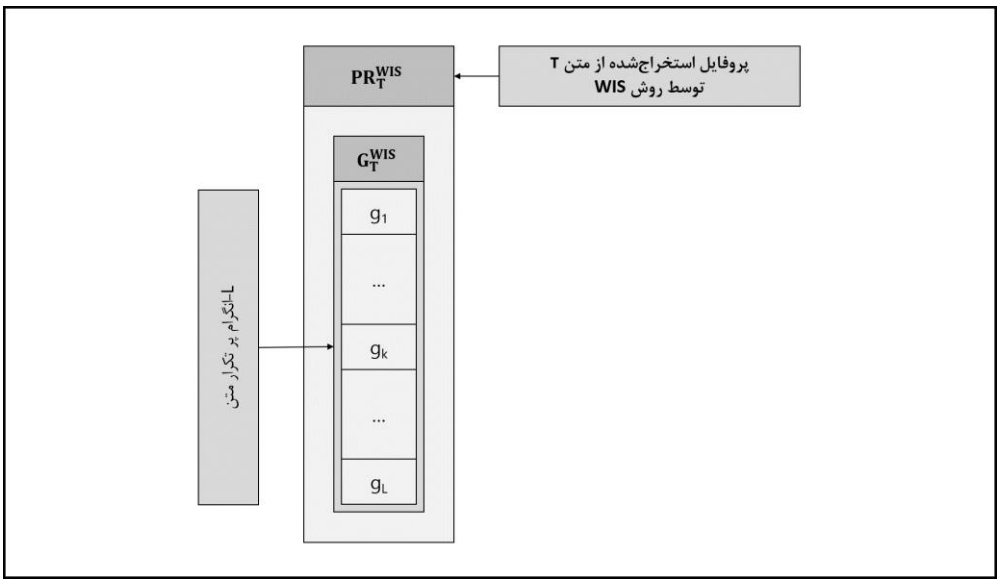
و مجموعه‌ی G_T^{WIS} دقیقاً همانند روش CNG-SPI تعریف و تشکیل می‌شود، یعنی: G_T^{WIS} عبارتست از مجموعه‌ای شامل L-انگرام پر تکرار متن T.

به عبارت دیگر روش WIS از پروفایل ساده‌شده (SP) که اولین‌بار در روش CNG-SPI معرفی شد، استفاده می‌نماید. یعنی:

$$PR_T^{WIS} = PR_T^{SPI} \qquad (۲-۴)$$

یک نمای کلی از پروفایل PR_T^{WIS} در شکل ۴-۱ نمایش داده شده است. همانطور که ملاحظه می‌شود در این پروفایل دیگر مقادیر فرکانس متناظر با هر انگرام وجود ندارند. روش WIS برای هر انگرام به جای فرکانس، کمیت دیگری به نام «اندیس[1]» (i) تعریف نموده که نمایانگر ترتیب قرار گرفتن یک انگرام در پروفایل می‌باشد. برای یک انگرام دلخواه (مثل g)، اندیس بدین صورت تعریف می‌شود:

$$i_T^{WIS}(g) = \begin{cases} k \quad , \quad \text{k=index of g in} G_T^{WIS} & if \;\; g \in G_T^{WIS} \\ L+1 & if \;\; g \notin G_T^{WIS} \end{cases} \qquad (۳-۴)$$



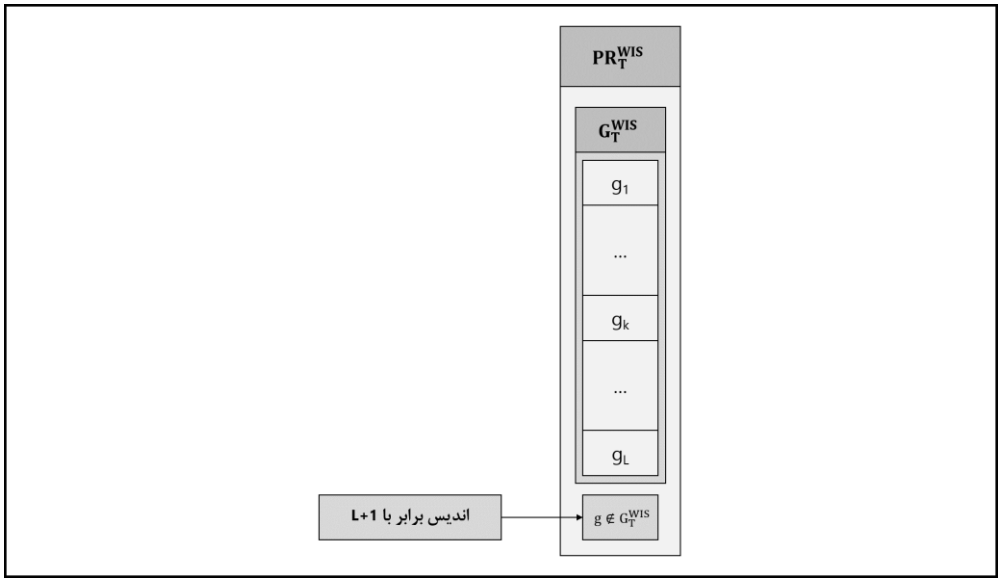
شکل ۴-۱: نمای کلی از پروفایل استخراج‌شده از متن T توسط روش WIS

[1] Index

به عبارت دیگر برای هر انگرام دلخواه (g):

- اگر g عضوی از G_T^{WIS} باشد:
  - $i_T^{WIS}(g)$ برابر است با اندیس قرار گرفتن g در مجموعه‌ی G (مقداری بین یک تا L)
- اگر g عضوی از G_T^{WIS} نباشد:
  - $i_T^{WIS}(g)$ برابر L+1 فرض می‌شود.

پس در مجموع می‌توان گفت که اگر انگرامی در پروفایل وجود داشته باشد، اندیس آن به راحتی از روش PR_T^{WIS} خوانده می‌شود؛ اما اگر انگرامی در پروفایل وجود نداشته باشد، فرض می‌کنیم در یک ردیف پایین‌تر از آخرین سطر PR_T^{WIS} قرار گرفته و اندیس آن برابر با L+1 می‌باشد. این نکته در شکل ۴-۲ نمایش داده شده است.



شکل ۴-۲: اندیس انگرامی خارج از پروفایل، در روش WIS

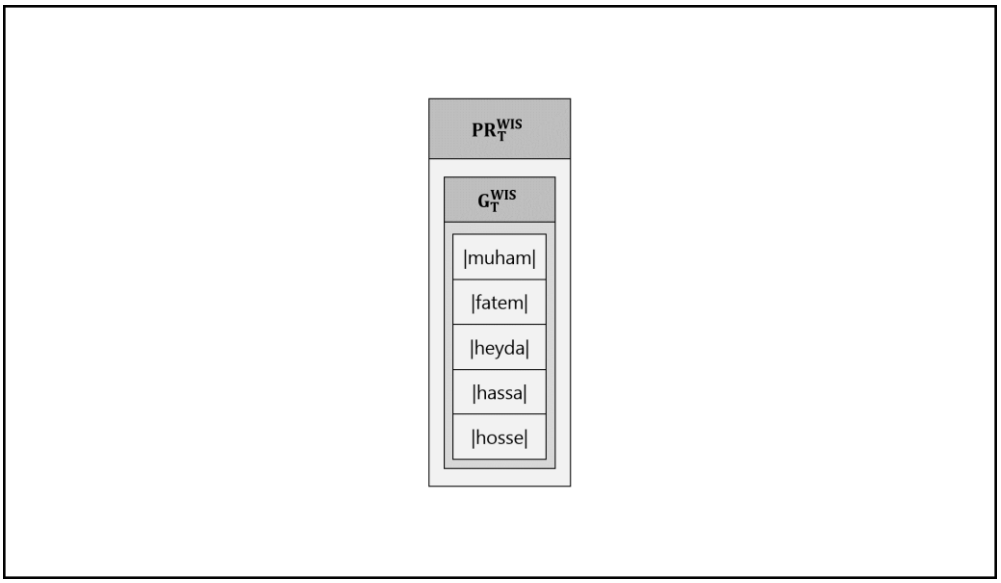
به منظور درک بهتر کمیت اندیس، مثالی ارائه می‌دهیم. فرض کنیم پروفایل استخراج‌شده از یک متن به صورت شکل ۴-۳ باشد:

در این حالت برای انگرام g_x=|fatem| (که در پروفایل وجود دارد)، داریم:

$$i(g_x)=2$$

اما برای انگرام |gx=|sajja (که در پروفایل وجود ندارد)، داریم:

$$i(g_y)=L+1=5+1=6$$



شکل ۴-۳: پروفایل استخراج‌شده از یک متن فرضی توسط روش WIS

۴-۲-۳ سیستم محاسبه‌ی فاصله

از فصل سوم به خاطر داریم که در روش CNG برای محاسبه‌ی فاصله‌ی میان پروفایل یک نویسنده (A) و پروفایل متن دیده‌نشده (U)، از رابطه‌ی زیر استفاده می‌شود:

$$Dist^{CNG}(U,A)=\sum_{g\in G} d^{CNG}(g) \qquad G=G_U^{CNG}\cup G_A^{CNG} \qquad (۴-۴)$$

که در آن:

$$d^{CNG}(g)=\left(\frac{f_U^{CNG}(g)-f_A^{CNG}(g)}{f_U^{CNG}(g)+f_A^{CNG}(g)}\right)^2 \qquad (۴-۵)$$

در روش WIS روند مشابهی جهت محاسبه‌ی فاصله‌ی میان دو پروفایل ارائه می‌شود؛ با این تفاوت که در این روند، به جای آنکه فاصله بر مبنای فرکانس انگرام‌ها محاسبه شود، بر مبنای اندیس آن‌ها تعیین می‌شود. بدین‌صورت که:

$$D^{WIS}(U,A)=\sum_{g\in G} d^{WIS}(g) \qquad G=G_U^{WIS}\cup G_A^{WIS} \qquad (۶-۴)$$

که در آن:

$$d^{WSI}(g)=\left(\frac{i_U^{WIS}(g)-i_A^{WIS}(g)}{i_U^{WIS}(g)+i_A^{WIS}(g)}\right)^2 \qquad (۷-۴)$$

به عبارت دیگر فاصله‌ی متناظر با یک انگرام (g) وابسته به اندیس آن انگرام در پروفایل U و A بوده و شامل دو بخش اساسی است:

۱) تفاضل اندیس‌ها: $i_U^{WIS}(g)-i_A^{WIS}(g)$

این عبارت در صورت کسر قرار گرفته است تا هرچه اختلاف میان اندیس انگرام g در دو پروفایل بیش‌تر باشد، فاصله‌ی دو پروفایل نیز افزایش یابد.

۲) وزن: $\frac{1}{i_U^{WIS}(g)+i_A^{WIS}(g)}$

این عبارت موجب می‌شود تا هر چه به سمت خانه‌های پایین‌تر جدول (انگرام‌های با فرکانس کمتر) پیش می‌رویم، اختلاف اندیس‌های متناظر با این خانه‌ها تأثیر کمتری در فاصله‌ی میان دو پروفایل داشته باشد.

به منظور درک بهتر نحوه‌ی وزن‌دهی، دو حالت زیر را در نظر می‌گیریم:

$$A: \quad i_U^{WIS}(g)=10 \ , \ i_A^{WIS}(g)=7$$
$$B: \quad i_U^{WIS}(g)=50 \ , \ i_A^{WIS}(g)=47$$

در هر دو حالت A و B، تفاضل اندیس‌ها یکسان و برابر با ۳ است. اما در حالت A وزنی معادل $\frac{1}{10+7}=\frac{1}{17}$ داریم، در صورتی که وزن حالت B برابر است با $\frac{1}{50+47}=\frac{1}{97}$. به عبارت دیگر از آن‌جا که در حالت B، اندیس‌ها مربوط به خانه‌های پایین‌تری از پروفایل بوده (و دارای اهمیت کمتری هستند)، لذا وزن کمتری به خود اختصاص داده تا در محاسبه‌ی فاصله‌ی میان دو پروفایل تأثیر کمتری داشته باشند. در انتها متذکر می‌شویم که با توجه به دو بخش اصلی موجود در رابطه‌ی $d_i^{WIS}(g)$ (یعنی تفاضل اندیس‌ها و وزن)، روش پیشنهادی را «تفاضل اندیس‌های وزن‌دار» نامیدیم.

۴-۳ روش انگرام‌های پراکنده

همانطور که در بخش ۴-۲ و قبل‌تر در فصل سوم دیدیم، پروفایلی که توسط روش CNG از یک متن استخراج می‌شود مبتنی بر پرتکرارترین انگرام‌های موجود در آن متن می‌باشد. حال می‌خواهیم این تعریف کلاسیک را تغییر داده و روشی را ارائه دهیم که در آن، مبنای دیگری (غیر از پرتکرارترین انگرام‌ها) برای استخراج پروفایل به کار رفته است. در این بخش نیز به همان روال قبلی ابتدا فلسفه‌ی ارائه‌ی روش را بیان نموده، سپس سیستم استخراج پروفایل را شرح داده و در نهایت سیستم محاسبه‌ی فاصله را بررسی می‌نماییم.

۴-۳-۱ فلسفه‌ی ارائه‌ی روش

به منظور شرح ایده‌ی اصلی این روش، ابتدا یک مثال ساده بیان می‌کنیم. فرض کنیم مجموعه‌متنی شامل ۳ نویسنده در اختیار داریم. پروفایل‌های استخراج‌شده توسط روش CNG برای این سه نویسنده به صورت شکل ۴-۴ می‌باشد.

G_{A1}^{CNG}	F_{A1}^{CNG}	G_{A2}^{CNG}	F_{A2}^{CNG}	G_{A3}^{CNG}	F_{A3}^{CNG}
g_a	0.050	g_l	0.045	g_y	0.057
g_b	0.043	g_x	0.040	g_p	0.051
g_x	0.040	g_y	0.039	g_q	0.047
g_c	0.037	g_m	0.037	g_x	0.040
g_y	0.029	g_n	0.035	g_r	0.03

شکل ۴-۴: پروفایل سه نویسنده‌ی موجود در مجموعه‌متن فرضی (این پروفایل‌ها توسط روش CNG استخراج شده اند)

در این مثال هر چند انگرام g_x در پروفایل هر سه نویسنده دارای فرکانس بالایی است (و در ردیف‌های ابتدایی جدول آمده)، اما فرکانس آن به ازای هر سه نویسنده دارای مقدار یکسانی (برابر با ۰,۰۴۰) بوده و لذا این انگرام تمایز خاصی میان این سه نویسنده ایجاد نمی‌کند و در نتیجه نمی‌تواند اطلاعات چندانی در اختیار ما دهد. ایده‌ی اصلی روش پیشنهادی این است که: به جای آنکه انگرام‌های دارای فرکانس بالا در پروفایل قرار گیرند، انگرام‌هایی را در پروفایل قرار دهیم که فرکانس استفاده‌ی آن‌ها توسط نویسندگان مختلف دارای تغییرات زیادی باشد. به عنوان مثال انگرام g_y که در پروفایل سه نویسنده به ترتیب دارای فرکانس‌های ۰,۰۲۹، ۰,۰۳۹ و ۰,۰۵۷ می‌باشد. به منظور سنجش شدت این تغییرات، استفاده از کمیت «واریانس[1]» پیشنهاد می‌شود، به همین جهت این روش را «انگرام‌های پراکنده[2]» نامیده و با VNG نمایش می‌دهیم.

۲-۳-۴ سیستم استخراج پروفایل

در روش VNG، یک گام مقدماتی برای استخراج پروفایل یک متن وجود دارد. چنین گامی در روش CNG وجود نداشته، اما شبیه آن در روش RLP مورد استفاده قرار گرفته است. هدف از این گام تهیه نمودن لیستی با نام «لیست انگرام‌های پراکنده[3]» می‌باشد. این لیست را اختصاراً با vList نمایش می‌دهیم. در ادامه ابتدا نحوه‌ی استخراج vList را شرح داده و سپس مراحل استخراج پروفایل یک متن دلخواه (T) را بیان می‌کنیم.

همانند سایر روش‌ها قبل از شروع فرایند استخراج پروفایل یک متن، می‌بایست دو پارامتر اساسی طول انگرام (N) و طول پروفایل (L) انتخاب شوند. این گام (تعیین پارامترهای N و L) را به روال سابق اصطلاحاً «گام صفرم» می‌نامیم تا در شمارش گام‌های اصلی ظاهر نشود.

---

[1] Variance
[2] Variant N-Grams (VNG)
[3] List of Variant N-grams List (VLIST)

الف) استخراج vList

vList با توجه به ابرمتن‌های موجود برای نویسندگان، تشکیل شده و مستقل از متن T می‌باشد. هدف از تشکیل این لیست تعیین انگرام‌های دارای بیش‌ترین واریانس می‌باشد. به منظور استخراج این لیست می‌بایست گام‌های زیر را طی نمود.

گام اول: محاسبه‌ی پروفایل CNG همه‌ی نویسنده‌ها

در این گام پروفایل تک تک ابرمتن‌ها را با استفاده از روش CNG استخراج می‌نماییم. بدین‌منظور پارامترهای N و L انتخاب‌شده در گام صفرم را مد نظر قرار می‌دهیم. در پایان این گام به ازای هر نویسنده، یک پروفایل CNG خواهیم داشت:

$$PR_{A_1}^{CNG},...,PR_{A_j}^{CNG},...,PR_{A_M}^{CNG}$$

گام دوم: تشکیل مجموعه‌ی G

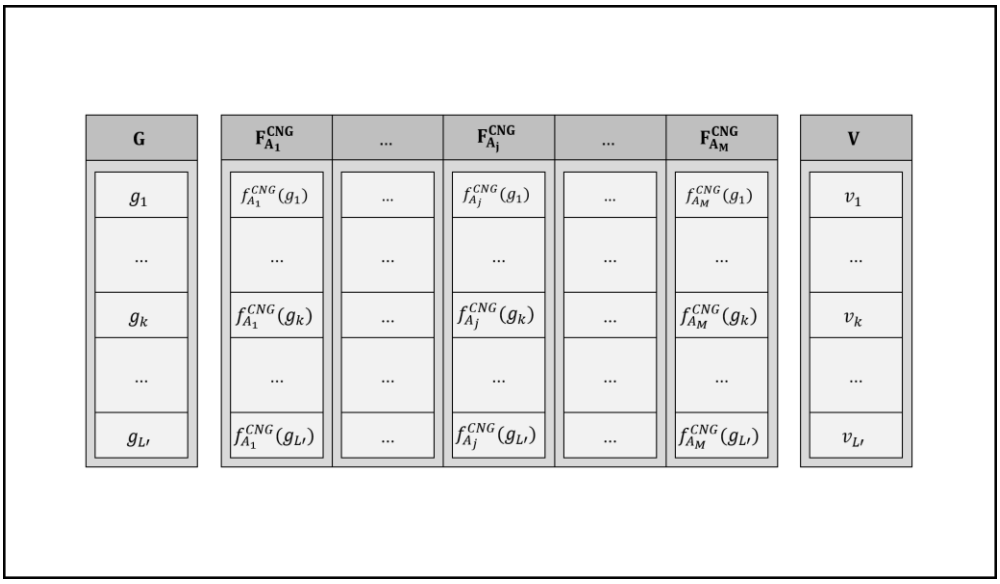
در این گام مجموعه‌ای حاصل از اجتماع تمامی پروفایل‌های استخراج‌شده در گام قبل، تشکیل داده و آن را G می‌نامیم:

$$G = \bigcup_{j=1}^{M} G_{A_j}^{CNG} \qquad (۸-۴)$$

گام سوم: محاسبه‌ی واریانس اعضای G

در این گام جدولی همانند شکل ۴-۵ تشکیل می‌دهیم که در آن:

- ستون اول شامل تمامی اعضای مجموعه‌ی G است.
- ستون‌های $F_{A_1}^{CNG}$ تا $F_{A_M}^{CNG}$ شامل فرکانس‌های انگرام‌های مجموعه‌ی G در پروفایل نویسنده‌ها می‌باشند. به عنوان مثال در سطر k-ام از ستون $F_{A_j}^{CNG}$، فرکانس متناظر با انگرام g_k از پروفایل نویسنده‌ی A_j خوانده شده و در جدول درج شده است.
- ستون آخر شامل واریانس هر انگرام روی ستون‌ها A_1 تا A_M است.



شکل ۴-۵: جدول تشکیل‌شده جهت استخراج لیست انگرام‌های پراکنده

به عبارت دیگر k-امین المان از ستون آخر جدول عبارتست از:

$$v_k = \text{variance}\left\{ f_{A_j}^{CNG}(g_k) \middle| j = 1,...,M \right\} \qquad (۹-۴)$$

با در نظر گرفتن تعریف واریانس در علم آمار متغیرهای گسسته می‌توان گفت:

$$v_k = \frac{1}{M}\sum_{j=1}^{M}\left[f_{A_j}^{CNG}(g_k) - \mathrm{m}_k \right]^2 \qquad (۱۰-۴)$$

که در آن:

$$m_k = \frac{1}{M}\sum_{j=1}^{M} f_{A_j}^{CNG}(g_k) \qquad (۱۱-۴)$$

در نهایت پس از تکمیل این جدول، L-انگرام دارای بالاترین واریانس را جدا نموده و در لیست vList قرار می‌دهیم. پس vList در واقع مجموعه‌ایست شامل L انگرام که فرکانس استفاده‌ی آن‌ها توسط نویسندگان مختلف، دارای بالاترین واریانس بوده است. این لیست را می‌توان به صورت یک مجموعه نمایش داد:

$$vList = \{vg_1,...,vg_k,...,vg_L\} \qquad (۱۲-۴)$$

ب) استخراج پروفایل متن

در روش VNG پروفایل یک متن دلخواه (T) دارای دو مجموعه است:

$$PR_T^{VNG} = \{vList, F_T^{VNG}\} \qquad (۱۳-۴)$$

مجموعه‌ی vList در قسمت الف و مستقل از متن T تولید شد. حال در این قسمت مجموعه‌ی F_T^{VNG} را محاسبه خواهیم نمود. این مجموعه در واقع حکم یک بردار عددی را دارد. به طور خلاصه می‌توان گفت بردار F_T^{VNG} عبارتست از فرکانس نرمال‌شده‌ی انگرام‌های موجود در لیست vList در متن T. به منظور روشن‌تر شدن مفهوم این بردار، مراحل محاسبه‌ی آن را در دو گام شرح می‌دهیم:

گام اول: تشکیل بردار F_T^{VNG}

ابتدا برداری به طول L المان تشکیل می‌دهیم که هر المان از آن، متناظر با یک انگرام موجود در vList می‌باشد. اگر المان k-ام از بردار F_T^{VNG} را با نماد $F_T^{VNG}[k]$ نمایش دهیم، آن‌گاه می‌توان گفت $F_T^{VNG}[k]$ متناظر است با انگرام vg_k.

گام دوم: پر کردن بردار F_T^{VNG}

برای پر کردن بردار تشکیل‌شده در گام قبل، از رابطه‌ی زیر استفاده می‌نماییم:

$$F_T^{VNG}[k] = f_T^{CNG}(vg_k) \qquad (۱۴-۴)$$

به عبارت دیگر k-امین المان از بردار F_T^{VNG} برابرست با فرکانس نرمال‌شده‌ی k-امین انگرام vList در متن T.

در انتها، به منظور رعایت روند متداول در تعریف پروفایل یک متن، پروفایل استخراج‌شده از متن T توسط روش VNG را در قالبی جدید معرفی می‌کنیم:

$$PR_T^{VNG} = \{G_T^{VNG}, F_T^{VNG}\} \quad , \quad G_T^{VNG} = vList \qquad (۱۵-۴)$$

پس در روش VNG نیز، همانند روش CNG هر پروفایل از دو مجموعه‌ی G و F تشکیل می‌شود. مجموعه‌ی G شامل انگرام‌های پراکنده و مجموعه‌ی F شامل فرکانس متناظر اعضای مجموعه‌ی G است. لازم بذکر

است بر خلاف مدل CNG که در آن هر متن دلخواه (مثل T)، مجموعه‌ی G متناظر با خود را داشت (G_T^{CNG})، در مدل VNG تمامی متون مربوط به یک مجموعه‌متن، یک مجموعه‌ی G یکسان دارند که آن همان vList است.

۳-۳-۴ سیستم محاسبه‌ی فاصله

فرض کنیم پروفایل یک نویسنده (A) و پروفایل متن دیده‌نشده (U) را در اختیار داریم. می‌خواهیم فاصله‌ی میان آن دو را محاسبه نماییم. در بخش قبلی بیان شد که تمامی متون مربوط به یک مجموعه‌متن، دارای مجموعه‌ی G یکسانی می‌باشند، لذا مجموعه‌های G_U^{VNG} و G_A^{VNG} نیز با یکدیگر برابر هستند.

پس می‌توان نتیجه گرفت المان‌های متناظر بردارهای F_U^{VNG} و F_A^{VNG} هم‌معنی می‌باشند. لذا می توان مسئله‌ی محاسبه‌ی فاصله میان پروفایل U و پروفایل A را به مسئله‌ی محاسبه‌ی فاصله‌ی میان دو بردار F_U^{VNG} و F_A^{VNG} تبدیل نمود. در نتیجه برای محاسبه‌ی فاصله‌ی میان U و A می‌توان از انواع معیارهای فاصله‌ی برداری استفاده نمود، از جمله: فاصله‌ی اقلیدسی[1]، فاصله‌ی کسینوسی[2]، فاصله‌ی مَنهَتَن[3] و ...

همانند مدل RLP، ما نیز معیار فاصله‌ی کسینوسی را برای محاسبه‌ی فاصله‌ی میان دو پروفایل بر می‌گزینیم. پس:

$$D^{VNG}(U,A) = 1 - \frac{\left\langle F_U^{VNG} \bullet F_A^{VNG} \right\rangle}{\left\| F_U^{VNG} \right\| \left\| F_A^{VNG} \right\|} \qquad (۱۶-۴)$$

که در آن:

- $\left\langle F_U^{VNG} \bullet F_A^{VNG} \right\rangle$ عبارتست از ضرب داخلی دو بردار F_U^{VNG} و F_A^{VNG}
- $\left\| F_U^{VNG} \right\|$ و $\left\| F_A^{VNG} \right\|$ به ترتیب اندازه‌ی بردارهای F_U^{VNG} و F_A^{VNG} می‌باشند.

[1] Euclidean Distance
[2] Cosine Distance
[3] Manhattan Distance

۴-۴ معرفی مجموعه‌متن‌های استفاده‌شده

همان‌طور که در مقدمه‌ی فصل بیان شد، در این پایان‌نامه از چهار مجموعه‌متن متفاوت استفاده شده است. در این بخش می‌خواهیم ضمن معرفی این مجموعه‌ها، تحلیلی کوتاه از متون موجود در آن‌ها ارائه دهیم. برای تعیین مجموعه‌متن دو رویکرد وجود دارد: گردآوری یک مجموعه‌متن جدید یا انتخاب یکی از مجموعه‌متن‌های آماده (که توسط سایر پژوهشگران جمع‌آوری شده اند). مستقل از اینکه کدام رویکرد مورد استفاده قرار گیرد، می‌بایست دو پرسش اساسی را مورد توجه قرار داد:

پرسش ۱: هنگام جمع‌آوری یک مجموعه‌متن جدید و یا انتخاب یکی از مجموعه‌های آماده، اولین سؤالی که در ذهن ایجاد می‌شود آنست که: «در حوزه‌ی تخصیص نویسنده، یک مجموعه‌متن ایده‌آل می‌بایست چه ویژگی‌هایی داشته باشد؟» پاسخ این پرسش در بخش ۴-۴-۱ ارائه شده است.

پرسش ۲: فرض کنیم با توجه به ویژگی‌های یک مجموعه‌متن ایده‌آل، مجموعه‌ی جدید را گردآوری نموده و یا یکی از مجموعه‌متن‌های آماده را انتخاب نمودیم. حال می‌خواهیم مجموعه‌متن تعیین‌شده را معرفی نماییم. سؤال مهمی که مطرح می‌شود آنست که: «هنگام معرفی یک مجموعه‌متن، می‌بایست چه اطلاعاتی از آن ارائه گردد؟» به عنوان مثال صرف اینکه اعلام کنیم یک مجموعه‌متن شامل ۱۰۰ متن از ۵ نویسنده‌ی فارسی‌زبان می‌باشد، کافیست؟ یا اطلاعات دیگری را باید ارائه گردد؟ این پرسش در بخش ۴-۴-۲ پاسخ داده می‌شود.

پس از پاسخ دادن به این پرسش‌ها، چهار مجموعه‌متن مورد استفاده در این پایان‌نامه معرفی می‌شوند. این معرفی در بخش‌های ۴-۴-۳ تا ۴-۴-۶ صورت می‌پذیرد.

۴-۴-۱ ویژگی‌های مجموعه‌متن ایده‌آل

بر مبنای مقالات [۳۱ و ۵۰] می‌توان ویژگی‌های یک مجموعه‌متن ایده‌آل را به طور خلاصه چنین بیان نمود: «متون موجود در یک مجموعه‌متن ایده‌آل می‌بایست به گونه‌ای انتخاب شوند که هویت نویسنده، تنها و یا مهم‌ترین عامل جداسازی متون باشد.» به منظور گردآوری چنین مجموعه‌ای، هنگام انتخاب نویسندگان و متن‌های آن‌ها باید سه ویژگی اصلی را رعایت نمود:

ویژگی اولِ مجموعه‌متن ایده‌آل:

به منظور دست‌یابی به یک مجموعه‌متن ایده‌آل، متون انتخاب‌شده می‌بایست در یک دوره‌ی زمانی تقریباً مشترک نوشته شده باشند، زیرا مقایسه‌ی متونی با فاصله‌ی تاریخی زیاد ارزش علمی چندانی ندارد. به عنوان مثال فرض کنیم این دو نویسنده برای مجموعه‌متن انتخاب گردند: جلال آل‌احمد (نویسنده‌ی معاصر) و ابوالفضل بیهقی (نویسنده‌ی قرن پنجم هجری شمسی). مسلماً متون این دو نویسنده چنان متمایز است که تخصیص یک متن جدید به یکی از این دو، ارزش علمی زیادی نخواهد داشت. دلیل اصلی این تمایز، «تَطَوُّر» زبان فارسی در طول زمان است.

ویژگی دومِ مجموعه‌متن ایده‌آل:

در حالت ایده‌آل تمامی متون گردآوری‌شده باید دقیقاً در ارتباط با یک موضوع[1] و در یک ژانر[2] خاص نوشته شده باشند. در غیر این‌صورت نمی‌توان هویت نویسنده را به عنوان تنها عامل جداساز متون در نظر گرفت.

ویژگی سومِ مجموعه‌متن ایده‌آل:

از آن‌جا که مواردی همچون سن، سطح و نوع تحصیلات، ویژگی‌های اخلاقی و جنسیت بر روی سبک نوشتاری افراد تأثیرگذارند، لذا در حالت ایده‌آل بهترست نویسندگان یک مجموعه‌متن به گونه‌ای انتخاب شوند که از نظر این ویژگی‌ها در وضعیت مشابهی قرار داشته باشند.

معمولاً هر کدام از مجموعه‌متن‌های حوزه‌ی تخصیص نویسنده فقط برخی از سه ویژگی فوق را دارا می‌باشند؛ زیرا گردآوری مجموعه‌متنی که تمامی موارد فوق را به طور همزمان رعایت کرده باشد، بسیار مشکل است. تلاش‌های صورت گرفته جهت تهیه‌ی مجموعه‌متن ایده‌آل حاکی از آنست که لحاظ کردن ویژگی‌های دوم و سوم به مراتب سخت‌تر از لحاظ نمودن ویژگی اول می‌باشد [۳۱]. به همین جهت معمولاً هنگام گردآوری یک مجموعه‌متن ویژگی اول را لحاظ نموده و از سایر ویژگی‌ها صرف نظر می‌کنند. با این وجود در بعضی از پژوهش‌های صورت‌گرفته در زبان‌های لاتین، مجموعه‌متن‌هایی بسیار نزدیک به حالت ایده‌آل ارائه گردیده‌اند. از جمله‌ی این مجموعه‌ها می‌توان به «نامه‌های فدرالیست[3]» اشاره نمود که در مقاله‌ی [۹۱] معرفی شده است. لازم بذکرست که تا کنون هیچ مجموعه‌متن کاملاً ایده‌آلی برای زبان فارسی ارائه نشده است [۵۰].

---

[1] Subject
[2] Genre
[3] Federalist Papers

۲-۴-۴ نحوه‌ی معرفی یک مجموعه‌متن

پس از آشنایی با ویژگی‌های یک مجموعه‌متن ایده‌آل برای حوزه‌ی تخصیص نویسنده، در این قسمت می‌خواهیم روند رایج معرفی یک مجموعه‌متن را مختصراً شرح دهیم. معمولاً پژوهشگران حوزه‌ی تخصیص نویسنده هنگام معرفی یک مجموعه‌متن، اطلاعاتی نظیر موارد زیر را ارائه می‌دهند:

- اطلاعاتی در مورد نویسندگان موجود در مجموعه: بازه‌ی حیات هر یک از نویسندگان، سبک نگارشی هر یک از نویسندگان و نام کتاب‌هایی که متون از آن‌ها استخراج شده است
- اطلاعاتی در مورد طول متون موجود در مجموعه: طول کوتاه‌ترین متن موجود (بر حسب کلمه و نویسه)، طول بلندترین متن موجود (بر حسب کلمه و نویسه) و متوسط طول متون موجود (بر حسب کلمه و نویسه)

ما نیز این روند رایج را هنگام معرفی مجموعه‌متن‌ها پی گرفته و سعی می‌کنیم برای هر مجموعه‌متن، چهار جدول ارائه دهیم:

- جدول A: در این جدول مشخصات مربوط به نویسندگان ارائه می‌گردد؛ اطلاعاتی همچون نام نویسندگان، بازه‌ی حیات، نام کتاب‌های مورد استفاده در مجموعه‌متن.
- جدول B: در این جدول اطلاعات آماری طول متون هر کدام از نویسندگان به صورت جدا درج می‌شود؛ به عنوان مثال برای هر نویسنده اطلاعاتی نظیر کمینه، بیشینه و متوسط طول متون مربوط به آن نویسنده ارائه می‌گردد.
- جدول C: در این جدول اطلاعات آماری کل متون موجود در مجموعه‌متن به صورت یک‌جا (و نه به تفکیک نویسندگان) ارائه می‌گردد. این جدول در واقع حاصل جمع‌بندی اطلاعات مندرج در جدول B است.
- جدول D: در این جدول اطلاعات مربوط به تعداد متون و تعداد نویسندگان به صورت خلاصه بیان می‌شود. این جدول نیز در واقع چکیده‌ای از جدول B است.

البته برای برخی از مجموعه‌متن‌ها، به دلیل عدم ارائه‌ی اطلاعات کافی توسط گردآورنده‌ی مجموعه، فقط تعدادی از جداول فوق قابل تشکیل است. به عنوان مثال برای برخی از مجموعه‌متن‌ها، هیچ گونه اطلاعاتی از نویسندگان ارائه نشده است و لذا نمی‌توان جدول A را برای آن مجموعه‌ها تشکیل داد.

در ادامه هنگام معرفی هر مجموعه‌متن، تلاش می‌کنیم تا در صورت امکان تمامی جداول A تا D را برای آن‌ها تشکیل دهیم. هدف از ارائه‌ی این جداول بدین صورت است:

- جداول A و B بیش‌تر به منظور آشنا شدن خواننده با فضای کلی هر مجموعه ارائه می‌شوند. جدول B در واقع مرجع استخراج جداول C و D می‌باشد و حذف آن تأثیر چندانی بر معرفی مجموعه‌متن ندارد؛ اما به دلیل تکمیل بحث و درک کامل‌تر خواننده از مجموعه‌متن، این جدول نیز برای مجموعه‌متن‌های معرفی‌شده ارائه می‌گردد.
- هدف از ارائه‌ی جداول C و D ایجاد امکان مقایسه‌ی مجموعه‌متن‌ها است. همانطور که در بخش ۴-۵ خواهیم دید اطلاعات مندرج در این جداول جهت مقایسه‌ی مجموعه‌ها بسیار سودمند است.

از آن‌جا که مقایسه‌ی مجموعه‌متن‌های معرفی‌شده در بخش ۴-۵ صورت می‌پذیرد، لذا در ادامه و هنگام تشکیل جداول C و D توضیحات چندانی ارائه نداده و تحلیل اطلاعات مندرج در این جداول را به پیوست شماره ۵ وا می‌گذاریم.

در انتها یادآور می‌شویم که برای سنجش طول یک متن از دو معیار مختلف استفاده می‌شود: تعداد کلمات و تعداد نویسه‌ها. هر چند عرف رایج آنست که هنگام معرفی یک مجموعه‌متن، اطلاعات مربوط به طول متون بر هر دو مبنا ارائه گردند، اما هنگام تحلیل این داده‌ها معمولاً تعداد کلمات بیش‌تر مورد توجه قرار می‌گیرند. به عنوان مثال برای مقایسه‌ی طول متون موجود در یک مجموعه، معمولاً متن دارای کمترین تعداد کلمه به عنوان کوتاه‌ترین متن شناخته می‌شود، نه متن دارای کمترین تعداد نویسه (البته ممکن است این دو همزمان در یک متن رخ دهند، یعنی متن دارای کمترین تعداد کلمه، دارای کمترین تعداد نویسه نیز باشد). ما نیز در این پایان‌نامه همین روند را پی گرفته و اطلاعات مربوط به طول متون را بر هر دو مبنا ارائه می‌دهیم، اما هنگام تحلیل بیش‌تر بر تعداد کلمات تکیه می‌کنیم.

۳-۴-۴ مجموعه‌متن CPPT

این مجموعه‌متن توسط نگارنده‌ی پایان‌نامه جمع‌آوری شده و شامل متونی از ۶ نویسنده‌ی معاصر فارسی‌زبان می‌باشد. این مجموعه با عنوان «متون منثور معاصر فارسی[1]» نام‌گذاری شده و با نماد اختصاری CPPT از آن یاد می‌شود. هنگام گردآوری این مجموعه، تلاش شد نخستین ویژگی یک مجموعه‌متن ایده‌آل رعایت شود، یعنی آثاری مورد استفاده قرار گرفتند که در یک بازه‌ی زمانی تقریباً مشترک نوشته شده‌اند. در ادامه جداول A تا D را برای مجموعه‌ی CPPT ارائه می‌دهیم:

[1] Contemporary Persian Prose Texts (CPPT)

الف) جدول A: اطلاعات نویسندگان

در این جدول نام، بازه‌ی حیات و تعداد متون هر نویسنده درج شده است (جدول ۴-۱). هم‌چنین نام کتاب‌هایی که متون نمونه از آن استخراج شده‌اند، به همراه تعداد متون هر یک نیز ارائه گردیده است. به عنوان مثال از ۱۹ متنی که برای «شهید مطهری» گردآوری شده است، ۶ متن مربوط به کتاب «انسان و ایمان»، ۴ متن مربوط به کتاب «جهان‌بینی توحیدی» و ۹ متن مربوط به کتاب «وحی و نبوت» می‌باشد.

فایل الکترونیکی مربوط به مجموعه‌متن بدین صورت تهیه شده است:

- به ازای هر نویسنده، یک پوشه[1]ی خاص به نام نویسنده ایجاد گردیده است.
- در هر پوشه، متون نویسنده به صورت فایل متنی[2] (با فرمت *.txt) قرار گرفته‌اند. این فایل‌های متنی به ترتیب از ۱ تا تعداد کل متون هر نویسنده شماره‌گذاری شده‌اند.

به عنوان مثال متون مربوط به شهید مطهری در پوشه‌ای با نام "Motahhari" ذخیره شده و از "1.txt" تا "19.txt" شماره‌گذاری شده‌اند. در ستون آخر جدول ۴-۱، شماره‌ی متون هر کتاب به صورت جداگانه ارائه شده است. به عنوان مثال از ۱۹ فایل متنی مربوط به شهید مطهری، فایل‌های ۱ تا ۶ مربوط به کتاب «انسان و ایمان»، فایل‌های ۷ تا ۱۰ مربوط به «جهان‌بینی توحیدی» و فایل‌های ۱۱ تا ۱۹ مربوط به کتاب «وحی و نبوت» می‌باشند.

جدول ۴-۱: جدول A (جدول اطلاعات نویسندگان مجموعه) برای مجموعه‌متن CPPT

ردیف	نام نویسنده	بازه‌ی حیات (شمسی)	تعداد کل متون	نام کتاب	تعداد متون	شماره متون
۱	جلال آل احمد	۱۳۰۲ تا ۱۳۴۸	۳۱	سه تار	۱۳	۱ تا ۱۳
				پنج داستان	۵	۱۴ تا ۱۸
				داستان‌های سیاست	۹	۱۹ تا ۲۷
				داستان‌های آسمانی	۴	۲۸ تا ۳۱
۲	سید علی موسوی گرمارودی	۱۳۲۰ تا کنون	۳۱	داستان‌های پیامبران (جلد ۱)	۲۱	۱ تا ۲۱
				داستان‌های پیامبران (جلد ۲)	۱۰	۲۲ الی ۳۱

[1] Folder
[2] Text file

۳	شهید مرتضی مطهری	۱۲۹۸ تا ۱۳۵۸	۱۹	انسان و ایمان	۶	۱ تا ۶
				جهان‌بینی توحیدی	۴	۷ تا ۱۰
				وحی و نبوت	۹	۱۱ تا ۱۹
۴	رسول پرویزی	۱۲۹۸ تا ۱۳۵۶	۲۰	شلوارهای وصله‌دار	۲۰	۱ تا ۲۰
۵	دکتر علی شریعتی	۱۳۱۲ تا ۱۳۵۶	۱۴	کویر	۳	۱ تا ۳
				حج	۷	۴ تا ۱۰
				وصیت‌نامه	۱	۱۱
				آخرین نامه به پدر	۱	۱۲
				نامه به مهندس بازرگان	۱	۱۳
				نامه به فرزند	۱	۱۴
۶	دکتر عبدالحسین زرین‌کوب	۱۳۰۱ تا ۱۳۷۸	۳۰	از کوچه‌ی رندان	۱۲	۱ تا ۱۲
				بامداد اسلام	۱۸	۱۳ تا ۳۰

ب) جدول B: جدول اطلاعات آماری طول متون هر نویسنده

در این بخش طول متون مربوط به هر نویسنده، تحلیل آماری شده و مقدار کمینه، بیشینه و متوسط آن در جدول ۴-۲ درج شده است. هم‌چنین مجموع طول تمامی متون مربوط به هر نویسنده نیز ارائه گردیده است. طول هر متن با دو معیار مختلف تعداد کلمات و تعداد نویسه‌ها محاسبه شده و نتایج به صورت جداگانه گزارش شده اند.

جدول ۴-۲: جدول B (اطلاعات آماری طول متون هر نویسنده) برای مجموعه‌متن CPPT

ردیف	تعداد متن	طول متون بر حسب کلمه				طول متون بر حسب نویسه			
		کمینه	بیشینه	میانگین	مجموع	کمینه	بیشینه	میانگین	مجموع
۱	31	1,035	5,256	3,000.42	93,013	4,624	25,926	14,374.03	445,595
۲	31	292	7,062	2,517.32	78,037	1,427	34,919	12,186.26	377,774
۳	19	525	10,618	3,602.79	68,453	2,549	50,651	17,327.63	329,225
۴	20	875	5,005	2,174.70	43,494	3,973	22,593	9,874.80	197,496

۵	14	1,173	3,833	2,199.93	30,799	5,930	18,712	10,859.64	152,035
۶	30	1,368	10,153	4,344.27	130,328	6,293	48,043	20,334.10	610,023

از جدول ۴-۲ می‌توان نتایج زیر را استنباط نمود:

- اگر تعداد کلمات را به عنوان مبنا برای سنجش طول متن فرض کنیم:
  - متون مربوط به نویسنده‌ی شماره‌ی ۴ (پرویزی) کوتاه‌تر و متون مربوط به نویسنده‌ی شماره‌ی ۶ (دکتر زرین‌کوب) بلندتر از متون سایر نویسندگان است.
  - کوتاه‌ترین و بلندترین متون موجود در مجموعه به ترتیب دارای ۲۹۲ و ۱۰۶۱۸ کلمه می‌باشند. کوتاه‌ترین متن مربوط به نویسنده‌ی شماره ۲ (گرمارودی) و بلندترین آن مربوط به نویسنده‌ی شماره ۳ (شهید مطهری) است.
- از نظر تعداد متون جمع‌آوری‌شده برای هر نویسنده:
  - نویسنده‌ی شماره‌ی ۵ (دکتر شریعتی) دارای کمترین متن می‌باشد.
  - بیش‌ترین متن مربوط به نویسندگان شماره‌ی ۱ و ۲ (آل‌احمد و گرمارودی) می‌باشد (۳۱ متن).
- از نظر حجم متون گردآوری‌شده برای هر نویسنده (مجموع تعداد کلمات):
  - کم‌ترین متن باز هم متعلق به نویسنده‌ی شماره‌ی ۵ می‌باشد.
  - بیش‌ترین متن مربوط به نویسنده‌ی شماره‌ی ۶ (زرین‌کوب) است. هر چند تعداد متون این نویسنده (۳۰ متن) کمتر از نویسنده‌های ۱ و ۲ می‌باشد، اما از آن‌جا که عموماً متون این نویسنده بلندتر از سایر نویسندگان است، در مجموع حجم متون وی بیش‌تر از سایرین می‌باشد.

ج) جدول C: جدول اطلاعات آماری طول متون کل مجموعه‌متن:

در این جدول کمینه، بیشینه، متوسط و مجموع طول متون موجود در مجموعه‌متن درج شده است (جدول ۴-۳). همانند جدول قبل، محاسبه‌ی طول یک متن بر دو مبنا (تعداد کلمات و تعداد نویسه‌ها) صورت پذیرفته و نتایج به صورت جدا ارائه شده است. همانطور که در بخش ۴-۲-۲ بیان شد، در این قسمت فقط جدول C را ارائه داده و تحلیل آن را به بخش ۴-۵ مؤکول می‌کنیم.

جدول ۴-۳: جدول C (اطلاعات آماری طول متون کل مجموعه) برای مجموعه‌متن CPPT

کمیت	بر حسب کلمه	بر حسب نویسه
کمینه‌ی طول متن	292	1,427
بیشینه‌ی طول متن	10,618	50,651
متوسط طول متون	2,973.24	14,159.41
مجموع طول متون	444,124	2,112,148

د) جدول D: جدول اطلاعات آماری تعداد متون کل مجموعه‌متن:

اطلاعات مربوط به تعداد متون موجود در مجموعه‌متن CPPT در جدول ۴-۴ ارائه شده اند. همانند جدول قبل، تحلیل این جدول نیز در بخش ۴-۵ بیان خواهد شد.

جدول ۴-۴: جدول D (اطلاعات آماری تعداد متون مجموعه) برای مجموعه‌متن CPPT

کمینه‌ی تعداد متن مربوط به یک نویسنده	۱۴
بیشینه‌ی تعداد متن مربوط به یک نویسنده	۳۱
تعداد کل متون	۱۴۵
تعداد کل نویسندگان	۶
متوسط تعداد متن به ازای هر نویسنده	۲۴,۱۶۷

۴-۴-۴ مجموعه‌متن RSK-40

این مجموعه متن توسط رمضانی و همکاران جمع‌آوری شده و اولین‌بار در مقاله‌ی [۵۱] ارائه گردیده است. در این مجموعه، متونی از ۴۰ نویسنده‌ی معاصر فارسی‌زبان گردآوری شده است. از جمله‌ی این نویسندگان می‌توان به بزرگ علوی، جلال آل‌احمد، صادق چوبک و فروغ فرخزاد اشاره کرد. متأسفانه لیست کامل این ۴۰ نویسنده در دسترس نبوده و لذا نمی‌توان برای این مجموعه جدول A (جدول اطلاعات نویسندگان) را تشکیل داد. هر چند گردآورندگان این مجموعه نام خاصی برای آن انتخاب نکرده‌اند، اما به منظور روان‌تر

شدن متن و سهولت ارجاع به این مجموعه‌متن، در ادامه از آن با نام RSK-40 یاد می‌کنیم. عبارت RSK در واقع سرواژه‌ی حاصل از نام خانوادگی گردآورندگان این مجموعه (رمضانی، شیدایی و کاهانی) و عدد ۴۰ بیان‌گر تعداد نویسندگان موجود در آن می‌باشد.

جداول B، C و D برای این مجموعه به ترتیب در جداول شماره‌ی ۴-۵، ۴-۶ و ۴-۷ آمده اند. نحوه‌ی درج اطلاعات در این جداول، دقیقاً به همان شیوه‌ی بیان‌شده در بخش ۴-۴-۳ (مجموعه متن CPPT) می‌باشد و لذا از تکرار مجدد آن‌ها صرف نظر نموده، فقط جداول را ارائه می‌دهیم. مجدداً تأکید می‌شود که اطلاعات جداول C و D در بخش ۴-۵ جهت مقایسه‌ی مجموعه‌متن‌های معرفی‌شده مورد استفاده قرار خواهند گرفت.

جدول ۴-۵: جدول B (اطلاعات آماری طول متون هر نویسنده) برای مجموعه‌متن RSK-40

ردیف	تعداد متن	طول متون بر حسب کلمه				طول متون بر حسب نویسه			
		کمینه	بیشینه	میانگین	مجموع	کمینه	بیشینه	میانگین	مجموع
۱	5	918	975	945.6	4,728	4,093	4,303	4,189.40	20,947
۲	5	905	932	921.8	4,609	4,190	4,340	4,260.40	21,302
۳	4	772	956	902	3,608	3,376	4,265	4,004.00	16,016
۴	5	919	946	937.2	4,686	4,134	4,321	4,227.20	21,136
۵	5	963	1,020	983.2	4,916	4,293	4,539	4,417.60	22,088
۶	5	909	946	922.8	4,614	4,016	4,184	4,129.20	20,646
۷	5	957	971	963.8	4,819	4,163	4,334	4,265.40	21,327
۸	5	924	975	946.2	4,731	4,312	4,557	4,411.40	22,057
۹	5	921	943	931.2	4,656	4,106	4,374	4,226.40	21,132
۱۰	5	933	993	959.6	4,798	4,304	4,436	4,355.60	21,778
۱۱	5	945	977	957.8	4,789	4,204	4,361	4,279.60	21,398
۱۲	4	556	963	852.5	3,410	2,548	4,302	3,826.25	15,305
۱۳	5	970	1,074	1,006.60	5,033	4,196	4,633	4,424.40	22,122
۱۴	5	892	930	910.2	4,551	4,022	4,311	4,171.00	20,855
۱۵	5	886	934	909.2	4,546	4,090	4,329	4,195.20	20,976

20,918	4,183.60	4,223	4,156	4,592	918.4	928	912	5	١٦
23,178	4,635.60	4,911	4,511	5,029	1,005.80	1,065	987	5	١٧
21,415	4,283.00	4,363	4,209	4,800	960	980	943	5	١٨
17,796	3,559.20	4,185	1,323	3,903	780.6	917	280	5	١٩
21,761	4,352.20	4,423	4,277	4,771	954.2	965	935	5	٢٠
21,220	4,244.00	4,328	4,124	4,667	933.4	945	925	5	٢١
21,808	4,361.60	4,469	4,290	4,887	977.4	1,002	964	5	٢٢
16,007	4,001.75	4,351	3,124	3,655	913.75	986	721	4	٢٣
22,096	4,419.20	4,599	4,277	4,901	980.2	1,005	963	5	٢٤
20,102	4,020.40	4,094	3,966	4,441	888.2	898	875	5	٢٥
22,442	4,488.40	4,631	4,374	4,870	974	1,019	953	5	٢٦
21,451	4,290.20	4,398	4,194	4,844	968.8	991	956	5	٢٧
20,944	4,188.80	4,229	4,137	4,662	932.4	941	925	5	٢٨
20,425	4,085.00	4,138	4,034	4,597	919.4	932	908	5	٢٩
18,139	4,534.75	5,106	3,726	4,146	1,036.50	1,171	848	4	٣٠
21,143	4,228.60	4,281	4,153	4,601	920.2	932	907	5	٣١
22,536	4,507.20	4,703	4,353	4,985	997	1,036	959	5	٣٢
22,299	4,459.80	4,614	4,334	4,793	958.6	986	931	5	٣٣
21,761	4,352.20	4,612	4,201	4,770	954	1,021	909	5	٣٤
15,822	3,955.50	4,844	1,877	3,220	805	971	360	4	٣٥
22,646	4,529.20	4,684	4,413	4,971	994.2	1,029	968	5	٣٦
21,319	4,263.80	4,328	4,167	4,750	950	959	942	5	٣٧
20,284	4,056.80	4,164	3,953	4,338	867.6	889	842	5	٣٨
23,050	4,610.00	4,835	4,423	4,908	981.6	1,006	950	5	٣٩
20,422	4,084.40	4,154	4,027	4,481	896.2	908	880	5	٤٠

جدول ۴-۶: جدول C (اطلاعات آماری طول متون کل مجموعه) برای مجموعه‌متن RSK-40

کمیت	بر حسب کلمه	بر حسب نویسه
متن با طول کمینه	280	1,323
متن با طول بیشینه	1,171	5,106
متوسط طول متون	937.929	4,251.96
مجموع طول متون	183,076	830,069

جدول ۴-۷: جدول D (اطلاعات آماری تعداد متون مجموعه) برای مجموعه‌متن RSK-40

کمینه‌ی تعداد متن مربوط به یک نویسنده	۴
بیشینه‌ی تعداد متن مربوط به یک نویسنده	۵
تعداد کل متون	۱۹۵
تعداد کل نویسندگان	۴۰
متوسط تعداد متن به ازای هر نویسنده	۴,۸۷۵

۵-۴-۴ مجموعه‌متن BASU

این مجموعه‌متن توسط فرهمندپور و همکاران در دانشگاه بوعلی سینای همدان گردآوری شده و برای اولین‌بار در مقاله‌ی [۴۹] ارائه گردیده است. به منظور گردآوری این مجموعه، از ۲۰ نفر از دانشجویان ترم اول کارشناسی مهندسی دانشگاه بوعلی سینا خواسته شده تا در مورد موضوعی خاص (دانشگاه) متونی با طول حداقل ۲۰۰۰ کلمه بنویسند. فرهمندپور و همکاران مجموعه‌متن گردآوری‌شده را مجموعه‌ی «دانشگاه بوعلی سینا[1]» نامیدند. در ادامه‌ی این پایان‌نامه از این مجموعه‌متن با عنوان BASU یاد می‌شود.

[1] BuAliSina University (BASU)

از آنجا که نویسندگان موجود در این مجموعه، دانشجویان دانشگاه بوعلی بوده و اطلاعات خاصی از آن‌ها در اختیار نمی‌باشد، لذا تشکیل جدول A برای این مجموعه تا حدی بی‌معنی است. در این مجموعه برای هر نویسنده فقط یک متن در اختیار داریم، لذا تحلیل آماری بر روی متون تک تک نویسنده‌ها عملی نبوده و جدول B نیز قابل تشکیل نمی‌باشد. به همین دلیل برای این مجموعه فقط جداول C و D ارائه می‌گردند (به ترتیب در جداول ۴-۸ و ۴-۹).

جدول ۴-۸: جدول C (اطلاعات آماری طول متون کل مجموعه) برای مجموعه‌متن BASU

کمیت	بر حسب کلمه	بر حسب نویسه
متن با طول کمینه	2,010	10,031
متن با طول بیشینه	4,808	23,840
متوسط طول متون	2,952.35	14,861.45
مجموع طول متون	59,047	297,229

جدول ۴-۹: جدول D (اطلاعات آماری تعداد متون مجموعه) برای مجموعه‌متن BASU

کمینه‌ی تعداد متن مربوط به یک نویسنده	۱
بیشینه‌ی تعداد متن مربوط به یک نویسنده	۱
تعداد کل متون	۲۰
تعداد کل نویسندگان	۲۰
متوسط تعداد متن به ازای هر نویسنده	۱

سؤالی که ممکن است در ذهن خواننده ایجاد شده باشد آنست که با توجه به وجود یک متن برای هر نویسنده، چگونه می‌توان متون آموزشی و آزمایشی را تشکیل داد؟ گردآوردندگان مجموعه‌ی BASU راهکار جالبی به منظور تشکیل متون آموزشی و آزمایشی ارائه داده اند. در این راهکار برای هر متن مراحل زیر را باید طی کرد:

- ابتدا فقط ۲۰۰۰ کلمه‌ی اول متن را حفظ، و مابقی را حذف می‌کنیم.
- سپس این متن ۲۰۰۰ کلمه‌ای را به چهار قسمت ۵۰۰ کلمه‌ای تقسیم می‌کنیم.
- در نهایت از هر نویسنده، سه قسمت ۵۰۰ کلمه‌ای را به عنوان نمونه‌های آموزشی و قسمت چهارم را به عنوان نمونه‌ی آزمایشی در نظر می‌گیریم.

در این صورت مجموعه‌ای در اختیار خواهیم داشت که در آن به ازای هر نویسنده، چهار متن با طول دقیقاً ۵۰۰ کلمه در دسترس است. برای تفاوت قائل شدن بین این مجموعه و مجموعه‌ی اصلی (که برای هر نویسنده یک متن داریم)، مجموعه‌ی اصلاح‌شده را اصطلاحاً «دانشگاه بوعلی سینا – نسخه‌ی دوم» نامیده و با نماد BASU-2 نمایش می‌دهیم.

به نظر می‌رسد دلیل انتخاب ۲۰۰۰ کلمه به عنوان مبنا آنست که کوتاه‌ترین متن موجود در مجموعه دارای ۲۰۱۰ کلمه بوده (جدول ۴-۸) و در صورت انتخاب مقدار بزرگتر (مثلاً ۲۵۰۰ کلمه، ۵ قسمت ۵۰۰ کلمه‌ای)، قادر به تولید قسمت‌های پنجم به بالاتر برای این متن نبودیم.

در ادامه جداول B، C و D برای مجموعه‌متن BASU-2 به ترتیب در جداول ۴-۱۰، ۴-۱۱ و ۴-۱۲ ارائه می‌شوند:

جدول ۴-۱۰: جدول B (اطلاعات آماری طول متون هر نویسنده) برای مجموعه‌متن BASU-2

ردیف	تعداد متن	طول متون بر حسب کلمه				طول متون بر حسب نویسه			
		کمینه	بیشینه	میانگین	مجموع	کمینه	بیشینه	میانگین	مجموع
۱	4	500	500	500	2,000	2,406	2,502	2,463.25	9,853
۲	4	500	500	500	2,000	2,459	2,545	2,508.00	10,032
۳	4	500	500	500	2,000	2,566	2,609	2,581.00	10,324
۴	4	500	500	500	2,000	2,560	2,609	2,587.25	10,349
۵	4	500	500	500	2,000	2,535	2,612	2,558.00	10,232
۶	4	500	500	500	2,000	2,450	2,670	2,541.00	10,164
۷	4	500	500	500	2,000	2,387	2,446	2,418.50	9,674
۸	4	500	500	500	2,000	2,456	2,545	2,503.25	10,013
۹	4	500	500	500	2,000	2,509	2,591	2,542.25	10,169
۱۰	4	500	500	500	2,000	2,510	2,562	2,530.50	10,122

۱۱	4	500	500	500	2,000	2,437	2,508	2,481.25	9,925
۱۲	4	500	500	500	2,000	2,578	2,649	2,629.75	10,519
۱۳	4	500	500	500	2,000	2,454	2,554	2,517.50	10,070
۱۴	4	500	500	500	2,000	2,432	2,515	2,476.25	9,905
۱۵	4	500	500	500	2,000	2,467	2,616	2,544.00	10,176
۱۶	4	500	500	500	2,000	2,458	2,570	2,502.00	10,008
۱۷	4	500	500	500	2,000	2,456	2,582	2,539.25	10,157
۱۸	4	500	500	500	2,000	2,413	2,560	2,494.00	9,976
۱۹	4	500	500	500	2,000	2,427	2,491	2,474.00	9,896
۲۰	4	500	500	500	2,000	2,489	2,566	2,524.50	10,098

جدول ۴-۱۱: جدول C (اطلاعات آماری طول متون کل مجموعه) برای مجموعه‌متن BASU-2

کمیت	بر حسب کلمه	بر حسب نویسه
متن با طول کمینه	500	2,387
متن با طول بیشینه	500	2,670
متوسط طول متون	500.00	2,520.78
مجموع طول متون	40,000	201,662

جدول ۴-۱۲: جدول D (اطلاعات آماری تعداد متون مجموعه) برای مجموعه‌متن BASU-2

کمینه‌ی تعداد متن مربوط به یک نویسنده	۴
بیشینه‌ی تعداد متن مربوط به یک نویسنده	۴
تعداد کل متون	۸۰
تعداد کل نویسندگان	۲۰
متوسط تعداد متن به ازای هر نویسنده	۴

۶-۴-۴ مجموعه‌متن AAAT

این مجموعه‌متن توسط اوآمر و سیود گردآوری شده و اولین‌بار در مقاله‌ی [۹۰] معرفی گردیده است. این مجموعه‌متن بر خلاف سایر مجموعه‌های معرفی‌شده در این فصل، دارای متونی با زبان عربی می‌باشد. این مجموعه شامل متونی از ۱۰ سفرنامه‌نویس کهن عربی بوده و در آن از هر نویسنده سه متن مجزا گردآوری شده است. گردآورندگان این مجموعه آن را به صورت «تخصیص نویسنده‌ی متون کهن عربی[1]» نام‌گذاری نموده و آن را به اختصار با نماد AAAT نمایش می‌دهند.

خوشبختانه پدیدآورندگان این مجموعه، هنگام معرفی AAAT اطلاعات مفیدی در مورد نویسندگان موجود در مجموعه ارائه داده‌اند و لذا جدول A برای این مجموعه قابل تشکیل بوده و در جدول ۴-۱۳ نمایش داده شده است.

جدول ۴-۱۳: جدول A (جدول اطلاعات نویسندگان مجموعه) برای مجموعه‌متن AAAT

ردیف	نام نویسنده	نام کتاب	تاریخ نگارش کتاب
۱	ابن بطوطه	رحله ابن بطوطه	۱۳۲۵ تا ۱۳۵۲ میلادی
۲	ابن جبیر	رحله ابن جبیر	۱۱۸۲ تا ۱۱۸۵ میلادی
۳	ناصر خسرو	سفرنامه ناصر خسرو (ترجمه‌شده از فارسی به عربی)	۱۰۴۵ میلادی
۴	ابن فضلان	رحله ابن فضلان	۹۲۱ میلادی
۵	ابن المجاور	تاریخ المستبصر	۱۲۳۳ میلادی
۶	الیوسی	المحاضرات فی اللغه و الادب	۱۶۸۴ میلادی
۷	لسان‌الدین ابن الخطیب	خطره الطیف فی رحله الشتاء و الصیف	۱۶۸۴ میلادی
۸	الآلوسی	غرائب الاغتراب	۱۸۵۲ میلادی
۹	محب‌الدین الحموی	حادی الأظعان النجدیه إلی دیار المصریه	۱۵۴۲ تا ۱۶۰۸ میلادی
۱۰	البلوی	تاج المفرق فی تحلیه علماء المشرق	چند سال قبل از ۱۳۶۴ میلادی

[1] Author-attribution of Ancient Arabic Texts (AAAT)

نکته‌ی جالب توجه، وجود یک کتاب ترجمه‌شده (سفرنامه‌ی ناصرخسرو، ترجمه‌شده از فارسی به عربی) در این مجموعه می‌باشد. چنین موردی کمتر در سایر مجموعه‌متن‌ها به چشم می‌خورد. به عبارت دیگر استفاده از متون ترجمه‌ای در مجموعه‌متن‌های تخصیص نویسنده چندان رایج نیست، زیرا متن ترجمه‌شده در واقع ترکیبی از سبک نوشتاری نویسنده‌ی اصلی و مترجم را در خود دارد و نمی‌توان آن را کاملاً به مترجم تخصیص داد.

جداول B، C و D برای این مجموعه‌متن به ترتیب در جداول شماره‌ی ۴-۱۴، ۴-۱۵ و ۴-۱۶ ارائه شده اند. نحوه‌ی درج اطلاعات در این سه جدول همانند روند سابق استفاده‌شده برای سایر مجموعه‌ها (روند بخش‌های ۴-۴-۳ تا ۴-۴-۵) است.

جدول ۴-۱۴: جدول B (اطلاعات آماری طول متون هر نویسنده) برای مجموعه‌متن AAAT

ردیف	تعداد متن	طول متون بر حسب کلمه				طول متون بر حسب نویسه			
		کمینه	بیشینه	میانگین	مجموع	کمینه	بیشینه	میانگین	مجموع
۱	3	307	632	515.333	1,546	1,748	3,481	2,901.67	8,705
۲	3	534	595	565.667	1,697	2,979	3,354	3,184.67	9,554
۳	3	290	652	516.667	1,550	1,566	3,423	2,793.00	8,379
۴	3	580	589	585	1,755	3,137	3,358	3,241.00	9,723
۵	3	458	717	562	1,686	2,173	3,343	2,673.67	8,021
۶	3	500	626	560	1,680	2,643	3,434	3,021.00	9,063
۷	3	451	594	525.667	1,577	2,695	3,445	3,141.67	9,425
۸	3	513	655	582.333	1,747	2,728	3,482	3,188.33	9,565
۹	3	317	630	496.333	1,489	1,842	3,478	2,835.00	8,505
۱۰	3	343	578	423	1,269	1,937	3,351	2,449.33	7,348

جدول ۴-۱۵: جدول C (اطلاعات آماری طول متون کل مجموعه) برای مجموعه‌متن AAAT

کمیت	بر حسب کلمه	بر حسب نویسه
متن با طول کمینه	290	1,566
متن با طول بیشینه	717	3,482
متوسط طول متون	533.20	2,942.93
مجموع طول متون	15,996	88,288

جدول ۴-۱۶: جدول D (اطلاعات آماری تعداد متون مجموعه) برای مجموعه‌متن AAAT

کمینه‌ی تعداد متن مربوط به یک نویسنده	۳
بیشینه‌ی تعداد متن مربوط به یک نویسنده	۳
تعداد کل متون	۳۰
تعداد کل نویسندگان	۱۰
متوسط تعداد متن به ازای هر نویسنده	۳

فصل پنجم

آزمایش‌ها، نتایج و تحلیل آن‌ها

۵-۱ مقدمه

در ابتدای این فصل مرور کوتاهی بر مطالب ارائه‌شده در فصول قبلی پایان‌نامه داریم:

- در فصل سوم، هشت مورد از مهم‌ترین روش‌های NDP تشریح شد.
- در ابتدای فصل چهام، دو روش جدید برای حل مسائل تخصیص نویسنده پیشنهاد گردید. این دو روش نیز در دسته‌ی روش‌های NDP جای می‌گیرند.
- در انتهای فصل چهارم، چهار مجموعه‌متن استفاده‌شده در این پایان‌نامه معرفی شدند.

پس در مجموع در این پایان‌نامه، ده روش NDP و چهار مجموعه‌متن در اختیار داریم. در این فصل می‌خواهیم این ده روش را بر روی چهار مجموعه‌متنِ موجود اعمال نموده و ضمن ارائه‌ی نتایجِ حاصل، تحلیلی از آن‌ها نیز ارائه دهیم. بدین‌منظور در این فصل چهار «آزمایش[1]» اصلی صورت گرفته است که هر کدام از آن‌ها شامل اعمال تمامی ده روش NDP بر روی یکی از چهار مجموعه‌متنِ موجود می‌باشند. به عنوان مثال در آزمایش اول (بخش ۵-۲)، ده روش NDP مطالعه‌شده بر روی مجموعه‌متن CPPT اعمال گردیده و نتایج بدست‌آمده ارائه شده‌اند.

مطالب بیان‌شده در این فصل به صورت زیر سازماندهی شده‌اند:

- ابتدا در پیوست ۶ (بخش ۶-۱) پرسش‌های بنیادینی که می‌بایست طی این فصل به آن‌ها پاسخ دهیم، مطرح می‌شوند. در واقع نتایج تمامی پژوهش‌های صورت‌گرفته در پایان‌نامه، در قالب پاسخ به سؤالات بنیادین خلاصه می‌شود.
- سپس در پیوست ۷ به بررسی نحوه‌ی ارائه‌ی یک آزمایش پرداخته شده است. این بخش در واقع یک روند استاندارد برای ارائه‌ی آزمایش‌ها معرفی نموده و تعیین می‌کند که هنگام ارائه‌ی یک آزمایش، چه اطلاعاتی می‌بایست به خواننده داده شود.
- پس از آن، در بخش‌های ۵-۲ تا ۵-۵، چهار آزمایش اصلی انجام‌شده در این پایان‌نامه (با همان روند استاندارد معرفی‌شده در بخش ۵-۲) ارائه می‌شوند.
- در نهایت و در پیوست ۶ (بخش ۶-۲) به پرسش‌های بنیادین مطرح‌شده ۲ پاسخ داده می‌شود.

---

[1] Experiment

۵-۲ آزمایش اول: مجموعه‌متن CPPT

هدف این آزمایش که بر روی مجموعه‌متنِ گردآوری‌شده توسط نگارنده صورت پذیرفته، سنجش قدرت روش‌های NDP در حل مسائل تخصیص نویسنده‌ی متون فارسی می‌باشد.

۵-۲-۱ مشخصات آزمایش

الف) مجموعه‌متن

در این آزمایش از مجموعه‌متن CPPT استفاده شده است که در کل شامل ۱۴۵ متن از ۶ نویسنده‌ی معاصر فارسی‌زبان می‌باشد. در رتبه‌بندی مجموعه‌متن‌ها بر مبنای سطح دشواری (که در فصل چهارم ارائه شد)، این مجموعه کمترین سطح دشواری را به خود اختصاص داده و لذا انتظار می‌رود دقت بالایی در این آزمایش حاصل شود.

ب) مجموعه‌های آموزش و آزمایش

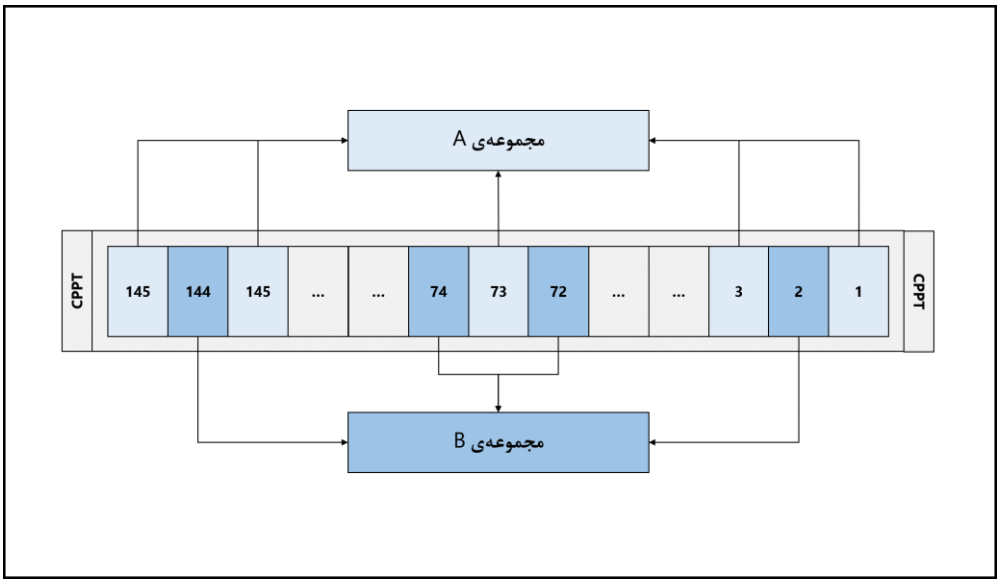
ما به منظور تشکیل مجموعه‌های آموزش و آزمایش در مجموعه‌متن CPPT، روند خاصی را پیشنهاد می‌دهیم؛ بدین صورت که ابتدا تمامی متون موجود در مجموعه را از ۱ تا ۱۴۵ شماره‌گذاری نموده و سپس مجموعه‌های A و B را به صورت زیر تشکیل می‌دهیم (شکل ۵-۳):

- مجموعه‌ی A: مجموعه‌ای شامل متون دارای شماره‌ی فرد
- مجموعه‌ی B: مجموعه‌ای شامل متون دارای شماره‌ی زوج

هنگام آزمایش هر یک از ده روش موجود، به ازای هر جفت از پارامترهای N و L می‌بایست دو فاز مجزا را اجرا کنیم:

- فاز اول: مجموعه‌ی A را به عنوان مجموعه‌ی آموزش و مجموعه‌ی B را به عنوان مجموعه‌ی آزمایش انتخاب نموده و سپس دقت روش مورد نظر را محاسبه می‌نماییم.
- فاز دوم: این بار مجموعه‌ی B را به عنوان مجموعه‌ی آموزش در نظر گرفته و مجموعه‌ی A را به عنوان مجموعه‌ی آزمایش انتخاب می نماییم. سپس دقت روش مورد نظر را محاسبه می‌کنیم.

در نهایت میانگین دقت‌های بدست‌آمده در فازهای اول و دوم را به عنوان دقت روش مذکور در نظر می‌گیریم.



شکل ۵-۱: تشکیل مجموعه‌های A و B از مجموعه‌متن CPPT

۵-۲-۲ نتایج آزمایش

نتایج این آزمایش (به همان صورتی که در بخش ۵-۲-۱ بیان شد)، در جدول ۵-۱ ارائه شده‌اند:

جدول ۵-۱: جدول نتایج آزمایش بر روی مجموعه‌متن CPPT

ردیف	نام روش	بالاترین دقت (درصد)	جفت‌های بهینه	
			L	N
۱	CNG	99.31	1600	9
۲	CNG-SPI	98.62	1000 , 1700 , 1800 , 1900 , 2000	8
			1600	9
۳	CNG-D1	98.62	2800 , 2900 , 3000	9
			2800	10
۴	CNG-D2	100	1600 , 1700 , 2000 , 2100, 2200	9
۵	CNG-WPI	97.24	1900 , 2000 , 2100	8

۶	RLP	99.31	1700	8
۷	RLP-IAF	99.31	1600 , 1700	8
۸	O-SPI	99.31	1500	9
۹	CNG-WIS	97.93	1900-2800	8
۱۰	VNG	97.93	1400-1900 , 2100 , 2900 , 3000	8

۵-۲-۳ تحلیل نتایج

الف) رتبه‌بندی روش‌ها

عملکرد روش‌های استفادشده در این آزمایش را می‌توان به صورت جدول ۵-۲ رتبه‌بندی نمود. در این جدول هر دو نوع رتبه‌بندی معرفی‌شده در بخش ۵-۲-۳ ارائه شده‌اند: رتبه‌بندی نوع اول (بر مبنای بالاترین دقت) و رتبه‌بندی نوع دوم (بر مبنای بالاترین دقت و تعداد جفت‌های بهینه).

جدول ۵-۲: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی CPPT

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	CNG-D2	100	5
۲	۲	RLP	99.31	2
	۳	CNG		1
	۳	RLP-IAF		1
	۳	O-SPI		1
۳	۴	SPI	98.62	6
	۵	CNG-D1		4
۴	۶	CNG-WIS	97.93	10
	۷	VNG		9
۵	۸	WPI	97.24	3

همانطور که در جدول فوق ملاحظه می‌شود، در آزمایش اول عموم روش NDP مطالعه‌شده عملکرد مطلوبی از خود ارائه می‌دهند. در این بین، روش CNG-D2 عملکرد بسیار مناسبی داشته و با دقتی معادل ۱۰۰ درصد رتبه‌ی نخست را به خود اختصاص داده است.

ب) ارزیابی روش‌های پیشنهادی

اگر دو روش پیشنهادی (روش CNG-WIS و روش VNG) را به صورت جداگانه (مستقل از عملکرد سایر روش‌ها) مورد بررسی قرار دهیم، با توجه به دقت حاصل‌شده توسط این دو روش (97.93%)، می‌توان نتیجه گرفت این روش‌ها قادرند به خوبی مسئله‌ی تخصیص نویسنده‌ی مربوط به آزمایش اول را حل نمایند.

اما اگر دو روش پیشنهادی را با سایر روش‌های مطالعه‌شده مقایسه نماییم، در رتبه‌های ۶ و ۷ قرار می‌گیرند که چندان رضایت‌بخش نمی‌باشد. البته می‌بایست توجه داشت که اختلاف دقت آن‌ها با روش قرار گرفته در مرتبه‌ی دوم (یعنی روش RLP) فقط 1.38% است.

ج) بازه‌ی معمول پارامترهای بهینه

اگر کمینه و بیشینه‌ی مقادیر پارامتر N موجود در جدول نتایج (جدول ۵-۲) را به ترتیب با نمادهای $N_{\min}$ و $N_{\max}$ نمایش داده و هم‌چنین $L_{\min}$ و $L_{\max}$ به ترتیب بیانگر مقدار کمینه و بیشینه‌ی پارامتر L در این جدول باشند، آن‌گاه می‌توان گفت:

Normal-Range in CPPT (Persian): $N_{\min} \le N \le N_{\max}\ ,\ L_{\min} \le L \le L_{\max}$ (۱-۵)

با این توصیف، با توجه به نتایج ارائه‌شده در جدول ۵-۱، بازه‌ی معمول برای پارامترهای N و L در آزمایش اول به صورت زیر است:

Normal-Range in CPPT (Persian): $8 \le N \le 10\ ,\ 1000 \le L \le 3000$ (۲-۵)

۵-۳ آزمایش دوم: مجموعه‌متن RSK-40

این آزمایش که بر روی مجموعه‌متن RSK-40 انجام شده است، در واقع عملکرد روش‌های NDP را در حالتی که تعداد نویسندگان موجود در مجموعه بسیار زیاد باشند، مورد ارزیابی قرار می‌دهد.

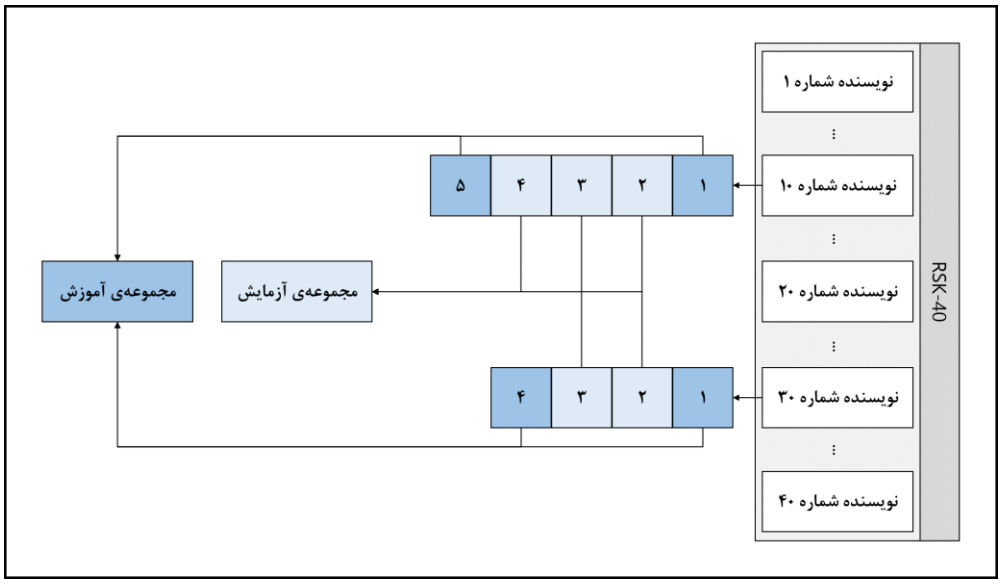
۵-۳-۱ مشخصات آزمایش

الف) مجموعه‌متن

این آزمایش بر روی مجموعه‌متن RSK-40 صورت پذیرفته که دارای ۱۹۵ متن از ۴۰ نویسنده‌ی معاصر فارسی‌زبان است. این مجموعه در رتبه‌بندی سطح دشواری در مکان سوم قرار داشته و لذا می‌توان انتظار دقت نسبتاً بالایی از آن داشت.

ب) مجموعه‌های آموزش و آزمایش

متأسفانه گردآورندگان RSK-40 روند مشخصی برای تشکیل مجموعه‌های آموزش و آزمایش ارائه نداده‌اند. لذا می‌بایست روندی جدید جهت تشکیل این دو مجموعه پیشنهاد دهیم. از فصل چهارم به یاد داریم که در مجموعه‌متن RSK-40، برای هر نویسنده ۴ یا ۵ متن وجود دارد. در این آزمایش متون ابتدایی و انتهایی هر نویسنده را در مجموعه‌ی آموزش و سایر متون را در مجموعه‌ی آزمایش قرار می‌دهیم (شکل ۵-۴). لازم به ذکر است که این آزمایش بر خلاف آزمایش اول (که شامل دو فاز بود)، تنها دارای یک فاز می‌باشد.



شکل ۵-۲: تشکیل مجموعه‌های آموزش و آزمایش از مجموعه‌متن RSK-40

۵-۳-۲ نتایج آزمایش

نتایج حاصل از این آزمایش در جدول ۵-۳ آورده شده است.

جدول ۵-۳: نتایج حاصل از آزمایش بر روی مجموعه‌متن RSK-40

ردیف	نام روش	بالاترین دقت (درصد)	جفت‌های بهینه	
			L	N
۱	CNG	100	1000 , 1500 , 1900 , 2000 , 2100	4
			1200	5
۲	CNG-SPI	100	1000 , 1500 , 1900 , 2000	4
			1200	5
۳	CNG-D1	100	1500 , 1700 , 1900 , 2000	4
			900	5
۴	CNG-D2	100	1900	3
			600 , 800-1000 , 1200 , 1400-2000	4
			700-900 , 1100-1900	5
			1000 , 1100 , 1300-2000	6
			1700-1900	7
۵	CNG-WPI	98.26	800	4
			700 , 1000 , 1100	5
۶	RLP	99.13	1000 , 1500	4
۷	RLP-IAF	99.13	1000 , 1200	4
۸	O-SPI	100	1200	4
			800 , 900 , 1100	5
۹	CNG-WIS	100	2000-2900	4
			1900-3000	5
			1900-3000	7
۱۰	VNG	99.13	900 , 1000	2

۵-۳-۳ تحلیل نتایج

الف) رتبه‌بندی روش‌ها

ده روش استفاده‌شده در این آزمایش را بر مبنای دقت عملکرد می‌توان به صورت جدول ۵-۴ رتبه‌بندی نمود:

جدول ۵-۴: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی RSK-40

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	CND-D2	100	38
	۲	CNG-WIS		34
	۳	CNG		6
	۴	CNG-SPI		5
	۴	CNG-D1		5
	۵	O-SPI		4
۲	۶	VNG	99.13	2
	۶	RLP		2
	۶	RLP-AIF		2
۳	۷	WPI	98.26	4

همانطور که در جدول فوق ملاحظه می‌شود، در آزمایش دوم نیز عموم روش NDP مطالعه‌شده عملکرد مطلوبی از خود ارائه می‌دهند. هم‌چنین در این آزمایش نیز روش CNG-D2 عملکرد بسیار مناسبی داشته و با دقتی معادل ۱۰۰ درصد مجدد رتبه‌ی نخست را به خود اختصاص داده است.

ب) ارزیابی روش‌های پیشنهادی

اگر دو روش پیشنهادی را به صورت جداگانه مورد بررسی قرار دهیم، با توجه به دقت‌های حاصل‌شده توسط این دو روش (100% و 99.13%)، می‌توان نتیجه گرفت این روش‌ها در حل مسئله‌ی تخصیص نویسنده‌ی مربوط به آزمایش دوم موفق عمل کرده‌اند.

اگر دو روش پیشنهادی را با سایر روش‌های مطالعه‌شده مقایسه نماییم، می‌توان گفت:

- اولین روش پیشنهادی (CNG-WIS) با دقت ۱۰۰ درصد و قرار گرفتن در رتبه‌ی ۲، عملکرد بسیار مناسبی از خود ارائه داده است. می‌بایست توجه داشت که این روش توانسته است عملکردی بسیار نزدیک به روش CNG-D2 از خود ارائه دهد و فقط چهار جفت بهینه از آن کمتر دارد.
- دومین روش پیشنهادی (VNG) با قرار گرفتن در رتبه‌ی ۶ عملکرد متوسطی از خود ارائه داده است. البته بایست توجه داشت عملکرد روش VNG در این آزمایش، با عملکرد روش RLP (که در آزمایش اول در رتبه‌ی ۲ قرار گرفته بود) یکسان می‌باشد.

ج) بازه‌ی معمول پارامترهای بهینه

با توجه به نتایج ارائه‌شده در جدول ۵-۳، در این آزمایش بازه‌ی معمول برای پارامترهای N و L به صورت زیر است:

$$\text{Normal-Range in RSK-40 (Persian):}\ \ 2 \le N \le 7\ ,\ 700 \le L \le 2900 \qquad (۵-۳)$$

۵-۴ آزمایش سوم: مجموعه‌متن BASU-2

هدف اصلی این آزمایش که بر روی مجموعه‌متن BASU-2 صورت پذیرفته، در واقع ارزیابی عملکرد روش‌های NDP در حل مسائل تخصیص نویسنده بر روی متون محاوره‌ای می‌باشد.

۵-۴-۱ مشخصات آزمایش

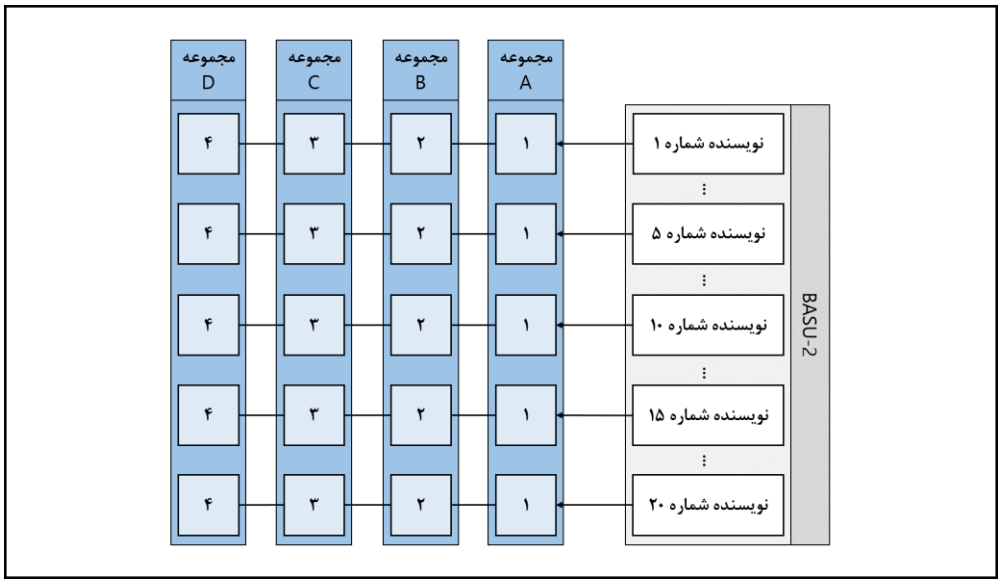
الف) مجموعه‌متن

در این آزمایش از مجموعه‌متن BASU-2 استفاده شده است که دارای ۸۰ متن از ۲۰ نفر از دانشجویان دانشگاه بوعلی می‌باشد (از هر نویسنده چهار متن). این مجموعه‌متن در رتبه‌بندی فصل چهارم به عنوان دشوارترین مجموعه انتخاب شده و لذا نمی‌بایست انتظار دقت چندان بالایی در آن داشت. با توجه به

محاوره‌ای بودن متون موجود در این مجموعه، با استفاده نتایج حاصل از این آزمایش می‌توان عملکرد روش‌های NDP را بر روی متون محاوره‌ای (و نه کتابی) ارزیابی نمود.

ب) مجموعه‌های آموزش و آزمایش

همانطور که در فصل چهارم بیان شد، روند پیشنهادشده توسط گردآورندگان مجموعه‌ی BASU-2 جهت تشکیل مجموعه‌های آموزش و آزمایش بدین‌صورت است که: در هر بار آزمایش، یکی از چهار متن هر نویسنده به در مجموعه‌ی آزمایش قرار گرفته و سه متن باقیمانده در مجموعه‌ی آموزش قرار می‌گیرند. به عبارت دیگر ابتدا می‌بایست طبق شکل ۵-۵، چهار مجموعه‌ی A، B، C و D را تشکیل دهیم. این مجموعه‌ها به ترتیب شامل متون اول، دوم، سوم و چهارم تمامی نویسنده‌ها می‌باشند. سپس می‌بایست چهار فاز را به صورت جدا اجرا نموده و دقت هر فاز به طور مجزا محاسبه کنیم. در هر فاز یکی از مجموعه‌ها به عنوان مجموعه‌ی آزمایش در نظر گرفته شده و مجموعه‌ی آموزش از ادغام سایر مجموعه‌ها حاصل می‌شود. به عنوان مثال در فاز نخست مجموعه‌ی A را به عنوان مجموعه‌ی آزمایش در نظر گرفته و مجموعه‌های B، C و D را با هم ادغام می‌کنیم تا مجموعه‌ی آموزش تشکیل شود. در انتها دقت نهایی با میانگین‌گیری بر روی نتایج بدست‌آمده در هر فاز محاسبه می‌شود.



شکل ۵-۳: تشکیل مجموعه‌های A، B، C و D از مجموعه‌متن BASU-2

۵-۴-۲ نتایج آزمایش

نتایج بدست‌آمده از آزمایش سوم در جدول ۵-۵ ارائه شده‌اند:

جدول ۵-۵: نتایج حاصل از آزمایش بر روی مجموعه‌متن BASU-2

ردیف	نام روش	بالاترین دقت (درصد)	پارامترهای بهینه	
			L	N
۱	CNG	85	1700	4
۲	CNG-SPI	85	1500	4
۳	CNG-D1	85	1700	3
			1800 , 3000	4
۴	CNG-D2	88.75	3000	4
۵	CNG-WPI	80	2900	4
۶	RLP	78.75	1000 , 1100 , 1200	3
۷	RLP-IAF	85	1000	3
۸	O-SPI	82.5	1500	3
۹	CNG-WIS	82.5	3000	5
۱۰	VNG	68.75	3000	3
			1300	5

۵-۴-۳ تحلیل نتایج

الف) رتبه‌بندی روش‌ها

ده روش استفادشده در این آزمایش را بر مبنای دقت عملکرد می‌توان به صورت جدول ۵-۶ رتبه‌بندی نمود:

جدول ۵-۶: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی RSK-40

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	CNG-D2	88.75	1
۲	۲	CNG-D1	85	3
	۳	CNG		1
	۳	CNG-SPI		1
	۳	RLP-IAF		1
۳	۴	CNG-WIS	82.5	1
	۴	O-SPI		1
۴	۵	CNG-WPI	80	1
۵	۶	RLP	78.75	3
۶	۷	VNG	68.75	1

همانطور که در جدول فوق ملاحظه می‌شود، در آزمایش سوم نیز با وجود سطح دشواری بالا، عموم روش NDP مطالعه‌شده عملکرد مطلوبی از خود ارائه می‌دهند. همان‌طور که در فصل دوم بیان شد گردآورندگان مجموعه‌متن BASU-2 با استفاده از روش‌های مبتنی بر سیستم‌های NLP، در بهترین حالت به دقتی معادل ۸۰ درصد رسیده‌اند؛ اما در این پایان‌نامه و با استفاده از روش‌های NDP دقتی معادل 88.75% کسب شده است که بیانگر بهبودی به اندازه‌ی 8.75% می‌باشد. هم‌چنین می‌بایست به این نکته نیز توجه نمود که در این آزمایش نیز (همانند دو آزمایش قبل) روش CNG-D2 بهترین عملکرد را ارائه داده و در رتبه‌ی نخست قرار گرفته است.

ب) ارزیابی روش‌های پیشنهادی

اگر روش پیشنهادی اول (CNG-WIS) را به صورت جداگانه مورد بررسی قرار دهیم، با توجه به دقت حاصل‌شده (82.5%) و توجه به این نکته که بالاترین دقت حاصل از روش‌های مبتنی بر سیستم‌های NLP برابر با ۸۰ درصد بوده است، می‌توان نتیجه گرفت این روش توانایی حل مسائل تخصیص نویسنده‌ی متون

محاوره‌ای فارسی را نیز دارد، زیرا توانسته است بهبودی به اندازه‌ی 2.5% ایجاد نماید. اما بررسی جداگانه‌ی روش پیشنهادی دوم (VNG) حاکی از عملکرد ضعیف آن در حل چنین مسائلی می‌باشد.

اگر دو روش پیشنهادی را با سایر روش‌های مطالعه‌شده مقایسه نماییم، می‌توان گفت:

- اولین روش پیشنهادی (CNG-WIS) با قرار گرفتن در رتبه‌ی چهارم عملکرد متوسطی از خود ارائه داده است.
- دومین روش پیشنهادی (VNG) با قرارگرفتن در رتبه‌ی ۷ (آخرین رتبه) عملکرد نامناسبی به نمایش گذاشته است!

با جمع‌بندی این نتایج می‌توان چنین گفت که:

- روش CNG-WIS قادر است مسائل تخصیص نویسنده‌ی فارسیِ محاوره‌ای را به خوبی حل نماید.
- روش VNG در حل مسائل زبان فارسیِ محاوره‌ای قدرت چندانی نداشته و در این‌گونه مسائل ضعیف عمل می‌نماید.

ج) بازه‌ی معمول پارامترهای بهینه

با توجه به نتایج ارائه‌شده در جدول ۵-۵، در این آزمایش بازه‌ی معمول برای پارامترهای N و L به صورت زیر است:

$$\text{Normal-Range in BASU-2 (Persian): } 3 \le N \le 5 \ , \ 1000 \le L \le 3000 \qquad (۴-۵)$$

۵-۵ آزمایش چهارم: مجموعه‌متن AAAT

این آزمایش که در آن از مجموعه‌متن AAAT استفاده شده، قدرت روش‌های NDP را در حل مسائل تخصیص نویسنده‌ی زبان عربی می‌سنجد.

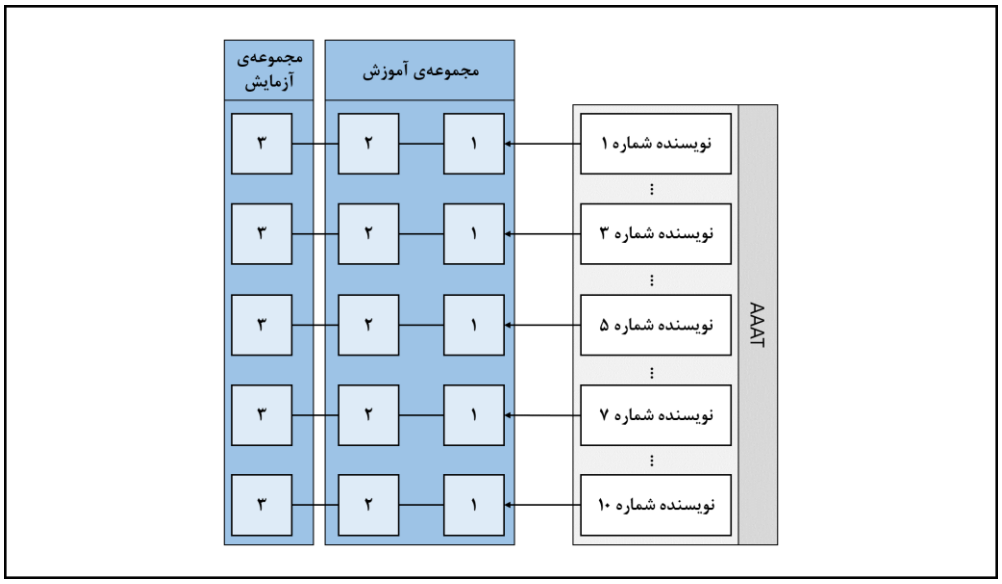
۵-۵-۱ مشخصات آزمایش

الف) مجموعه‌متن

در این آزمایش مجموعه‌متن AAAT مورد استفاده قرار گرفته که دارای متونی با زبان عربی می‌باشد. در این مجموعه ۱۰ نویسنده و از هر نویسنده ۳ متن وجود دارند. در رتبه‌بندی فصل چهارم، این مجموعه در مکان دومِ دشوارترین مجموعه‌ها قرار گرفته و لذا انتظار نمی‌رود دقت چندان بالایی در آن حاصل شود.

ب) مجموعه‌های آموزش و آزمایش

در این آزمایش مجموعه‌های آموزش و آزمایش طبق شیوه‌ی ارائه‌شده توسط گردآورندگان مجموعه تشکیل شده‌اند. طبق پیشنهاد گردآورندگان، مجموعه‌های آموزش و آزمایش بدین‌صورت تعیین می‌گردند: برای هر نویسنده دو متن اول را در مجموعه‌ی آموزش و متن سوم را در مجموعه‌ی آزمایش قرار می‌دهیم (شکل ۵-۶). لذا این آزمایش نیز همانند آزمایش دوم، تنها دارای یک فاز است.



شکل ۵-۴: تشکیل مجموعه‌های آموزش و آزمایش از مجموعه‌متن AAAT

۵-۵-۲ نتایج آزمایش

جدول نتایج برای این آزمایش به شرح زیر است:

جدول ۵-۷: نتایج حاصل از آزمایش بر روی مجموعه‌متن AAAT

ردیف	نام روش	بالاترین دقت (درصد)	پارامترهای بهینه	
			L	N
۱	CNG	80	1700 , 1800	3
			200 , 400-700	4
			1100 , 1300-3000	5
			1200-1400 , 1600 , 1700 , 1900-2300 , 3000	6
			2300-2600 , 2900 , 3000	7
			2600 , 2900 , 3000	8
۲	CNG-SPI	80	1800 , 1900	3
			200 , 400 , 500 , 700 ,	4
			1100-1300 , 1700-3000	5
			1400 , 1600-2300 , 3000	6
			2300-2600 , 2900 , 3000	7
			2600 , 2900 , 3000	8
۳	CNG-D1	80	50-3000	1
			2100	3
			200 , 400-700 , 3000	4
			1100 , 1300-3000	5
			1300 , 1600 , 1700 , 1900-2300 , 3000	6
			2300-2600 , 2900 , 3000	7
			2600 , 2900 , 3000	8

۴	CNG-D2	90	1700 , 2200-2800 , 2900 , 3000	4
۵	CNG-WPI	90	3000	4
			2800-3000	5
۶	RLP	80	500-700 , 1600-1800	4
			2000-2400	5
۷	RLP-IAF	90	1750	3
			800 , 900 , 1500 , 1750 , 200	4
۸	O-SPI	80	200 , 400-700	4
۹	CNG-WIS	80	300 , 400 , 800-2800	4
			1500-3000	5
۱۰	VNG	90	1000 , 1400-1600 , 1800-2100 , 2500 , 2600	4
			200 , 1100	5

۳-۵-۵ تحلیل نتایج

همانند آزمایش‌های قبلی، تحلیل این آزمایش نیز در سه بخش صورت می‌پذیرد:

الف) رتبه‌بندی روش‌ها

ده روش استفادشده در این آزمایش را بر مبنای دقت عملکرد می‌توان این‌گونه رتبه‌بندی نمود:

جدول ۵-۸: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی AAAT

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	VNG	90	12
	۲	CNG-D2		9
	۳	RLP-IAF		6

	۴	CNG-WPI		4
۲	۵	CNG-D1	80	75
	۶	CNG		46
	۷	CNG-SPI		42
	۸	CNG-WIS		39
	۹	RLP		10
	۱۰	O-SPI		5

همانطور که در جدول فوق ملاحظه می‌شود، بر روی مجموعه‌متن زبان عربی نیز با وجود سطح دشواری بالا، عموم روش NDP مطالعه‌شده عملکرد مطلوبی از خود ارائه می‌دهند. آن‌طور که در فصل دوم بیان شد گردآورندگان مجموعه‌متن AAAT با استفاده از روش‌های مبتنی بر سیستم‌های NLP، در بهترین حالت به دقتی معادل ۸۰ درصد رسیده‌اند؛ اما در این پایان‌نامه و با استفاده از روش‌های NDP دقتی معادل 90% کسب شده است که بیانگر بهبودی معادل 10% است.

ب) ارزیابی روش‌های پیشنهادی

اگر روش پیشنهادی اول (CNG-WIS) را به صورت جداگانه مورد بررسی قرار دهیم، با توجه به دقت حاصل‌شده (80%) و هم‌چنین در نظر گرفتن این موضوع که بالاترین دقت حاصل از روش‌های مبتنی بر سیستم‌های NLP برابر با ۸۰ درصد بوده است، می‌توان نتیجه گرفت این روش قادر به حل مسائل تخصیص نویسنده‌ی متون عربی نیز می‌باشد.

اگر دو روش پیشنهادی را با سایر روش‌های مطالعه‌شده مقایسه نماییم، می‌توان گفت:

- اولین روش پیشنهادی (CNG-WIS) با قرار گرفتن در رتبه‌ی ۸ عملکرد تقریباً متوسطی (متوسط رو به ضعیف) از خود ارائه داده است. البته می‌بایست توجه داشت که تمامی روش‌های قرار گرفته در رتبه‌های ۵ تا ۱۰ دارای دقت یکسانی بوده و تنها در تعداد جفت‌های بهینه متفاوت می‌باشند؛ لذا عملکرد این روش چندان هم نامناسب نمی‌باشد.

- روش VNG عملکردی بسیار عالی داشته و با ارائه‌ی دقتی معادل 90% به ازای ۱۲ جفت از پارامترها، در رتبه‌ی اول قرار گرفته است، حتی بالاتر از روش CNG-D2 (که در هر سه آزمایش اول در رتبه‌ی نخست قرار گرفته بود)!

ج) بازه‌ی معمول پارامترهای بهینه

با توجه به نتایج ارائه‌شده در جدول ۵-۷، در این آزمایش بازه‌ی معمول برای پارامترهای N و L به صورت زیر است:

Normal-Range in AAAT (Arabic): $1 \le N \le 8 \ , \ 50 \le L \le 3000$ (۵-۵)

فصل ششم

مطالعات موردی در حکایات و غزلیات

۶-۱ مقدمه

در این فصل می‌خواهیم دو مسئله‌ی خاص در ادبیات فارسی را مطرح نموده و توانایی روش‌های NDP در حل آن‌ها را ارزیابی نماییم: نظیره‌های گلستان و اشعار سبک هندی. شرح بیشتر این دو حوزه‌ی ادبیاتی و اطلاعات تکمیلی در مورد آن‌ها در پیوست ۸ آمده است.

به منظور جلوگیری از طولانی شدن این فصل، در آزمایش‌های انجام‌شده فقط ۴ روش از ۱۰ روش موجود را بررسی نموده‌ایم: CNG (پایه و مبنای تمامی روش‌ها)، CNG-D2 (روش دارای بهترین عملکرد روی آزمایش‌های زبان فارسی در فصل پنجم) و دو روش پیشنهادی در این پایان‌نامه (CNG-WIS و VNG).

۶-۲ نظیره‌های گلستان

۶-۲-۱ گردآوری مجموعه‌متن

از میان انبوه نظیره‌هایی که تا کنون بر گلستان نوشته شده‌اند، «بهارستان» و «پریشان» را انتخاب نموده و مجموعه‌متنی با نام «گلستان، بهارستان و پریشان» و نماد اختصاری GBP گردآوری نمودیم. همانطور که در جدول ۶-۱ آمده، برای گردآوری این مجموعه، از هر کتاب ۲۵ حکایت به عنوان نمونه انتخاب شده است.

جدول ۶-۱: اطلاعات مربوط به مجموعه‌متن GBP

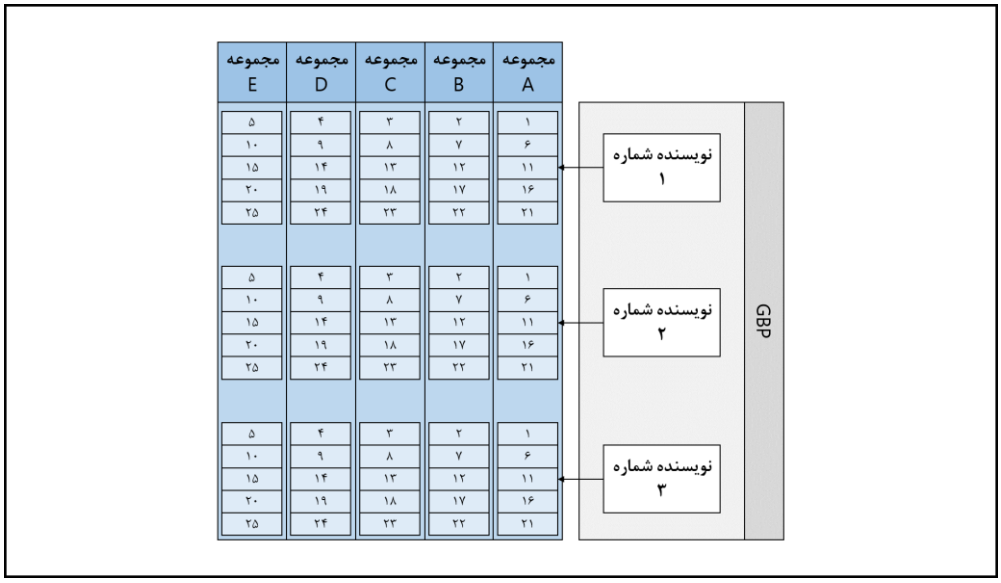
ردیف	نام اثر	نویسنده	قرن حیات	تعداد متن
۱	گلستان	سعدی	۷	۲۵
۲	بهارستان	جامی	۹	۲۵
۳	پریشان	قاآنی	۱۳	۲۵

۶-۲-۲ آزمایش و نتایج

پس از گردآوری مجموعه‌متن، می‌بایست نحوه‌ی تشکیل مجموعه‌های آموزش و آزمایش را تعیین نماییم. پیشنهاد می‌کنیم این دو مجموعه بدین‌صورت تشکیل شوند:

۱) ابتدا متون مربوط به هر نویسنده را از ۱ تا ۲۵ شماره‌گذاری نموده و سپس پنج مجموعه‌ی A، B، C، D و E را به صورت زیر تشکیل می‌دهیم (شکل ۶-۱):
   - مجموعه‌ی A: شامل متون ۱، ۶، ۱۱، ۱۶ و ۲۱ از هر سه نویسنده
   - مجموعه‌ی B: شامل متون ۲، ۷، ۱۲، ۱۷ و ۲۲ از هر سه نویسنده
   - مجموعه‌ی C: شامل متون ۳، ۸، ۱۳، ۱۸ و ۲۳ از هر سه نویسنده
   - مجموعه‌ی D: شامل متون ۴، ۹، ۱۴، ۱۹ و ۲۴ از هر سه نویسنده
   - مجموعه‌ی E: شامل متون ۵، ۱۰، ۱۵، ۲۰ و ۲۵ از هر سه نویسنده
۲) سپس آزمایش در پنج فاز جداگانه انجام می‌دهیم. در هر فاز یکی از پنج مجموعه‌ی تشکیل‌شده را به عنوان مجموعه‌ی آزمایش در نظر گرفته و سایر مجموعه‌ها را در کنار هم قرارمی‌دهیم تا مجموعه‌ی آموزش ایجاد شود.

دقت نهایی با میانگین‌گیری بر روی دقت‌های پنج فازِ انجام‌شده محاسبه می‌شود.



شکل ۶-۱: نحوه‌ی تشکیل مجموعه‌های A تا E از روی مجموعه‌متن GBP

نتایج این آزمایش در جدول ۶-۲ نمایش داده شده است.

جدول ۶-۲: نتایج حاصل از آزمایش بر روی مجموعه‌متن GBP

ردیف	نام روش	بالاترین دقت (درصد)	پارامترهای بهینه	
			L	N
۱	CNG	97.33	1200 , 1700-1900 , 2300-2600	3
۲	CNG-D2	97.33	1300	3
			3000	4
۳	CNG-WIS	97.33	1200-2600	3
۴	VNG	98.66	1900-3000	5
			2100-2700 , 2900 , 3000	6

عملکرد چهار روش بررسی‌شده در این آزمایش را می‌توان به صورت جدول ۶-۳ رتبه‌بندی نمود:

جدول ۶-۳: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی GBP

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	VNG	98.66	21
۲	۲	WIS	97.33	15
	۳	CNG		8
	۴	CNG-D2		2

۶-۲-۳ تحلیل نتایج

با استفاده از جدول ۶-۳ می‌توان تحلیل‌های زیر را ارائه داد:

۱) روش‌های NDP استفاده‌شده در این آزمایش توانسته‌اند با دقت مناسبی مسئله‌ی تخصیص نویسنده‌ی حکایات فارسی را حل نمایند؛ لذا می‌توان این فرضیه را مطرح نمود که روش‌های NDP علاوه بر متون نثر، بر روی حکایات نیز عملکرد مناسبی از خود ارائه می‌دهند. البته برای اثبات این

فرضیه نیاز به آزمایش‌های بیش‌تری بر روی چندین مجموعه‌متن متشکل از حکایات فارسی می‌باشد.

۲) روش‌های پیشنهادی (CNG-WIS و VNG) عملکرد بسیار خوبی از خود ارائه داده و در رتبه‌های اول و دوم قرار گرفته‌اند.

۳) روش CNG-D2 که در آزمایش‌های متون نثر فارسی (سه آزمایش اول فصل پنجم) همواره در رتبه‌ی نخست قرار داشت، در این آزمایش (که بر روی حکایات فارسی) بدترین عملکرد را بین ۴ روش بررسی‌شده از خود ارائه می‌دهد. این نتیجه تأکید دیگری‌ست بر این موضوع که نمی‌توان ادعا نمود که یک روش NDP، همواره و در تمامی آزمایش‌ها عملکردی بهتر از سایر روش‌ها دارد.

۶-۳ غزلیات سبک هندی

۶-۳-۱ گردآوری مجموعه‌متن

از میان شاعرانی که در سبک هندی شعر سروده‌اند، «صائب تبریزی»، «بیدل دهلوی» و «حزین لاهیجی» را انتخاب نموده و مجموعه‌متنی با نام «صائب، بیدل و حزین» و نماد اختصاری SBH گردآوری نمودیم. اطلاعات مربوط به این مجموعه‌متن در جدول ۶-۱ آمده است. لازم به ذکرست که در جمع آوری اشعار به گونه‌ای عمل شد تا علاوه بر سبک، قالب اشعار گردآوری شده نیز یکسان باشد. می‌بایست توجه داشت که از هرکدامِ این شعرا، ۳۰ غزل در مجموعه‌ی SBH وجود دارد.

جدول ۶-۴: اطلاعات مربوط به مجموعه‌متن SBH

ردیف	شاعر	قرن حیات	سبک	قالب	تعداد متن
۱	صائب تبریزی	۱۱	هندی	غزل	۳۰
۲	بیدل دهلوی	۱۱ و ۱۲	هندی	غزل	۳۰
۳	حزین لاهیجی	۱۲	هندی	غزل	۳۰

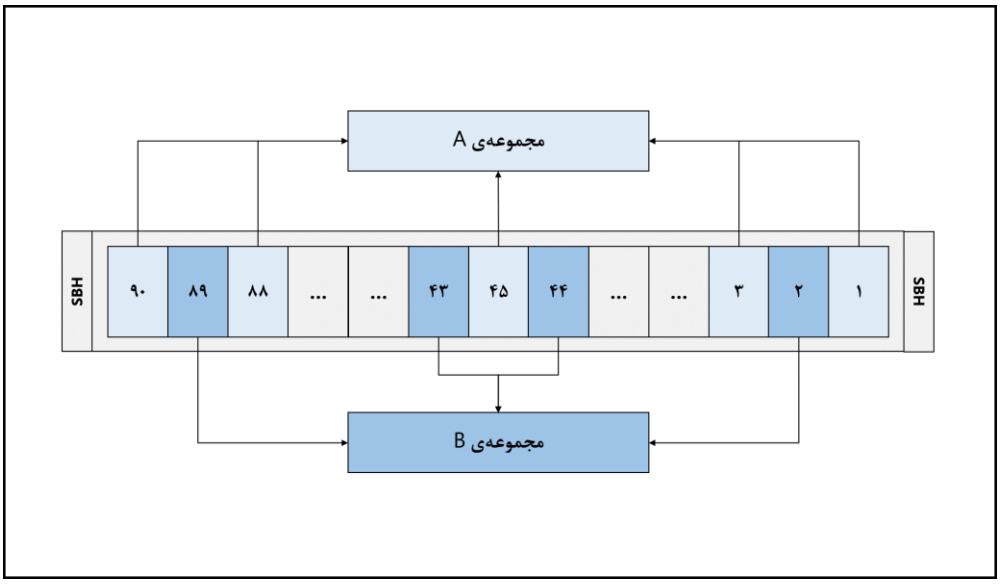
۶-۳-۲ آزمایش و نتایج

پس از گردآوری مجموعه‌متن، می‌بایست نحوه‌ی تشکیل مجموعه‌های آموزش و آزمایش را تعیین نماییم. پیشنهاد می‌کنیم این دو مجموعه بدین‌صورت تشکیل شوند:

۱) ابتدا متون موجود در مجموعه از ۱ تا ۹۰ شماره‌گذاری شده و سپس دو مجموعه‌ی A و B مطابق شکل ۶-۲ تشکیل شوند؛ یعنی متون با شماره‌ی فرد در مجموعه‌ی A قرار گرفته و متون با شماره‌ی زوج در مجموعه‌ی B.

۲) پس از تشکیل دو مجموعه‌ی A و B، آزمایش در دو فاز مختلف صورت می‌پذیرد. در هر فاز یکی از دو مجموعه‌ی تشکیل‌شده به عنوان مجموعه‌ی آزمایش و دیگری به عنوان مجموعه‌ی آموزش در نظر گرفته می‌شود.

دقت نهایی با میانگین‌گیری بر روی دقت‌های دو فاز انجام‌شده محاسبه می‌شود.



شکل ۶-۲: نحوه‌ی تشکیل مجموعه‌های A و B از روی مجموعه‌متن SBH

نتایج این آزمایش در جدول ۶-۵ نمایش داده شده است.

جدول ۶-۵: نتایج حاصل از آزمایش بر روی مجموعه‌متن SBH

ردیف	نام روش	بالاترین دقت (درصد)	پارامترهای بهینه	
			L	N
۱	CNG	98.88	1000 , 1200 , 1300 , 2200	3
			800 , 2000 , 2500-2700 , 2900	4
			700-1000	5
۲	CNG-D2	100	2800-3000	5
۳	CNG-WIS	100	700-1600 , 2000	4
۴	VNG	97.77	500-3000	3

عملکرد چهار روش بررسی‌شده در این آزمایش را می‌توان به صورت جدول ۶-۶ رتبه‌بندی نمود:

جدول ۶-۶: رتبه‌بندی عملکرد روش‌ها بر روی مجموعه‌ی SBH

رتبه نوع اول	رتبه نوع دوم	نام روش	بالاترین دقت (درصد)	تعداد جفت‌های بهینه
۱	۱	CNG-WIS	100	11
	۲	CNG-D2		3
۲	۳	CNG	98.88	14
۳	۴	VNG	97.77	26

۶-۳-۳ تحلیل نتایج

با توجه به جدول ۶-۶ می‌توان تحلیل‌های زیر را ارائه داد:

۱) روش‌های NDP استفاده‌شده در این آزمایش توانسته‌اند عملکرد بسیار مناسبی در حل مسئله‌ی تخصیص نویسنده‌ی اشعار فارسی ارائه دهند؛ لذا فرضیه‌ی دیگری را نیز می‌توان مطرح نمود:

روش‌های NDP علاوه بر متون نثر و حکایات، بر روی اشعار نیز عملکرد مناسبی از خود ارائه می‌دهند. البته اثبات این فرضیه نیز نیازمند آزمایش‌ها و پژوهش‌های بیش‌تری است.

۲) روش‌های پیشنهادی اول (CNG-WIS) با ارائه‌ی عملکردی بسیار خوب در رتبه‌ی نخست قرار گرفته است؛ اما روش پیشنهادی دوم (VNG) عملکرد ضعیفی نسبت به سایر روش‌ها داشته و در رتبه‌ی ۴ قرار گرفته است.

۳) روش CNG-D2 که در آزمایش قبل در رتبه‌ی آخر قرار گرفته بود، در این آزمایش مجدداً عملکرد مناسبی داشته و در رتبه‌ی ۲ قرار گرفته است.

فصل هفتم

نتیجه‌گیری و پیشنهادات

۱-۷ فعالیت‌ها و یافته‌های پایان‌نامه

آنچه در این پایان‌نامه انجام گرفت را می‌توان به صورت زیر خلاصه نمود:

۱) بررسی کوتاه و دسته‌بندی صورت‌مسئله‌های مختلف حوزه‌ی پردازش متن. (فصل ۱)
۲) بررسی کوتاه انواع مسائل تخصیص نویسنده. (فصل ۲)
۳) بررسی و دسته‌بندی انواع روش‌های حل مسائل تخصیص نویسنده و مقایسه‌ی آن‌ها. (فصل ۲)
۴) بررسی هشت مورد از مهم‌ترین روش‌های NDP جهت حل مسائل تخصیص نویسنده. (فصل ۳)
۵) گردآوری یک مجموعه‌متن (CPPT) شامل ۱۴۵ متن از ۶ نویسنده‌ی معاصر فارسی‌زبان. (فصل ۴)
۶) پیشنهاد دو روش جدید در حوزه‌ی تخصیص نویسنده (CNG-WIS و VNG). (فصل ۴)
۷) اعمال ده مورد از روش‌های NDP بر روی چهار مجموعه‌متن متفاوت و بررسی نتایج بدست‌آمده. (فصل ۵)
۸) ارائه‌ی بازه‌ی معمول پیشنهادی برای پارامترهای N و L در روش‌های NDP برای چهار مجموعه‌متن موردآزمایش. (فصل ۵)
۹) گردآوری دو مجموعه‌متن خاص در حوزه‌ی ادبیات فارسی (فصل ۶):
   - مجموعه‌متن GBP: شامل ۷۵ حکایت (ترکیبی از نثر و نظم) از سه نویسنده.
   - مجموعه‌متن SBH: شامل ۹۰ شعر (با سبک هندی و قالب غزل) از سه شاعر.
۱۰) اعمال چهار مورد از روش‌های NDP بر روی مجموعه‌متن‌های GBP و SBH و بررسی نتایج بدست‌آمده. (فصل ۶)

یافته‌های این پایان‌نامه به شرح زیر می‌باشند:

۱) در حوزه‌ی تخصیص نویسنده‌ی متونی با زبان فارسی:

- اگر نویسنده‌ها صاحب‌سبک باشند، روش‌های NDP عملکرد بسیار مناسبی از خود نشان می‌دهند. به عنوان مثال برخی از روش‌های NDP روی مجموعه‌متن‌های CPPT و RSK-40 (که نویسندگان موجود در هر دو مجموعه، نویسندگانی صاحب‌سبک هستند)، دارای عملکردی خیره‌کننده با دقت ۱۰۰ درصد می‌باشند.
- اگر نویسنده‌ها صاحب‌سبک نباشند، عملکرد روش‌های NDP نسبت به حالت قبل دچار افت شده، اما هم‌چنان عملکرد آن‌ها قابل قبول است. به عنوان مثال روش‌های NDP روی

مجموعه‌متن BASU-2 (که نویسندگان آن دانشجویان مهندسی دانشگاه بوعلی می‌باشند) در بهترین حالت دقتی نزدیک به ۹۰ درصد را ارائه می‌دهند.

۲) در حوزه‌ی تخصیص نویسنده‌ی متونی با زبان عربی نیز روش‌های NDP عملکرد نسبتاً مناسبی از خود ارائه می‌دهند. هر چند با توجه به اینکه در این پایان‌نامه فقط از یک مجموعه‌متن عربی استفاده شده است (AAAT)، نمی‌توان نتایج آن را به کل متون عربی تعمیم داد. به منظور اظهار نظر جامع در مورد زبان عربی، نیاز به پژوهش‌های بیش‌تری می‌باشد.

۳) موفقیت روش CNG-WIS حاکی از آنست که ترتیب قرار گرفتن انگرام‌ها در پروفایل می‌تواند اطلاعات مفیدی جهت حل مسائل تخصیص نویسنده ارائه دهد.

۴) موفقیت روش VNG بیان‌گر آنست که علاوه بر انگرام‌های پرتکرار، انگرام‌های پراکنده (فصل چهارم) نیز در حل مسائل تخصیص نویسنده سودمند می‌باشند.

۵) با استفاده از نتایج بدست‌آمده در آزمایش‌های فصل پنجم و میانگین‌گیری بر روی بازه‌های پیشنهادی بر روی آزمایش‌های زبان فارسی (سه آزمایش اول)، می‌توان گستره‌های زیر را به عنوان بازه‌ی معمول برای پارامترهای N (طول انگرام) و L (طول پروفایل) در زبان فارسی پیشنهاد داد:

$$\text{Normal-Range in Persian: } 4 \le N \le 8$$

$$\text{Normal-Range in Persian: } 900 \le L \le 3000$$

البته این دو بازه نیاز به ارزیابی دقیق‌تری دارند که آن را به پژوهش‌های آینده واگذار می‌نماییم.

۷-۲ پیشنهادات

در انتهای این پایان‌نامه، پیشنهادات زیر جهت کارها و پژوهش‌های آیندگان ارائه می‌شود:

۱) ایجاد یک پایگاه اینترنتی برای متون فارسی:

در تحقیقات حوزه‌ی تخصیص نویسنده‌ی زبان انگلیسی معمولاً از پایگاه اینترنتی «گوتِنبرگ[1]» جهت گردآوری مجموعه‌متن جدید استفاده می‌شود. این پایگاه متون بیش از ۴۶ هزار کتاب را در خود جای داده است که به صورت رایگان می‌توان آن‌ها را بارگیری[2] نمود. حال پیشنهاد می‌شود چنین پایگاهی برای متون فارسی نیز تشکیل شود تا پژوهش‌گران حوزه‌ی تخصیص نویسنده‌ی متون فارسی بتوانند به راحتی مجموعه‌متنی با ویژگی‌های موردنظر خود را تشکیل داده و روش‌های مختلف را بر روی آن آزمایش نمایند.

۲) توسعه‌ی روش‌های مبتنی بر اندیس:

روش CNG-WIS که در فصل چهارم پیشنهاد شد، در واقع به جای فرکانس از اندیس انگرام‌های درون پروفایل جهت محاسبه‌ی فاصله استفاده می‌نماید. حال همین ایده (یعنی شرکت دادن اندیس انگرام‌ها در محاسبه‌ی فاصله) را می‌توان بر روی سایر روش‌های ارائه شده در فصل سوم پیاده نمود. به عنوان مثال می‌توان با ایجاد تغییراتی در روش RLP، آن را بگونه‌ای تغییر داد که اندیس انگرام‌ها در فاصله تأثیرگذار باشند.

۳) توسعه‌ی روش‌های مبتنی بر انگرام‌های پراکنده:

روش VNG که در فصل چهارم پیشنهاد شد از دو جهت قابل توسعه است:

- در این روش برای تعیین انگرام‌های پراکنده از کمیت «واریانس» استفاده نمودیم، اما می‌توان کمیت‌های دیگری را جهت تعیین میزان پراکندگی انگرام‌ها به کار برد.
- در این روش برای محاسبه‌ی فاصله‌ی میان دو پروفایل از معیار فاصله‌ی کسینوسی استفاده کردیم، اما می‌توان معیارهای دیگری را جهت محاسبه‌ی فاصله به کار برد.

---

[1] http://www.gutenberg.org/
[2] Download

۴) گسترش مجموعه‌ی پارامترهای مورداستفاده در آزمایش:

در آزمایش‌های فصل پنجم از گستره‌ی وسیعی از پارامترهای N و L استفاده شد:

$$1 \le N \le 10$$
$$50 \le L \le 3000$$

هر چند گسترش بازه‌ی N به مقادیر بیش از ۱۰ چندان معقول به نظر نمی‌رسد، اما بازه‌ی L را می‌توان گسترش داد؛ به ویژه پیشنهاد می‌شود در پژوهش‌های آینده بازه‌ی $3000 \le L \le 5000$ نیز مورد آزمایش قرار گیرد.

۵) آزمودن روش O-SPI به ازای مقادیر آستانه‌ی غیر از یک:

در آزمایش‌های فصل پنجم همواره از روش O-SPI به ازای $rc_{th} = 1$ استفاده شد (یعنی انگرام‌هایی که تعداد تکرار آن‌ها برابر ۱ بود حذف می‌شدند). حال پیشنهاد می‌شود در تحقیقات آینده این روش به ازای مقادیر مختلف rc_{th} نیز آزموده شود.

پیوست‌ها

پیوست ۱:

مشخصات تکمیلی حوزه‌ی تخصیص نویسنده

۱-۱ بررسی دو قالب اصلی

با توجه به نویسنده‌های موجود در مجموعه‌متن و نویسنده‌ی واقعی متن دیده‌نشده دو قالب می‌توان برای مسئله‌ی تخصیص نویسنده در نظر گرفت [۲۲]:

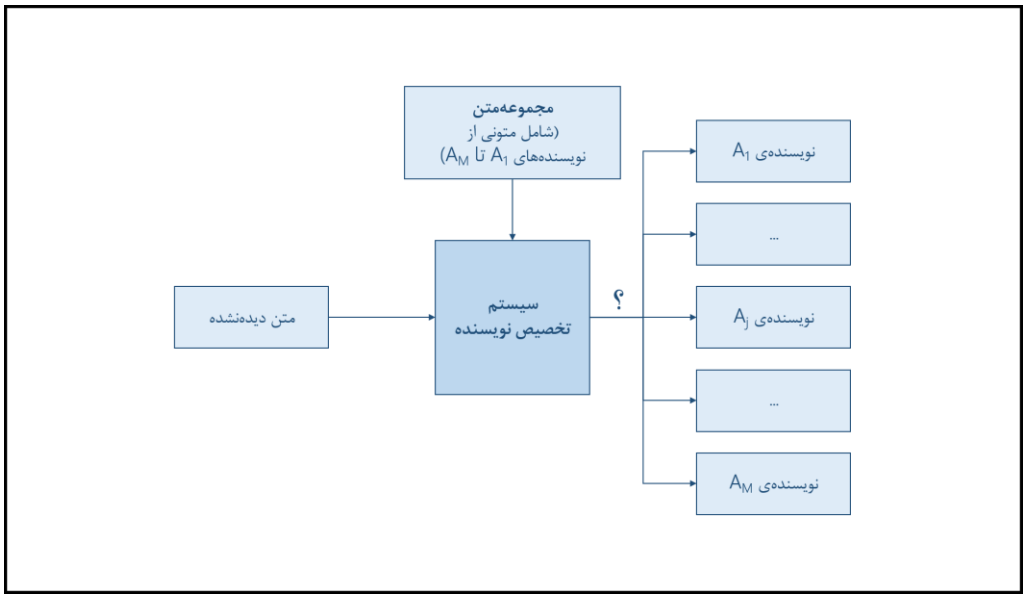
- **تخصیص مجموعه-بسته**[1]: در این حالت مطمئن هستیم که نویسنده‌ی واقعی متن دیده‌نشده، یکی از نویسنده‌های موجود در مجموعه‌متن است. لذا به ازای هر متن دیده‌نشده، سیستم AAS یکی از نویسنده‌های موجود در مجموعه‌متن به عنوان نویسنده‌ی واقعی اعلام می‌نماید. یک نمای کلی از چنین مسئله‌ای در شکل ۱-۶ نمایش داده شده است.

- **تخصیص مجموعه-باز**[2]: در این نوع، الزامی وجود ندارد که نویسنده‌ی واقعی متن دیده‌نشده، یکی از نویسنده‌های موجود در مجموعه‌متن باشد. لذا به ازای هر متن دیده‌نشده، سیستم AAS یا یکی از نویسنده‌های موجود در مجموعه‌متن به عنوان نویسنده‌ی واقعی اعلام می‌نماید، و یا اعلام می‌کند که متن دیده‌نشده توسط هیچ‌کدام از نویسنده‌های موجود در مجموعه‌متن نوشته نشده است. شکل ۱-۷ یک نمونه از این مسئله را نمایش می‌دهد.

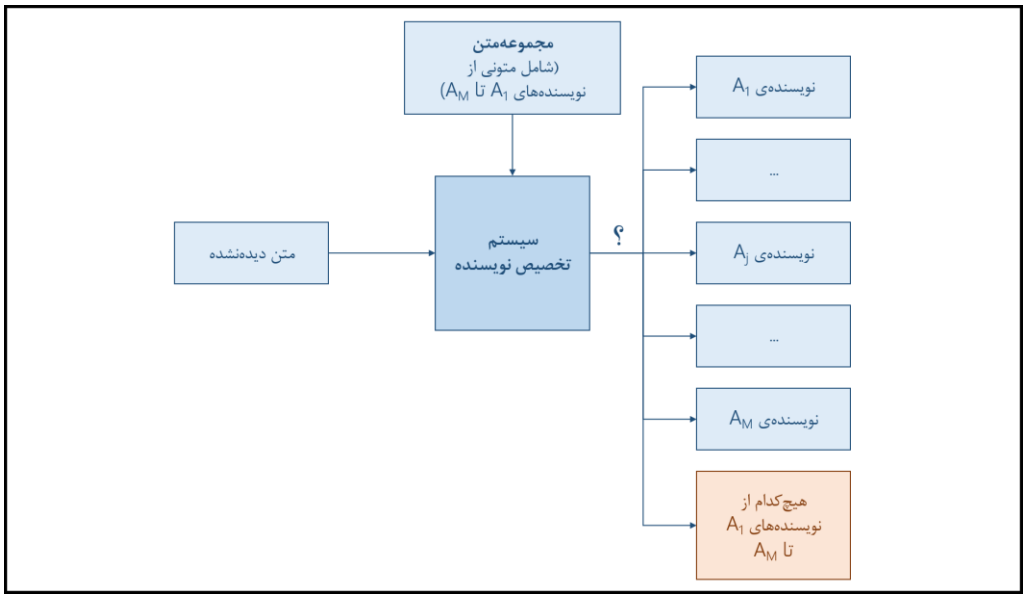
به طور کلی حل مسئله‌ی مجموعه-باز مشکل‌تر از مجموعه-بسته می‌باشد. هر چند اکثر پژوهش‌های صورت گرفته در حوزه‌ی تخصیص متن (از جمله این پایان‌نامه) بر روی قالب مجموعه-بسته متمرکز شده اند، اما اخیرا پژوهش‌های زیادی بر روی قالب مجموعه-بسته صورت گرفته و نتایج جالبی بدست آمده است. از جمله‌ی این پژوهش‌ها می‌توان به [۲۲ و ۳۰] اشاره نمود.

همانطور که گفته شد در این پایان‌نامه بر روی قالب مجموعه-بسته تمرکز شده است، اما با توجه به پیشرفت‌های ایجاد شده در قالب مجموعه-باز و توجه روزافزون پژوهشگران به آن، در بخش ۱-۳-۲ دو مسئله‌ی خاص از این قالب را مختصراً بررسی می‌نماییم.

[1] Closed-Set Attribution
[2] Open-Set Attribution



شکل ۱-۱: نمای کلی از مسئله‌ی تخصیص نویسنده، قالب مجموعه-بسته

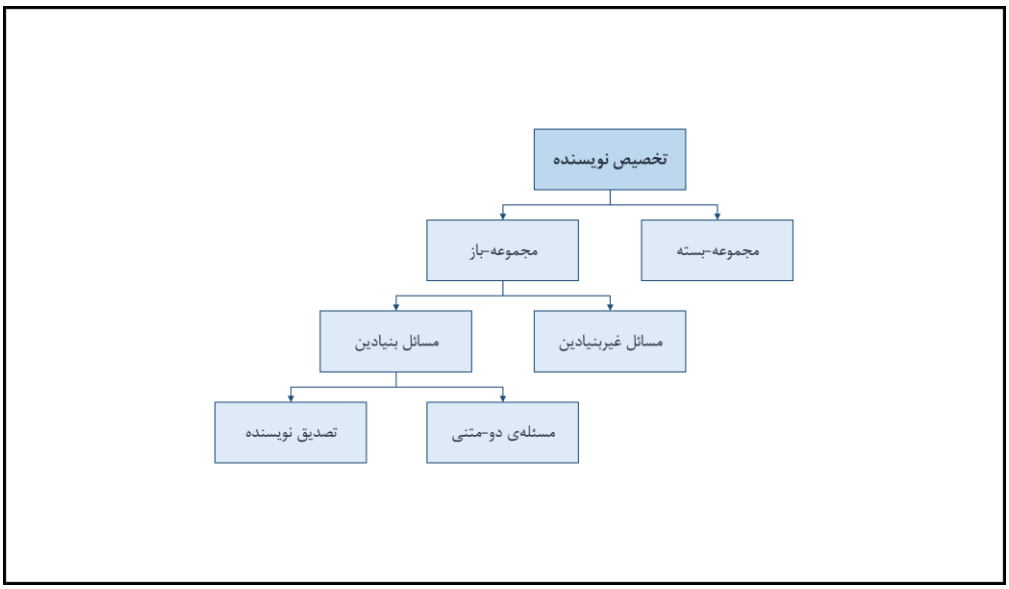


شکل ۱-۲: نمای کلی از مسئله‌ی تخصیص نویسنده، قالب مجموعه-باز

۱-۲ بررسی دو مسئله‌ی بنیادین

در حوزه‌ی تخصیص نویسنده، به مسئله‌ای که با دانستن راه‌حل آن بتوان سایر مسائل حوزه را حل نمود، «مسئله‌ی بنیادی» می‌گویند. از آنجا که تخصیص مجموعه-باز مسئله‌ی جامع‌تری از تخصیص مجموعه-

بسته است، مسائل بنیادین تخصیص نویسنده معمولاً در حالت تخصیص مجموعه-باز تعریف می‌شوند. در این بخش دو مورد از این مسائل را بررسی می‌نماییم. هم‌چنین تقسیم‌بندی قالب‌ها و مسائل بنیادین به صورت نموداری در شکل ۱-۸ آمده است.



شکل ۱-۳: تقسیم‌بندی قالب‌ها و مسائل بنیادین

الف) مسئله‌ی تصدیق نویسنده:

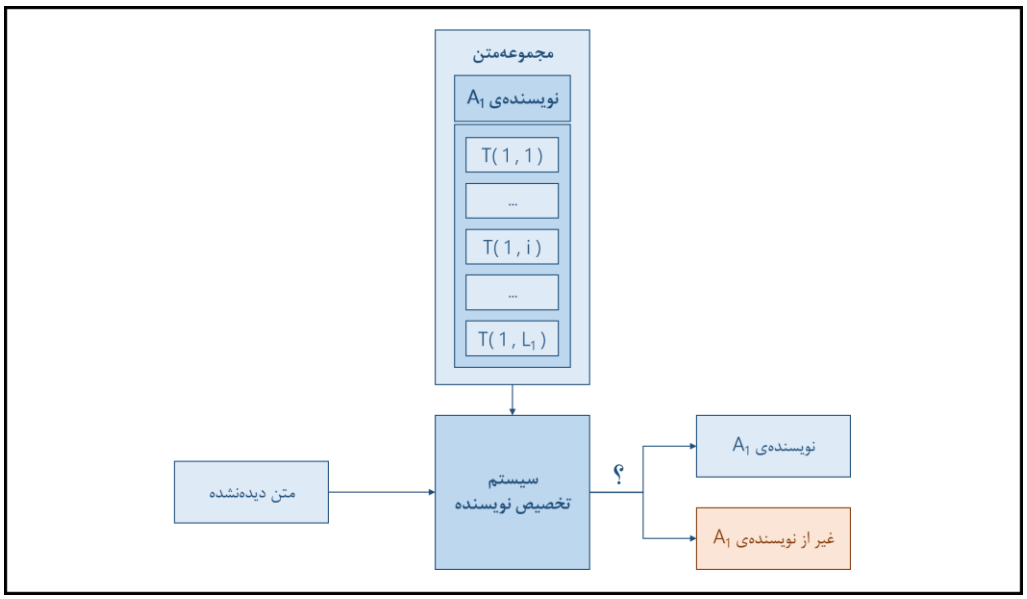
همانطور که گفته شد، در تخصیص مجموعه-بسته نویسنده‌ی واقعی متن دیده‌نشده حتماً یکی از نویسنده‌های موجود در مجموعه‌متن است، لذا در این حالت حداقل می‌بایست دو نویسنده در مجموعه‌متن حضور داشته باشند و مسئله‌ی تخصیص مجموعه-بسته‌ی تک-نویسنده بدون معناست. اما در تخصیص مجموعه-باز می‌توان حالتی را درنظر گرفت که مجموعه‌متن فقط شامل یک نویسنده باشد. در این حالت خاص که «تصدیق نویسنده[1]» نامیده می‌شود، صورت مسئله را می‌توان این‌گونه بیان کرد: تعدادی متن از یک نویسنده‌ی مشخص در اختیار داریم. حال متن جدیدی به ما داده می‌شود، آیا این متن توسط نویسنده‌ی موردنظر نوشته شده است یا خیر؟ مجموعه‌متن و سیستم AAS این مسئله در شکل ۱-۹ نمایش داده شده‌اند.

[1] Author Verification

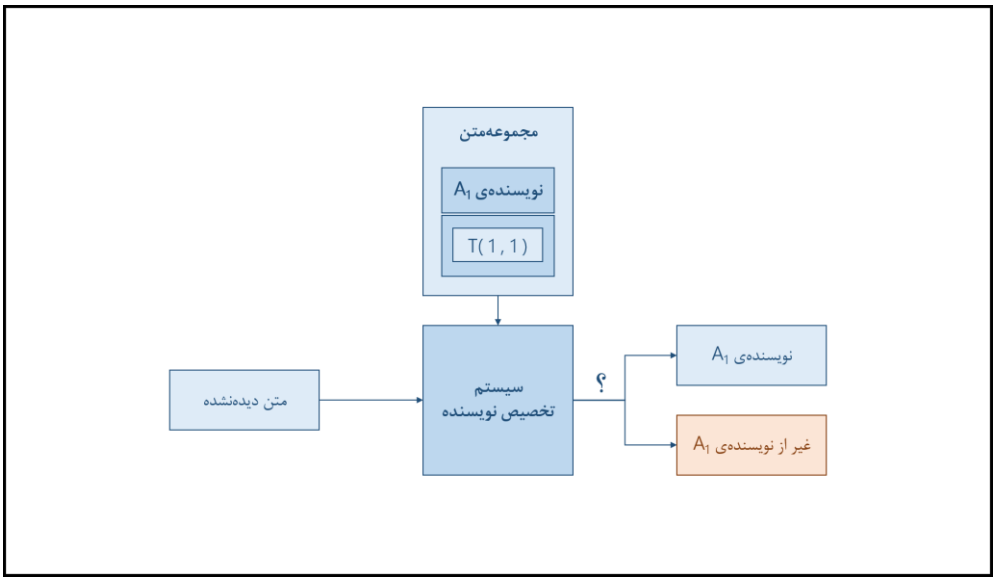
ب) مسئله‌ی دو-متنی

گفتیم که در مسئله‌ی تصدیق نویسنده فقط یک نویسنده در مجموعه‌متن وجود دارد. حال فرض کنیم که فقط یک متن از این نویسنده در مجموعه‌متن وجود داشته باشد. صورت مسئله در این حالت این‌گونه است: یک متن با نویسنده‌ی مشخص در اختیار داریم. متن جدید به ما داده می‌شود، آیا این متن توسط همان نویسنده نوشته شده است یا خیر؟ به عبارت دیگر: دو متن در اختیار داریم. می‌خواهیم بدانیم آیا این دو متن توسط یک نفر نوشته شده اند یا خیر؟ مجموعه‌متن و سیستم این حالت در شکل ۱-۱۰ نمایش داده شده اند.

دو مسئله‌ی بنیادین فوق دارای درجه‌ی سختی بالاتری نسبت به مسئله‌ی مجموعه-باز معمولی می‌باشند [۲۲]. اما با دانستن نحوه‌ی حل آن‌ها می‌توان هر مسئله‌ی مجموعه-باز دیگری را حل نمود. پوسا در پژوهش [۲۲] روشی برای حل مسئله‌ی بنیادین تصدیق نویسنده ارائه داده است. هم‌چنین تحقیقات [۲۳ و ۲۴] بر روی مسئله‌ی دو-متنی متمرکز شده‌اند.



شکل ۱-۴: نمای کلی مسئله‌ی تصدیق نویسنده



شکل ۱-۵: نمای کلی مسئله‌ی دو-متنی

۱-۳ بررسی دو نوع زبان متن

پس از بررسی مسائل مختلف حوزه‌ی پردازش متن، حال به بررسی انواع مختلف متن موجود در این حوزه از دیدگاه زبان می‌پردازیم. همانگونه که در شکل ۱-۱۱ نمایش داده شده است، یک متن بر اساس نوع زبان بر دو نوع تقسیم‌بندی می‌شود:

- متون زبان طبیعی[1]: این متون در واقع متن‌های نوشته شده به یکی از زبان‌های رایج بشری می‌باشد، به عنوان مثال: فارسی، عربی، انگلیسی، اسپانیایی، چینی و ...
- متون زبان برنامه‌نویسی[2]: این متون که در واقع کدهای نوشته‌شده در زبان‌های مختلف برنامه‌نویسی می‌باشند، به دو صورت تقسیم‌بندی می‌شوند:
  - بر مبنای زبان برنامه‌نویسی: متون به دسته‌هایی نظیر موارد زیر تقسیم‌بندی می‌شوند: .NET، Visual Basic، C++، C#، Java و ...
  - بر مبنای وجود توضیح[3]: از آنجا معمولاً متن نوشته شده در توضیح از نوع متون زبان طبیعی است، لذا وجود یا عدم وجود توضیح در متون برنامه‌نویسی بسیار حائز اهمیت است. لذا متون بر این مبنا به دو دسته تقسیم شده و برخی از روش‌های ارائه‌شده فقط بر

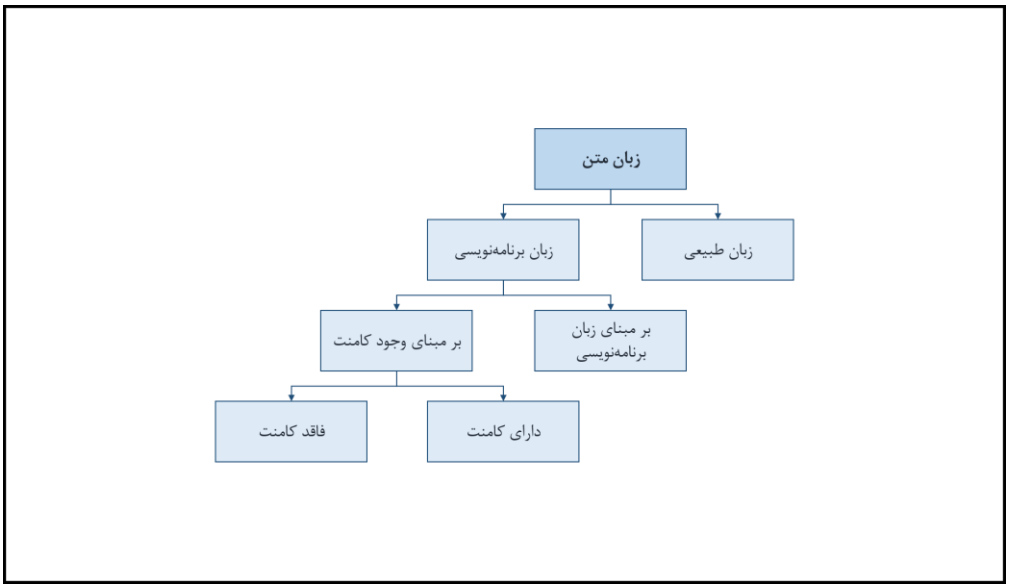
[1] Natural-Language Texts
[2] Programming-Language Texts
[3] Comment

روی یکی از این دسته‌ها موفقیت‌آمیز عمل می‌نمایند: برنامه‌های دارای توضیح و برنامه‌های فاقد توضیح.

هرچند دو حوزه‌ی فوق در ظاهر کاملاً متفاوت به نظر می‌رسند و معمولاً نیز روش‌های ارائه شده برای تخصیص نویسنده فقط بر روی یکی از این دو نوع زبان (طبیعی یا برنامه‌نویسی) پاسخ می‌دهد؛ اما در برخی از مواقع روش‌های ارائه شده برای یکی از این نوع زبان‌ها، برای نوع دیگر نیز پاسخ مطلوب می‌دهد. لذا برخی روش‌های ارائه شده در حوزه‌ی تخصیص نویسنده، بر روی هر دو نوع زبان قابل اجراست، اما برخی دیگر فقط مخصوص یکی از انواع زبان‌هاست.

به عبارت دیگر روش‌های ارائه‌شده در حوزه‌ی تخصیص نویسنده از نظر جامعیت زبانی انواع مختلفی دارند. به عنوان نمونه می‌توان موارد زیر را در نظر گرفت:

- برخی روش‌ها فقط بر روی زبان‌های طبیعی، آن هم فقط یکی از آن‌ها قابل اجرا هستند. به عنوان مثال: روشی مخصوص زبان انگلیسی
- برخی روش‌ها فقط بر زبان‌های طبیعی، اما تمامی آنها قابل اجرا هستند. به عنوان مثال: روشی که هم روی زبان انگلیسی، هم زبان فارسی و هم سایر زبان‌ها قابل اجراست.
- برخی روش‌ها علاوه بر آنکه بر روی تمامی زبان‌های طبیعی قابل اجرا هستند، بر روی تمامی زبان‌های برنامه‌نویسی نیز قابل اجرا می‌باشند.



شکل ۱-۶: تقسیم‌بندی زبان متن

پیوست ۲:

مرور پژوهش‌های صورت‌گرفته در زبان فارسی

طبق بررسی‌های صورت‌گرفته در این پایان‌نامه، از میان پژوهش‌هایی که تا کنون در حوزه‌ی تخصیص نویسنده بر روی متون فارسی صورت پذیرفته است، شش پژوهشی که در ادامه بیان شده، در دسترس می‌باشند. لازم بذکرست که تمامی این پژوهش‌ها با استفاده از الگووارهٔ مبتنی بر نمونه (و به طور خاص‌تر با استفاده از مدل فضای برداری) صورت پذیرفته و بیش‌تر بر ویژگی‌های حاصل از ابزار NLP متکی هستند.

۲-۱ پژوهش اول: شاهمیری و همکاران، ۱۳۸۴

اولین پژوهش تخصیص نویسنده، توسط شاهمیری و همکاران بر روی شاهنامه‌ی فردوسی صورت گرفت. هدف این پژوهش [۴۶] طراحی سیستمی جهت تشخیص اشعار شاهنامه از سایر اشعار ادبیات فارسی بود. لازم بذکرست که در این تحقیق هر بیت شعر، به عنوان یک نمونه متن در نظر گرفته شده است.

- پایگاه داده: شامل ۱۰۰ بیت از اشعار شاهنامه و ۱۰۰ بیت از اشعاری خارج از شاهنامه
- ویژگی‌های مورد استفاده: سه دسته ویژگی مورداستفاده قرار گرفته است:
  - ویژگی‌های فیزیکی: ترتیب حروف، هم‌قافیه بودن مصراع‌ها، تعداد حروف قافیه‌ی مصراع‌ها و ...
  - ویژگی‌های مفهومی: تعداد واژگان ناایرانی، تعداد واژگان ناایرانی و ...
  - ویژگی‌های آوایی: وزن شعر، تعداد هجا، تفاضل هجای پو مصراع و ...

در مجموع، ۴۷ ویژگی برای تبدیل متن به بردار انتخاب شده است.

- روش‌های طبقه‌بندی مورد استفاده: روش آنالیز آماری و روش شبکه‌های عصبی مصنوعی
- نتایج: بهترین نتیجه دقتی معادل ۱۰۰ درصد بوده و با استفاده از شبکه‌ی عصبی حاصل شده است.

۲-۲ پژوهش دوم: شاهمیری و همکاران، ۱۳۸۵

این پژوهش [۴۷]، ادامه‌ی تحقیق قبلی است و در آن پایگاه‌داده گسترش و تعداد ویژگی‌ها افزایش یافته اند. در ضمن یک روش جدید (درخت تصمیم) نیز آزمایش شده است.

- پایگاه داده: شامل ۴۰۰ بیت از اشعار شاهنامه و ۴۰۰ بیت از اشعاری خارج از شاهنامه
- ویژگی‌های مورد استفاده: همان دسته‌بندی بیان‌شده در پژوهش قبلی را داراست، اما در این پژوهش در مجموع، ۶۴ ویژگی برای تبدیل متن به بردار انتخاب شده است.
- روش‌های طبقه‌بندی مورد استفاده: روش درخت تصمیم و روش شبکه‌های عصبی مصنوعی
- نتایج: در این پژوهش نیز، بهترین نتیجه دقتی معادل ۱۰۰ درصد بوده و با استفاده از شبکه‌ی عصبی حاصل شده است.

۲-۳ پژوهش سوم: شاهمیری و همکاران، ۱۳۸۶

در این پژوهش [۴۸] صورت مسئله گسترش و به‌جای تشخیص اشعار شاهنامه از غیر شاهنامه، به تشخیص شاعر یک بیت از میان چهار گزینه‌ی موجود تبدیل شده است.

- پایگاه داده: شامل ۸۰۰ بیت از اشعار فردوسی، خیام، حافظ و مولوی
- ویژگی‌های مورد استفاده: همان دسته‌بندی بیان‌شده در پژوهش قبلی را داراست، اما در این پژوهش در مجموع، ۶۸ ویژگی برای تبدیل متن به بردار انتخاب شده است.
- روش‌های طبقه‌بندی مورد استفاده: روش درخت تصمیم و روش شبکه‌های عصبی مصنوعی
- نتایج: بهترین نتیجه دقتی معادل ۹۵/۹ درصد بوده و با استفاده از شبکه‌ی عصبی حاصل شده است.

۲-۴ پژوهش چهارم: فرهمندپور و همکاران، تابستان ۱۳۹۱

در این پژوهش [۴۹] که در دانشگاه بوعلی سینای همدان صورت پذیرفته است، برخلاف سه پژوهش قبل، هدف تشخیص نویسنده‌ی متون نثر فارسی است نه متون نظم.

- پایگاه داده: دو پایگاه‌داده تهیه شده است:
  - دانشگاه بوعلی سینا: شامل متونی از ۲۰ دانشجوی دانشگاه بوعلی سینا. به ازای هر نویسنده ۲۰۰۹ کلمه، ۱۵۰۰ کلمه برای آموزش، ۵۰۹ کلمه برای آزمایش
  - نویسندگان هم‌عصر: شامل متونی از هشت نویسنده‌ی فارسی زبان هم عصر. به ازای هر نویسنده ۷۷۵۰ کلمه، ۵۰۰۰ کلمه برای آموزش، ۲۵۰۰ کلمه برای آزمایش
- ویژگی‌های مورد استفاده: چهار دسته ویژگی مورداستفاده قرار گرفته است:

- ویژگی‌های واژگانی: پرمایگی واژگانی، میانگین طول کلمات، میانگین طول جملات (بر حسب نویسه)
- ویژگی‌های نحوی: توزیع تکرار عبارات اسمی، قیدی و صفت و ...
- ویژگی‌های معنایی: تعداد تکرار انواع افزوده‌های ربطی و ...
- ویژگی‌های وابسته به کاربرد: نحوه‌ی توزیع نویسه‌هایی خاص، همچون: فاصله، Enter، Tab و ...

در مجموع، بیش از ۱۵۰۰ ویژگی برای تبدیل متن به بردار انتخاب شده است.

- روش‌های طبقه‌بندی مورد استفاده: پنج روش استفاده شده است: استفاده از روش LDA، استفاده از شبکه‌ی عصبی، استفاده از درخت تصمیم، استفاده از روش دلتا[1] و استفاده از روش kNN
- نتایج: بهترین نتایج بدست آمده در این پژوهش عبارتند از:
  - پایگاه‌داده دانشگاه بوعلی: دقت ۸۱/۶ درصد، با استفاده از روش LDA
  - پایگاه‌داده نویسندگان هم‌عصر: دقت ۱۰۰ درصد، با استفاده از روش LDA

۵-۲ پژوهش پنجم: فرهمندپور و همکاران، زمستان ۱۳۹۱

این پژوهش [۵۰] در واقع گسترش تحقیق قبلی است به ازای روش‌های جدید روی همان پایگاه‌داده‌ها.

- پایگاه داده: همان دو پایگاه‌داده مورد استفاده قرار گرفته اند.
- ویژگی‌های مورد استفاده: از همان ویژگی‌ها استفاده شده است.
- روش‌های طبقه‌بندی مورد استفاده: چهار روش استفاده شده است: استفاده از روش kNN، استفاده از روش دلتا، ترکیب الگوریتم ژنتیک و روش kNN و ترکیب الگوریتم ژنتیک و روش دلتا
- نتایج: بهترین نتایج بدست آمده در این پژوهش عبارتند از:
  - پایگاه‌داده دانشگاه بوعلی: دقت ۸۰ درصد، با استفاده از ترکیب الگوریتم ژنتیک و روش kNN
  - پایگاه‌داده نویسندگان عم‌عصر: دقت ۱۰۰ درصد، با استفاده از ترکیب الگوریتم ژنتیک و روش kNN

[1] Delta Method

۲-۶ پژوهش ششم: رمضانی و همکاران، ۱۳۹۲

در این پژوهش [۵۱] که در دانشگاه فردوسی مشهد پذیرفته است، همانند دو پژوهش قبل، هدف تشخیص نویسنده‌ی متون نثر فارسی است. اما از پایگاه‌داده‌هایی با تعداد نویسنده‌ی مختلف استفاده شده است.

- پایگاه داده: پنج پایگاه‌داده تهیه شده است: پایگاهی شامل ۲، ۵، ۱۰، ۲۰ و ۴۰ نویسنده‌ی معاصر فارسی
- ویژگی‌های مورد استفاده: ویژگی‌های مختلفی استفاده شده است، از جمله:
  - ویژگی‌های نویسه‌ای: تعداد نویسه‌ها، انگرام‌های در سطح نویسه و ...
  - ویژگی‌های لغوی: تعداد لغات، طول جملات و ...
  - ویژگی‌های نحوی: اطلاعات افعال، اطلاعات قیدها و ...

در این پژوهش، چندین مجموعه از ویژگی‌های فوق به صورت مجزا مورد آزمایش قرار گرفته اند، تا تأثیر و میزان موفقیت هریک در تخصیص نویسنده مشخص گردد.

- روش‌های طبقه‌بندی مورد استفاده: سه روش استفاده شده است: استفاده از روش SVM، استفاده از روش kNN و استفاده از روش C5
- نتایج: اطلاعات مربوط به بهترین نتایج بدست‌آمده روی هر پایگاه‌داده در جدول ۲-۱ نمایش داده شده است. این اطلاعات شامل موارد مقابل است: بالاترین دقت، ویژگی‌ای که بالاترین دقت را نتیجه داده و طبقه‌بندی که منجر به تولید بهترین نتیجه شده است.

جدول ۲-۱: بهترین نتایج بدست‌امده در پژوهش ششم به تفکیک پایگاه‌داده

	۲ نویسنده	**۵ نویسنده**	**۱۰ نویسنده**	**۲۰ نویسنده**	**۴۰ نویسنده**
بالاترین دقت	۱۰۰ درصد	۹۶ درصد	۸۰ درصد	۷۴ درصد	۶۴ درصد
ویژگی	طول کلمات	فرکانس تکرار نویسه‌ها	فرکانس تکرار کلمات	فرکانس تکرار نویسه‌ها	فرکانس تکرار نویسه‌ها
طبقه‌بند	kNN	C5	C5	C5	C5

پیوست ۳

مقایسه‌ی ویژگی‌ها و روش‌های تخصیص نویسنده

در بخش‌های قبلی، ویژگی‌ها و روش‌های ارائه‌شده در حوزه‌ی تخصیص نویسنده را از سه دیدگاه تقسیم‌بندی نمودیم (این تقسیم‌بندی به صورت خلاصه در شکل ۲-۱ آمده است):

۱) از دیدگاه ویژگی‌های مورد استفاده
۲) از دیدگاه نحوه‌ی استخراج ویژگی
۳) از دیدگاه نحوه‌ی رفتار با متون مجموعه‌متن (الگوواره‌ی تخصیص)

حال می‌خواهیم حوزه‌ی مطالعاتی خود را از نظر این سه دیدگاه تعیین نماییم، یعنی به این سؤالات پاسخ دهیم که می‌خواهیم در این پایان‌نامه:

۱) از کدام دسته از ویژگی‌ها استفاده نماییم؟
۲) ویژگی‌ها را به صورت ایستا استخراج کنیم یا به صورت پویا؟
۳) کدام الگوواره را در دستور کار خود قرار دهیم؟ مبتنی بر نمونه یا مبتنی بر پروفایل؟

به منظور پاسخ به این سؤالات، شش بررسی متفاوت انجام داده و از جمع‌بندی نتایج این شش بررسی به پاسخ سؤالات فوق می‌رسیم.

۳-۱ بررسی اول: نیازهای پایان‌نامه و پژوهش‌های گذشته

در تعیین روند کلی روش مورد استفاده در این پایان‌نامه، می‌بایست به دو نکته توجه کرد:

۱) تحقیقات این پایان‌نامه بر روی متون فارسی و عربی صورت می‌پذیرد؛ لذا می‌بایست روش استفاده‌شده در این پایان‌نامه تا جای ممکن «مستقل از زبان[1]» باشد.
۲) همانطور که در بخش ۲-۵ دیدیم، پژوهش‌های گذشته در حوزه‌ی تخصیص نویسنده‌ی متون زبان فارسی ، عموماً بر روی ویژگی‌های حاصل از ابزارهای NLP متمرکز شده اند. از طرفی ابزارهای NLP وابسته به زبان بوده و دشواری بالایی دارند. لذا اگر در این پایان‌نامه روی ویژگی‌های دیگری

[1] Language-Independent

غیر ویژگی‌های حاصل از ابزارهای NLP متمرکز شویم، علاوه بر اجتناب از دشواری موجود در استفاده از این ابزار، حوزه‌ی نوینی در پژوهش‌های تخصیص متن فارسی را آغاز کرده ایم.

پس به طور خلاصه می‌توان نتیجه‌ی زیر را از بررسی اول استخراج نمود:

نتیجه‌ی ۱: هنگام تعیین روش مورد استفاده در این پایان‌نامه، بهترست روش‌هایی را انتخاب نماییم که:

۱) مستقل از زبان باشند.
۲) بر مبنای ویژگی‌هایی غیر از ویژگی‌های حاصل از ابزار NLP بنا شده باشند.

۲-۳ بررسی دوم: ویژگی‌های لغوی و نویسه‌ای در برابر ویژگی‌های نحوی و معنایی

در بخش ۲-۲ دیدیم که استفاده از ویژگی‌های نحوی و معنایی، عموماً مستلزم در اختیار داشتن ابزارهای خاص NLP بوده و لذا وابستگی زیادی به زبان متن دارند. از طرفی با بررسی مثال‌های ارائه‌شده در بخش‌های ۲-۲-۱ و ۲-۲-۲ به راحتی می‌توان دریافت که روش‌های مبتنی بر استفاده از ویژگی‌های لغوی و نویسه‌ای:

۱) عموماً مستقل از زبان می‌باشند. به عنوان مثال شمارش تعداد تکرار یک کلمه‌ی خاص در یک متن، به هیچ عنوان وابسته به زبان متن نمی‌باشد.
۲) معمولاً نیاز به ابزار NLP خاصی ندارند.
۳) دارای پیچیدگی و دشواری به مراتب کمتری نسبت به ویژگی‌های لغوی و معنایی می‌باشند.

در بررسی اول بیان شد که با توجه به هدف این پایان‌نامه می‌بایست به دنبال روشی باشیم که هم مستقل از زبان بوده و هم نیاز به ابزار NLP خاصی نداشته باشد، اما ویژگی‌های نحوی و معنایی هیچ‌کدام از این دو ویژگی را ندارند. از طرفی با توجه به موارد فوق ویژگی‌های لغوی و نویسه‌ای عموماً دارای این دو خصیصه هستند. لذا در مجموع، می‌توان نتیجه‌ی بررسی دوم را این‌گونه بیان نمود:

نتیجه‌ی ۲: از میان چهار دسته‌ی اصلی ویژگی‌ها (یعنی لغوی، نویسه‌ای، نحوی و معنایی)، ویژگی‌های لغوی و نویسه‌ای برای تحقیقات این پایان‌نامه مناسب‌تر می‌باشند.

۳-۳ بررسی سوم: ویژگی‌های نویسه‌ای در برابر ویژگی‌های لغوی

در بررسی اول به این نتیجه رسیدیم که از میان چهار نوع ویژگی اصلی، ویژگی‌های لغوی و نویسه‌ای را برگزینیم. حال می‌خواهیم بررسی کنیم که کدام‌یک از این دو نوع ویژگی برای اهداف این پایان‌نامه مناسب‌تر می‌باشد؟! طبق بررسی‌های صورت‌گرفته در پژوهش [۳۱]:

- از میان ویژگی‌های لغوی، ویژگی‌های کلمات دستوری و کلمات پرتکرار در مجموع (از نظر دقت و سادگی) ویژگی‌های مناسب‌تری می‌باشند.
- از میان ویژگی‌های نویسه‌ای، ویژگی‌های مبتنی بر انگرام در مجموع ویژگی‌های مناسب‌تری هستند.

پس بررسی خود را بر مقایسه‌ی ویژگی‌های کلمات دستوری، کلمات پرتکرار و ویژگی‌های مبتنی بر انگرام متمرکز می‌کنیم. به منظور روان‌تر شدن متن در ادامه از ویژگی‌های کلمات دستوری و کلمات پرتکرار با عنوان «ویژگی‌های اول» و از ویژگی‌های مبتنی بر انگرام به عنوان «ویژگی‌های دوم» یاد می‌کنیم. پژوهش‌های صورت گرفته در حوزه‌ی تخصیص متن نشان داده است که:

- پیاده‌سازی روش‌های اول بر روی زبان‌هایی که تعریف دقیقی از کلمه ندارند (همچون زبان چینی) با مشکل مواجه است. اما پیاده‌سازی روش‌های دوم چنین محدودیتی ندارند، زیرا نیازی به تعریف کلمه نداشته و مستقیماً از روی نویسه‌ها استخراج می‌گردند.
- برای پیاده‌سازی روش‌های اول نیاز به سیستمی داریم که بتواند کلمات یک متن را از هم جدا نماید. زمانی که ادبیات متن رسمی باشد، تولید چنین سیستمی بسیار آسان است و لذا به عنوان نقطه‌ضعف روش‌های اول به حساب نمی‌آید. اما زمانی که متن ادبیات محاوره‌ای داشته باشد، تولید چنین سیستمی پیچیده خواهد شد. اما روش‌های دسته‌ی دوم چنین محدودیتی ندارند، زیرا هیچ سیستم خاصی برای جداسازی نویسه‌های متن نیاز نیست.
- روش‌های اول ترتیب قرار گرفتن کلمات در متن را نادیده می‌گیرند. این ترتیب عموماً اطلاعات مهمی را در خود دارد که اصطلاحاً به آن‌ها «اطلاعات متنی[1]» گفته می‌شود. اما ویژگی‌های روش‌های دوم این اطلاعات را نیز با خود بهمراه دارند.
- روش‌های دوم در برابر نویز بسیار مقاوم‌تر از روش‌های اول می‌باشند. به عنوان مثال فرض کنیم یک ایراد املایی در متن وجود داشته باشد: به جای کلمه‌ی ‘simplistic’ اشتباهاً کلمه‌ی ‘simpilstc’ نوشته شده باشد. در روش‌های اول این دو کلمه دو ویژگی کاملاً متفاوت به حساب

[1] Contextual Information

می‌آیند، اما در روش‌های دوم این دو کلمه با حفظ تفاوت، تعدادی ویژگی (انگرام) مشترک نیز تولید می‌کنند.

- نقطه‌ضعف اصلی روش‌های دوم تعداد بالای ویژگی‌های تولید شده توسط آن‌ها نسبت به روش‌های اول می‌باشد. به عنوان مثال عبارت 'The arrival of the Savior' فقط دارای دو ۵ کلمه است، اما ۲۵ انگرام تولید می‌نماید (به ازای هر نویسه، یک انگرام).

و در نهایت:

- گریو در سال ۲۰۰۷ میلادی، با آزمایش ویژگی‌های نویسه‌ای و لغوی مختلف بر روی یک مجموعه‌متن یکسان، به این نتیجه رسید که ویژگی‌های مبتنی بر انگرام در سطح نویسه بهترین عملکرد را از خود نشان می‌دهند. [۷۱]

در نهایت با جمع‌بندی موارد فوق می‌توان نتیجه‌ی بررسی سوم را بدین‌صورت خلاصه نمود:

نتیجه‌ی ۳: از میان ویژگی‌های لغوی و نویسه‌ای، ویژگی‌های نویسه‌ای و از میان انواع مختلف ویژگی‌های نویسه‌ای، ویژگی‌های مبتنی بر انگرام برای تحقیقات این پایان‌نامه مناسب‌تر می‌باشند.

۳-۴ بررسی چهارم: استخراج ایستا در برابر استخراج پویا

پس از آنکه مشخص شد ویژگی‌های مبتنی بر انگرام برای تحقق اهداف این پایان‌نامه مناسب‌ترند، حال می‌بایست به این سؤال پاسخ داد که آیا این ویژگی‌ها را به صورت ایستا استخراج نماییم، یا به صورت پویا؟!

بدین منظور می‌بایست دو نکته را مورد توجه قرار دهیم:

- طبق بررسی انجام‌شده در [۲۸]، هرچند در گذشته استخراج ایستا رواج بسیار زیادی در بین پژوهش‌گران داشته و اکثر تحقیقات حوزه‌ی تخصیص متن از آن بهره می‌برده است، اما جدیداً استخراج پویا جایگزین آن‌ها شده اند، زیرا بررسی‌ها نشان داده است که عموماً استخراج پویا جواب بهتری در مسئله‌ی تخصیص نویسنده ارائه می‌دهد.
- از طرفی همانطور که در بخش ۲-۳-۲ بیان شد، زمانی که از ویژگی‌های مبتنی بر انگرام استفاده می‌شود، عموماً روش استخراج پویا برای حل مسئله انتخاب می‌گردد.

پس به راحتی می‌توان به نتیجه‌ی بررسی چهارم پی برد:

نتیجه‌ی ۴: از میان دو نحوه‌ی استخراج ویژگی‌ها (ایستا و پویا)، استخراج پویا برای تحقیقات این پایان‌نامه مناسب‌تر می‌باشند.

۳-۵ بررسی پنجم: الگوواره‌ی مبتنی بر نمونه در برابر الگوواره‌ی مبتنی بر پروفایل

در بررسی‌های قبلی به این نتیجه رسیدیم که می‌بایست از ویژگی‌های مبتنی بر انگرام استفاده نموده و آن‌ها را به صورت پویا استخراج نماییم. حال می‌بایست از میان دو الگوواره‌ی تخصیص موجود، یکی را برگزینیم. به منظور روان‌تر شدن متن، در ادامه‌ی این بخش از الگوواره‌ی مبتنی بر نمونه و الگوواره‌ی مبتنی بر پروفایل به ترتیب با عناوین «الگوواره‌ی اول» و «الگوواره‌ی دوم» یاد می‌کنیم.

بررسی‌های [۳۱] نشان می‌دهد که اکثر پژوهش‌گران حوزه‌ی تخصیص نویسنده، تحقیقات خود را بر روی الگوواره‌ی اول متمرکز کرده‌اند. اما توجه به سه نکته در مورد الگوواره‌ی دوم ضروری است:

۱) روش‌هایی که از ویژگی‌های مبتنی بر انگرام استفاده می‌کنند، اکثراً الگوواره‌ی دوم را در دستور کار خود قرار می‌دهند. غالب بودن الگوواره‌ی دوم تا جایی‌ست که لایتون در سال ۲۰۱۲ در تقسیم‌بندی کلی روش‌های تخصیص نویسنده، آن‌ها را به دو دسته‌ی «مدل فضای برداری» و «پروفایل‌های مبتنی بر انگرام محلی[1]» (در فصل سوم شرح داده خواهد شد) گروه‌بندی نموده و می‌نویسد [۴۵]:
«هر چند روش‌های تولید پروفایل، الزاماً به یک روش خاص محدود نمی‌شوند، اما پژوهش‌های گذشته بر روی استفاده از انگرام جهت تولید پروفایل متمرکز شده اند.»
هر چند در الگوواره‌ی اول نیز از ویژگی‌های مبتنی بر انگرام استفاده می‌شود، اما عمده‌ی استفاده‌ی این نوع ویژگی‌ها در الگوواره‌ی دوم است.

۲) پوسا و استاماتاتوس در مقاله‌ی مروری خود سال ۲۰۱۴ میلادی [۲۲]، با وجود اذعان به غالب بودن پژوهش‌ها در مورد الگوواره‌ی اول، بهتر بودن همیشگی این الگوواره را رد کرده و روشی با استفاده الگوواره‌ی دوم ارائه می‌دهند که پاسخ بهتری از سایر روش‌هایی که از الگوواره‌ی اول استفاده می‌کنند، می‌دهد.

[1] Local N-gram Profiles (LNG)

۳) همان‌گونه که در بخش ۲-۵ دیدیم، تمامی پژوهش‌های صورت‌گرفته در حوزه‌ی تخصیص نویسنده در زبان فارسی بر روی الگوواره‌اول متمرکز شده و الگواره‌ی دوم با آنکه در متون لاتین (به خصوص انگلیسی) عملکرد بسیار موفقیت‌آمیزی داشته، مورد بی‌مهری قرار گرفته است. لذا لذا جای خالی تحقیقات الگوواره‌ی دوم در زبان فارسی به شدت حس می‌شود و با تمرکز بر روی آن، می‌توان فصلی جدید را در حوزه‌ی تخصیص نویسنده‌ی متون فارسی آغاز کرد.

با جمع‌بندی سه مورد فوق، نتیجه‌ی بررسی پنجم را بدین صورت بیان می‌کنیم:

نتیجه‌ی ۵: از میان دو الگوواره‌ی تخصیص موجود (مبتنی بر نمونه و مبتنی بر پروفایل)، الگوواره‌ی مبتنی بر پروفایل برای تحقیقات این پایان‌نامه مناسب‌تر می‌باشند.

۳-۵-۱ بررسی ششم: جمع‌بندی نتایج و نتیجه‌گیری نهایی

نتایج حاصل از بررسی‌های اول تا پنجم را بدین صورت جمع‌بندی می‌نماییم:

نتیجه‌ی نهایی: از میان تمامی ویژگی‌ها و روش‌های موجود، روش‌های زیر برای تحقیقات این پایان‌نامه مناسب می‌باشند: روش‌هایی که:

۱) از ویژگی‌های مبتنی بر انگرام استفاده نمایند.
۲) ویژگی‌ها را به صورت پویا استخراج کنند.
۳) الگوواره‌ی مبتنی بر پروفایل را در دستور کار خود قرار دهند.

این روش‌ها را به صورت قراردادی «روش‌های NDP»[1] می‌نامیم.

تعداد زیادی از روش‌های ارائه‌شده تا کنون را می‌توان در دسته‌ی روش‌های NDP تقسیم‌بندی نمود که مهم‌ترین آن‌ها را در فصل سوم بررسی می‌نماییم.

[1] N-gram-based, Dynamic and Profile-based (NDP)

پیوست ۴:

نسخه‌های توسعه‌یافته از روش انگرام‌های متداول

در این بخش به بررسی هفت مورد از روش‌های NDP که بر مبنای روش CNG ارائه شده‌اند، می‌پردازیم. پیش از آغاز این بررسی‌ها می‌بایست چند نکته را بیان نماییم:

۱) در این بخش نیز به منظور معرفی هر روش، سه مشخصه از آن تشریح می‌شود: فلسفه‌ی ارائه‌ی روش، سیستم استخراج پروفایل و سیستم محاسبه‌ی فاصله

۲) در تمامی مطالب این بخش، نمادگذاری زیر را رعایت می‌نماییم:

- یک متن دلخواه: T
- یک نویسنده‌ی موجود در مجموعه‌متن: A
- متن دیده‌نشده: U
- تعداد نویسندگان موجود در مجموعه‌متن: M

پس می‌توان هدف دو سیستم اساسی این روش‌ها را این‌گونه بیان نمود:

- سیستم استخراج پروفایل: استخراج پروفایل از T
- سیستم محاسبه‌ی فاصله: محاسبه‌ی یک مقدار عددی به عنوان فاصله‌ی میان پروفایل استخراج‌شده از U و پروفایل استخراج‌شده از A (یا به طور خلاصه: محاسبه‌ی فاصله‌ی میان U و A)

۳) هنگام تشریح هر روش فرض می‌کنیم که پارامترهای N (طول انگرام) و L (طول پروفایل) از قبل تعیین شده‌اند و لذا از بیان گام صفرم صرف نظر می‌نماییم.

۴) تمامی روش‌ها در این بخش به صورت مختصر شرح داده شده‌اند و خواننده به منظور آشنایی کامل‌تر با این روش‌ها، می‌بایست به مقاله‌ی اصلی هر یک مراجعه نماید (جدول ۳-۲).

۵) از آن‌جا که مبنای تمامی این روش‌ها، روش CNG می‌باشد، گاهی عبارت CNG به ابتدای نام روش‌ها افزوده می‌شود. به عنوان مثال از روش «اشتراک وزن‌دار پروفایل‌ها» در برخی مقالات با عنوان مستقل WPI و در برخی دیگر با عنوان CNG-WPI یاد می‌شود. در این بخش هنگام نوشتن متن از عناوین مستقل روش‌ها استفاده شده، اما در نوشتن روابط ریاضی عنوان کامل هر روش به کار رفته است.

۶) در این بخش روش‌ها به همان ترتیب ارائه‌شده در جدول ۳-۲ معرفی می‌شوند (یعنی بر مبنای سال ارائه). از میان هشت روش موجود در این جدول، روش اول (روش CNG) در بخش ۳-۴ به طور کامل تشریح شد. حال هفت روش باقی‌مانده (از روش دوم تا روش هشتم) در این بخش معرفی می‌شوند.

۴-۱ روش دوم: روش SPI

روش SPI برای اولین‌بار در سال ۲۰۰۷ میلادی و در مقاله‌ی [۲۷] ارائه گردید. هرچند این روش در ابتدا برای مسائل تخصیص نویسنده‌ی کدهای برنامه‌نویسی معرفی شد، اما بعدها در متون زبان طبیعی نیز مورداستفاده قرار گرفته و عملکرد رضایت‌بخش آن موجب شد تا در زمره‌ی مهم‌ترین روش‌های NDP قرار گیرد. با وجود اینکه گاهی از این روش با عنوان «پروفایل نویسنده‌ی کد مبدأ» یا به طور خلاصه SCAP یاد می‌شود، اما در این پایان‌نامه همواره نام SPI مورداستفاده قرار می‌گیرد.

۴-۱-۱ فلسفه‌ی ارائه‌ی روش

روش SPI در واقع حاصل ایجاد برخی ساده‌سازی‌ها بر روی روش CNG می‌باشد. این روش نوع جدیدی از پروفایل را (که نسخه‌ی ساده‌شده‌ی پروفایل کلاسیک است) معرفی نموده و هم‌چنین معیار جدیدی برای محاسبه‌ی فاصله‌ی میان U و A پیشنهاد می‌کند.

۴-۱-۲ سیستم استخراج پروفایل

همانطور که در بخش ۳-۴ بیان شد، روش CNG، نوعی خاص از پروفایل موسوم به پروفایل کلاسیک (CP) را مورد استفاده قرار می‌دهد. این نوع پروفایل در واقع شامل دو مجموعه است: مجموعه‌ای شامل L-انگرام پرتکرار و مجموعه‌ای شامل فرکانس نرمال‌شده‌ی هر یک از L-انگرام پرتکرار.

روش SPI تغییر کوچکی در این نوع پروفایل ایجاد نموده و با حذف مجموعه‌ی فرکانس‌ها، تعریف جدیدی از پروفایل یک متن ارائه می‌دهد: «پروفایل یک متن عبارتست از L-انگرام پرتکرار آن». به عبارت دیگر این

نوع پروفایل، از حذف مجموعه‌ی فرکانس‌ها از پروفایل CP ایجاد می‌گردد، لذا پروفایلی که روش SPI معرفی می‌نماید، فقط شامل یک مجموعه است که همان مجموعه‌ی G ارائه‌شده در روش CNG می‌باشد. یعنی:

$$PR_T^{CNG-SPI} = \left\{ G_T^{CNG} \right\} \qquad (۱-۴)$$

این نوع پروفایل را که در واقع نسخه‌ی ساده‌شده‌ی پروفایل کلاسیک است، «پروفایل ساده‌شده[1]» نامیده و با نماد SP نمایش می‌دهیم.

شاید در ذهن خواننده چنین فرضیه‌ای ایجاد شده باشد که هنگام استخراج پروفایل SP نیازی به شمارش انگرام‌ها وجود ندارد. این فرضیه صحیح نمی‌باشد، زیرا در این روش نیز (همانند روش CNG) می‌بایست فرایند شمارش انگرام‌ها صورت پذیرد تا در اثر آن مجموعه‌ی L-انگرام پرتکرار متن تشکیل شود. البته پس از تشکیل این مجموعه دیگر نیازی به محاسبه‌ی فرکانس نرمال‌شده‌ی اعضای آن نمی‌باشد، زیرا مجموعه‌ی F_T^{CNG} دیگر در پروفایل قرار ندارد.

۴-۱-۳ سیستم محاسبه‌ی فاصله

به یاد داریم که در روش CNG برای محاسبه‌ی فاصله‌ی میان U و A از معیار فاصله‌ی نسبی (RD) استفاده می‌شد. همان‌طور که در رابطه‌ی ۳-۷ آمده است، این معیار مبتنی بر فرکانس انگرام‌ها می‌باشد. پس با توجه به حذف شدن مجموعه‌ی فرکانس‌ها در پروفایل SP، دیگر نمی‌توان از معیار فاصله‌ی نسبی (RD) جهت محاسبه‌ی فاصله‌ی میان دو پروفایل استفاده نمود. لذا می‌بایست روش جدید ارائه گردد. در روش SPI به منظور محاسبه‌ی فاصله‌ی میان U و A باید مراحل زیر را طی نمود:

گام اول: ابتدا می‌بایست با استفاده از رابطه‌ی زیر، مجموعه‌ی $SPI(U,A)$ را تشکیل داد:

$$SPI(U,A) = G_U^{CNG} \cap G_A^{CNG} \qquad (۲-۴)$$

مجموعه‌ی $SPI(U,A)$ متشکل از انگرام‌های مشترک میان پروفایل‌های $PR_U^{CNG-SPI}$ و $PR_A^{CNG-SPI}$ می‌باشد.

گام دوم: سپس فاصله‌ی میان U و A با استفاده از معیار «فاصله‌ی اشتراک پروفایل‌های ساده‌شده[2]» (یا SPID) محاسبه می‌شود:

[1] Simplified Profile (SP)
[2] Simplified Profile Intersection Distance (SPID)

$$Dist^{CNG-SPI}(\mathrm{U},\mathrm{A}) = SPID(U,A) = 1 - \frac{|SPI(\mathrm{U},\mathrm{A})|}{L} \qquad (۳-۴)$$

که $|SPI(\mathrm{U},\mathrm{A})|$ برابرست با تعداد اعضای مجموعه‌ی $SPI(\mathrm{U},\mathrm{A})$. با توجه به تعریف $SPI(\mathrm{U},\mathrm{A})$، عبارت $|SPI(\mathrm{U},\mathrm{A})|$ را می‌توان به صورت «تعداد انگرام‌های مشترک بین پروفایل‌های $PR_U^{CNG-SPI}$ و $PR_A^{CNG-SPI}$» تعبیر نمود.

۴-۲ روش سوم: روش CNG-D1

این روش برای اولین‌بار در سال ۲۰۰۷ میلادی در مقاله‌ی [۲۶] معرفی شد. این روش روند جدید برای استخراج پروفایل ارائه نداده و فقط تغییر کوچکی در نحوه‌ی محاسبه‌ی فاصله‌ی میان U و A ایجاد می‌نماید.

۴-۲-۱ فلسفه‌ی ارائه‌ی روش

آزمایش‌های صورت‌گرفته بر روی مجموعه‌متن‌های متفاوت، نشان داده است که اگر متون موجود برای یک یا چند مورد از نویسندگان دارای طول کوتاهی باشند، روش CNG (خصوصاً به ازای مقادیر بزرگ L) عملکرد چندان مناسبی از خود ارائه نمی‌دهد. به منظور رفع این محدودیت، روش D1 تغییراتی در معیار فاصله‌ی مورداستفاده در روش CNG (یعنی معیار RD) ایجاد می‌نماید. از بخش ۳-۳ به یاد داریم که معیار RD به هر انگرام موجود در مجموعه‌ی $G = G_U^{CNG} \cup G_A^{CNG}$ یک فاصله‌ی خاص تخصیص داده و در نهایت تمامی این فواصل را جمع می‌نمود. همان‌طور که در ادامه خواهیم دید، روش D1 تغییری در نحوه‌ی تعریف این مجموعه‌ی G ارائه می‌دهد.

۴-۲-۲ سیستم استخراج پروفایل

روش D1 دقیقاً همان روند استخراج پروفایل ذکرشده برای روش CNG را پی می‌گیرد. به عبارت دیگر روش D1 نیز از پروفایل CP استفاده می‌نماید. پس می‌توان چنین نوشت:

$$PR_T^{CNG-D1} = \left\{ G_T^{CNG}, F_T^{CNG} \right\} \qquad (۴-۴)$$

۴-۲-۳ سیستم محاسبه‌ی فاصله

همان‌طور که گفته شد روش D1 معیار جدیدی را برای محاسبه‌ی فاصله‌ی میان U و A ارائه می‌دهد. این معیار فاصله که اصطلاحاً RD1 نامیده می‌شود، عبارتست از:

$$Dist^{CNG-D1}(U,A) = RD1(U,A) = \sum_{g \in G} d^{CNG}(g) \qquad (۵-۴)$$

که در آن $d^{CNG}(g)$ از رابطه‌ی ۳-۱۰ قابل محاسبه است. همان‌طور که ملاحظه می‌گردد قالب کلی دو معیار RD و RD1 یکسان می‌باشد، اما در نحوه‌ی تعریف مجموعه‌ی G با هم تفاوت دارند. در معیار RD1 این مجموعه به صورت اجتماع دو مجموعه‌ی G_U^{CNG} و G_A^{CNG} تعریف می‌شود (رابطه‌ی ۳-۸)، در حالیکه در معیار RD1 مجموعه‌ی G به صورت زیر تعریف می‌گردد:

$$G = G_U^{CNG} \qquad (۶-۴)$$

به عبارت دیگر در معیار RD1، فقط فواصل متناظر با انگرام‌های موجود در پروفایل متن دیده‌نشده تجمیع شده و (بر خلاف معیار RD) انگرام‌های موجود در پروفایل نویسنده نادیده انگاشته می‌شوند.

۴-۳ روش چهارم: روش CNG-D2

همانطور که در جدول ۳-۲ آمده است، روش‌های D1 و D2 در یک مقاله ارائه گردیده‌اند. روش D2 نیز همانند روش D1 همان پروفایل CP را در دستور کار قرار داده، اما معیار جدیدی برای محاسبه‌ی فاصله ارائه می‌دهد که به معیار RD2 معروف است.

۴-۳-۱ فلسفه‌ی ارائه‌ی روش

روش D2 را در واقع می‌توان نسخه‌ای تکامل‌یافته از روش D1 به حساب آورد. در این روش ابتدا به ازای هر انگرام یک مقدار فرکانس استاندارد محاسبه می‌شود. سپس تجمیع فاصله‌های متناظر با انگرام‌های موجود در مجموعه‌ی G به صورت وزن‌دار صورت می‌پذیرد. ایده‌ی اصلی این وزن‌دهی بدین صورت است: هر چه اختلاف میان فرکانس یک انگرام در متن دیده‌نشده و فرکانس استاندارد آن انگرام بیش‌تر باشد، وزن بیش‌تری به خود اختصاص خواهد داد. این نوع وزن‌دهی موجب می‌شود تا انگرام‌هایی که فرکانس استفاده

از آن‌ها با مقدار معمول (فرکانس استاندارد) تفاوت زیادتری دارد، تأثیر بیش‌تری در محاسبه‌ی فاصله‌ی میان U و A داشته باشند. نحوه‌ی محاسبه‌ی فرکانس استاندارد یک انگرام و نحوه‌ی وزن‌دهی در ادامه شرح داده خواهد شد.

۴-۳-۲ سیستم استخراج پروفایل

همانند روش D1، در روش D2 نیز پروفایل CP در دستور کار قرار می‌گیرد. به عبارت دیگر:

$$PR_T^{CNG-D2} = \left\{ G_T^{CNG}, F_T^{CNG} \right\} \qquad (۴-۷)$$

۴-۳-۳ سیستم محاسبه‌ی فاصله

معیار RD2 جهت محاسبه‌ی فاصله‌ی میان U و A روندی دارای دو گام اصلی پیشنهاد می‌دهد. گام اول کاملاً مستقل از متن دیده‌نشده بوده و با توجه به متون موجود در مجموعه‌متن صورت می‌پذیرد. این گام‌ها عبارتند از:

گام اول: استخراج پروفایل مجموعه‌متن

این گام خود شامل دو مرحله است:

۱) ابتدا می‌بایست ابرمتن‌های تمامی نویسندگان را به هم متصل نموده تا «ابرمتن مجموعه‌متن[1]» یا «ابرمتن کل[2]» تشکیل شود. این ابرمتن را با نماد C نمایش می‌دهیم.
۲) سپس با استفاده از روش CNG، پروفایل ابرمتن کل را استخراج می‌نماییم.

در پایان این مرحله، پروفایل استخراج‌شده از C را در اختیار خواهیم داشت که به صورت زیر است:

$$PR_C^{CNG} = \left\{ G_C^{CNG}, F_C^{CNG} \right\} \qquad (۴-۸)$$

مقادیر استاندار بیان‌شده در ۳-۵-۳-۱ در واقع فرکانس‌های بدست آمده توسط این پروفایل می‌باشند. به عبارت دیگر به ازای هر انگرام دلخواه (g)، مقدار فرکانس استاندارد آن برابرست با:

[1] Corpus Super-Text
[2] Total Super-Text

$$f^{standard}(g) = f_C^{CNG}(g) \qquad (۹-۴)$$

به طور خلاصه می‌توان گفت: «فرکانس استاندارد[1] انگرام g عبارتست از فرکانس نرمال‌شده‌ی آن انگرام در پروفایل استخراج‌شده از ابرمتن کل»

گام دوم: محاسبه‌ی فاصله

در این گام فاصله‌ی میان U و A با استفاده از معیار RD2 محاسبه می‌گردد. معیار RD2 که می‌توان آن را نسخه‌ی وزن-دار معیار RD1 در نظر گرفت، به صورت زیر می‌باشد:

$$Dist^{CNG-D2}(U,A) = RD2(U,A) = \sum_{g \in G} d^{CNG}(g) \times w^{D2}(g) \qquad (۱۰-۴)$$

که در آن $d^{CNG}(g)$ همان عبارت موجود در معیار RD بوده و از رابطه‌ی ۳-۱۰ محاسبه می‌شود. هم‌چنین $w^{D2}(g)$ در واقع وزن متناظر با انگرام g بوده و بدین‌صورت تعریف می‌گردد:

$$w^{D2}(g) = \left[\frac{f_U^{CNG}(g) - f_C^{CNG}(g)}{f_U^{CNG}(g) + f_C^{CNG}(g)} \right]^2 \qquad (۱۱-۴)$$

روابط فوق را می‌توان بدین صورت تفسیر نمود: به ازای هر انگرام مثل g، فاصله‌ی متناظر با آن ($d^{CNG}(g)$) دارای یک وزن می‌باشد. این وزن که با نماد $w^{D2}(g)$ نمایش داده می‌شود، در واقع بیان‌گر اختلاف میان فرکانس g در متن دیده‌نشده ($f_U^{CNG}(g)$) و مقدار استاندارد فرکانس این انگرام ($f_C^{CNG}(g)$) می‌باشد. لازم بذکرست که اگر فرکانس یک انگرام (مثلاً gx) در متن دیده‌نشده دقیقاً برابر با مقدار استاندارد فرکانس این انگرام باشد، وزن متناظر با انگرام gx برابر صفر شده و لذا این انگرام تأثیری در فاصله‌ی کل بین U و A نخواهد داشت.

۴-۴ روش پنجم: روش WPI

روش WPI که اولین‌بار در سال ۲۰۱۱ میلادی و در مقاله‌ی [۲۷] معرفی شد، در واقع نسخه‌ای ارتقاءیافته از روش SPI می‌باشد. این روش همانند روش SPI از پروفایل ساده‌شده (SP) استفاده نموده، اما معیار

[1] Standard Frequency

فاصله‌ی جدیدی به جای SPID (رابطه‌ی ۳-۱۴) معرفی می‌نماید. این معیار جدید را «فاصله‌ی اشتراک پروفایل وزن-دار[1]» نامیده و با نماد WPID نمایش می‌دهیم.

۴-۴-۱ فلسفه‌ی ارائه‌ی روش

به یاد داریم که در روش SPI، تعداد انگرام‌های مشترک بین دو پروفایل به عنوان مبنای سنجش فاصله مورداستفاده قرار می‌گرفت. روش WPI نیز همین رویه را در پیش می‌گیرد، با این تفاوت که به هر انگرام یک وزن اختصاص می‌دهد. ایده‌ی اصلی این وزن‌دهی بدین‌صورت قابل بیان است: انگرامی که در پروفایل تعداد کم‌تری از نویسندگان قرار داشته باشد، دارای اهمیت بیش‌تری (وزن بیش‌تری) است. به عناون مثال اگر انگرام gx در پروفایل ۴ نویسنده و انگرام gy در پروفایل ۲ نویسنده وجود داشته باشد، آن‌گاه انگرام gy خاص‌تر بوده و لذا وزن بیش‌تری نسبت به gx به خود اختصاص خواهد داد.

در نهایت، فاصله‌ی بین U و A به جای آن‌که بر مبنای تعداد انگرام‌های مشترک بین دو پروفایل محاسبه گردد (معیار SPID) بر مبنای میانگین وزن‌های متناظر با انگرام‌های مشترک بین دو پروفایل محاسبه می‌گردد. روند این وزن‌دهی و محاسبه‌ی فاصله در ادامه تشریح خواهد شد.

۴-۴-۲ سیستم استخراج پروفایل

در این روش (همانند روش SPI) برای استخراج پروفایل T، از پروفایل ساده‌شده (SP) استفاده می‌شود. لذا می‌توان نوشت:

$$PR_T^{CNG-WPI} = PR_T^{SPI} = \left\{G_T^{CNG}\right\} \qquad (۱۲-۴)$$

۴-۴-۳ سیستم محاسبه‌ی فاصله

روندی که معیار WPID جهت محاسبه‌ی فاصله‌ی میان U و A ارائه می‌دهد، دارای دو گام اصلی است. همانند معیار RD2، در این روند نیز گام اول کاملاً مستقل از متن دیده‌نشده بوده و با توجه به متون موجود در مجموعه‌متن صورت می‌پذیرد. این گام‌ها عبارتند از:

[1] Weighted Profile Intersection Distance (WPID)

گام اول: محاسبه‌ی وزن متناظر با هر انگرام

این گام خود شامل سه مرحله است:

۱) ابتدا پروفایل تک تک نویسنده‌ها را با استفاده از روش CNG استخراج می‌نماییم.
۲) سپس مجموعه‌ی G_C را با اجتماع بر روی انگرام‌های موجود در تمامی پروفایل‌های استخراج‌شده در مرحله‌ی قبل تشکیل می‌دهیم. به عبارت دیگر:

$$G_C = \bigcup_{j=1}^{M} G_{A_j}^{CNG} \qquad (۱۳-۴)$$

۳) در نهایت به ازای هر انگرام موجود در مجموعه‌ی G_C، یک مقدار وزن محاسبه می‌کنیم:

$$w_g = \frac{1}{\sum_{j=1}^{M} I_{g \in G_{A_j}^{CNG}}} \qquad (۱۴-۴)$$

که در آن:

$$I_{g \in G_{A_j}^{CNG}} = \begin{cases} 1, & if\ g \in G_{A_j}^{CNG} \\ 0, & if\ g \notin G_{A_j}^{CNG} \end{cases} \qquad (۱۵-۴)$$

به عبارت دیگر وزن متناظر با انگرام g عبارتست از: «معکوس تعداد نویسندگانی که انگرام g در پروفایل آن‌ها قرار دارد.» لذا انگرامی که در پروفایل تعداد کمتری از نویسندگان وجود داشته باشد، دارای وزن بالاتری خواهد بود.

از آن‌جا که هر انگرام موجود در مجموعه‌ی G_C، حداقل در پروفایل یکی از نویسندگان وجود دارد، لذا مقدار وزن متناظر با یک انگرام همواره در بازه‌ی زیر خواهد بود:

$$\frac{1}{M} \le w_g \le 1 \qquad (۱۶-۴)$$

گام دوم: محاسبه‌ی فاصله

این گام شامل دو مرحله است:

۱) ابتدا با استفاده از رابطه‌ی زیر شباهت میان پروفایل U و A را محاسبه می‌نماییم:

$$WPI(U,A)=\frac{1}{|SPI(U,A)|}\sum_{g\in SPI(U,A)} w_g \qquad (۱۷-۴)$$

مجموعه‌ی $SPI(U, A)$ در واقع همان مجموعه‌ی معرفی‌شده در روش SPI بوده و به منظور تشکیل آن می‌بایست از رابطه‌ی ۳-۱۳ استفاده نمود. هم‌چنین $|SPI(U,A)|$ برابرست با تعداد اعضای مجموعه‌ی $SPI(U,A)$.

لازم به ذکر است که $WPI(U,A)$ در واقع برابرست با میانگین وزن‌های متناظر با انگرام‌های موجود در مجموعه‌ی $SPI(U,A)$. لذا می‌توان گفت:

$$\frac{1}{M}\le WPI(U,A)\le 1 \qquad (۱۸-۴)$$

۲) با توجه به رابطه‌ی ۳-۲۹ برای تبدیل کمیت $WPI(U,A)$ (که در واقع یک معیار سنجش شباهت است) به یک معیار فاصله، کافیست آن را معکوس نماییم. این معیار را WPID نامیده و بدین‌صورت تعریف می‌نماییم:

$$Dist^{CNG-WPI}(U,A)=WPID(U,A)=\frac{1}{WPI(U,A)} \qquad (۱۹-۴)$$

۴-۵ روش ششم: روش RLP

روش RLP اولین‌بار در سال ۲۰۱۱ میلادی و در مقاله‌ی[۲۸] معرفی شد. این روش نوع جدیدی از پروفایل را ارائه نموده و معیار جدیدی نیز برای محاسبه‌ی فاصله‌ی میان این نوع پروفایل‌ها پیشنهاد می‌دهد. هرچند این روش نیز بر مبنای روش CNG ارائه شده است، اما تغییرات زیادی در آن ایجاد نموده است. با توجه به حجم تغییرات صورت گرفته، روش RLP را می‌توان نقطه‌ی عطفی در روش‌های NDP دانست.

۴-۵-۱ فلسفه‌ی ارائه‌ی روش

مبنای اصلی روش RLP نوع جدیدی از فرکانس است که جایگزین فرکانس ارائه‌شده در روش CNG می‌گردد. این فرکانس که «فرکانس باز-کانونی‌شده[1]» نام دارد، خود با استفاده از روش CNG محاسبه

[1] Re-centered Frequency

می‌شود. به منظور محاسبه‌ی این نوع از فرکانس برای یک انگرام دلخواه (g) در متن T، می‌بایست سه مرحله را طی نمود:

۱) ابتدا پروفایل کامل (یعنی به ازای L=Infinite) متن را با استفاده از روش CNG استخراج می‌نماییم: PR_T^{CNG}

۲) سپس پروفایل کامل ابرمتن کل را (که از کنار هم قرار دادن ابرمتن‌های تمامی نویسندگان ایجاد می‌شود)، با استفاده از روش CNG استخراج می‌کنیم: PR_C^{CNG}

۳) در نهایت فرکانس باز-کانونی‌شده‌ی متناظر با انگرام g در متن T بدین‌صورت محاسبه می‌گردد:

$$f_T^{RLP}(g) = f_T^{CNG}(g)\Big|_{L=\infty} - f_C^{CNG}(g)\Big|_{L=\infty} \qquad (۲۰-۴)$$

به عبارت دیگر فرکانس باز-کانونی‌شده‌ی یک انگرام در یک متن، در واقع حاصل تفاضل فرکانس نرمال‌شده‌ی آن انگرام در دو متن خواهد بود: متن T و ابرمتن کل (C).

با توجه به موارد فوق، مفهوم فرکانس باز-کانونی‌شده را می‌توان چنین شرح داد: ابتدا با محاسبه‌ی فرکانس نرمال‌شده‌ی انگرام g در ابرمتن کل، می‌توان به این نکته پی برد که به طور معمول انگرام g با چه فرکانسی در متون ظاهر می‌شود. پس $f_C^{CNG}(g)\Big|_{L=\infty}$ در واقع میزان متداول استفاده از انگرام g را نشان می‌دهد. حال می‌توان گفت مقدار عددی فرکانس باز-کانونی‌شده بیان می‌کند که فرکانس استفاده از انگرام g در متن T، تا چه حد مطابق میزان متداول استفاده از این انگرام می‌باشد.

در انتها ذکر این نکته ضروری به نظر می‌رسد که فرکانس باز-کانونی‌شده (بر خلاف فرکانس نرمال‌شده که همواره در بازه‌ی صفر تا یک است) می‌تواند مقادیر منفی نیز به خود بگیرد و در حالت کلی بین بازه‌ی بین ۱- تا ۱+ تغییر می‌کند. در ادامه خواهیم دید که این نوع جدید از فرکانس، مبنای استخراج پروفایل در روش RLP خواهد بود.

۴-۵-۲ سیستم استخراج پروفایل

از مباحث قبلی به یاد داریم که در روش CNG پروفایل یک متن بدین صورت تعریف می‌شود: «مجموعه‌ی L-انگرام پرتکرار (یعنی دارای بیش‌ترین فرکانس نرمال‌شده)، به همراه فرکانس نرمال‌شده‌ی تک تک آن‌ها». در روش RLP مشابه این تعریف اما این‌بار بر مبنای فرکانس باز-کانونی‌شده مورد استفاده قرار می‌گیرد: «پروفایل یک متن عبارتست از مجموعه‌ی L-انگرام دارای بیش‌ترین فرکانس باز-کانونی‌شده، به همراه

مجموعه‌ی دیگری که شامل فرکانس باز-کانونی‌شده‌ی تک تک این انگرام‌ها باشد». به عبارت این نوع پروفایل که آن را «پروفایل محلی باز-کانونی‌شده[1]» نامیده و با نماد RLP نمایش می‌دهند، به صورت زیر تعریف می‌شود:

$$PR_T^{RLP} = \left\{ G_T^{RLP}, F_T^{RLP} \right\} \qquad (۲۱-۴)$$

که در آن:

- G_T^{RLP} مجموعه‌ای شامل L-انگرامِ دارای بیش‌ترین فرکانس باز-کانونی‌شده می‌باشد.
- F_T^{RLP} مجموعه‌ای شامل فرکانس باز-کانونی‌شده‌ی تک تک اعضای G_T^{RLP} است.

۴-۵-۳ سیستم محاسبه‌ی فاصله

در روش RLP به منظور محاسبه‌ی فاصله‌ی میان U و A می‌بایست سه گام اساسی را طی نماییم:

گام اول: تشکیل مجموعه‌ی G

ابتدا مجموعه‌ی G را که در واقع حاصل اجتماع انگرام‌های موجود در پروفایل‌های U و A می‌باشد، تشکیل می‌دهیم:

$$G = G_U^{RLP} \cup G_A^{RLP} = \{ g_1, ..., g_k, ..., g_{L'} \} \qquad (۲۲-۴)$$

تعداد اعضای این مجموعه را با نماد 'L نمایش می‌دهیم.

گام دوم: محاسبه‌ی بردارهای P_U^{RLP} و P_A^{RLP}

بردارهای P_U^{RLP} و P_U^{RLP}، بردارهایی شامل 'L المان می‌باشند که هر کدام از این المان‌ها متناظر با یکی از انگرام‌های موجود در مجموعه‌ی G است. به عبارت دیگر المان‌های k-ام از بردارهای P_U^{RLP} و P_A^{RLP} متناظرند با انگرام k-ام موجود در مجموعه‌ی G. این بردارها بدین‌صورت محاسبه می‌شوند:

- المان k-ام از بردار P_U^{RLP}، برابرست با فرکانس باز-کانونی‌شده‌ی انگرام g_k در متن U.
- المان k-ام از بردار P_A^{RLP}، برابرست با فرکانس باز-کانونی‌شده‌ی انگرام g_k در ابرمتن A.

[1] Re-centered Local Profile (RLP)

به عبارت دیگر:

$$P_U^{RLP}[k] = f_U^{RLP}(g_k) \qquad (۲۳-۴)$$

$$P_A^{RLP}[k] = f_A^{RLP}(g_k) \qquad (۲۴-۴)$$

گام سوم: محاسبه‌ی فاصله

در روش RLP، فاصله‌ی میان U و A برابر با فاصله‌ی میان دو بردار P_U^{RLP} و P_U^{RLP} در نظر گرفته می‌شود. به عبارت دیگر مسئله‌ی محاسبه‌ی فاصله‌ی میان دو پروفایل (PR_U^{RLP} و PR_A^{RLP}) به مسئله‌ی محاسبه‌ی فاصله‌ی میان دو بردار (P_U^{RLP} و P_A^{RLP}) تبدیل می‌شود. لذا می‌توان از هر کدام از معیارهای معرفی‌شده برای محاسبه‌ی فاصله‌ی بین دو بردار (از جمله فاصله‌ی اقلیدسی[1]، منهتن[2] و ...) استفاده نمود. ارائه‌دهندگان روش RLP، معیاری با نام «نسخه‌ی اصلاح‌شده‌ی فاصله‌ی کسینوسی[3]» پیشنهاد داده‌اند. طبق این معیار که آن را با نماد MCD نمایش می‌دهیم، فاصله‌ی میان U و A بدین‌صورت محاسبه می‌شود:

$$Dist^{RLP}(U,A) = MCD(U,A) = 1 - \frac{\left\langle P_U^{RLP} \bullet P_A^{RLP} \right\rangle}{\left\| P_U^{RLP} \right\|^2 \ \left\| P_A^{RLP} \right\|^2} \qquad (۲۵-۴)$$

که در آن:

- منظور از $\left\langle P_U^{RLP} \bullet P_A^{RLP} \right\rangle$ در واقع همان ضرب داخلی دو بردار P_U^{RLP} و P_A^{RLP} می‌باشد.
- $\| P_T \|$ و $\| P_A \|$ به ترتیب نمایانگر اندازه‌ی بردارهای P_U^{RLP} و P_A^{RLP} هستند.

۴-۶ روش هفتم: روش RLP-IAF

روش RLP-IAF که در واقع نسخه‌ی تکامل‌یافته‌ی روش RLP می‌باشد، برای اولین‌بار در سال ۲۰۱۲ میلادی و در مقاله‌ی [۴۵] معرفی شد. این روش نوع جدیدی از پروفایل را ارائه نمی‌دهد، بلکه فقط در نحوه‌ی محاسبه‌ی فاصله تغییراتی ایجاد می‌کند.

[1] Euclidean Distance
[2] Manhattan Distance
[3] Modifier Version of Cosine Distance (MCD)

۴-۶-۱ فلسفه‌ی ارائه‌ی روش

از مطالب گذشته به یاد داریم که در روش RLP، دو بردار P_U^{RLP} و P_U^{RLP} مبنای محاسبه‌ی فاصله می‌باشند. ایده‌ی اصلی روش RLP-IAF در واقع وزن‌دار کردن المان‌های بردارهای P_U^{RLP} و P_A^{RLP} می‌باشد. از آن‌جا که هر یک از المان‌های این دو بردار متناظر با یک انگرام می‌باشند، لذا این روش در واقع انگرام‌ها را وزن‌دهی می‌نماید. مبنای این وزن‌دهی (مشابه با روش WPI)، تعداد نویسندگانی است که از آن انگرام استفاده نموده‌اند. نحوه‌ی محاسبه‌ی وزن متناظر با هر انگرام در ادامه شرح داده خواهد شد.

۴-۶-۲ سیستم استخراج پروفایل

همانطور که گفته شد روش RLP-IAF برای استخراج پروفایل یک متن دلخواه (T)، همان روند معرفی‌شده در روش RLP را مورداستفاده قرار می‌دهد. به عبارت دیگر:

$$PR_T^{RLP-IAF} = PR_T^{RLP} = \left\{G_T^{RLP}, F_T^{RLP}\right\} \qquad (۲۶-۴)$$

۴-۶-۳ سیستم محاسبه‌ی فاصله

در روش RLP-IAF برای محاسبه‌ی فاصله‌ی میان U و A می‌بایست چهار گام را طی نمود:

گام اول: تشکیل مجموعه‌ی G

در این گام، همانند روش RLP مجموعه‌ی G با اجتماع بر روی انگرام‌های موجود در دو پروفایل تشکیل می‌شود:

$$G = G_U^{RLP} \cup G_A^{RLP} = \{g_1, ..., g_k, ..., g_{L'}\} \qquad (۲۷-۴)$$

گام دوم: محاسبه‌ی iaf به ازای هر عضو G

در گام دوم، به ازای هر انگرام موجود در مجموعه‌ی G یک مقدار به عنوان «فرکانس معکوس نویسنده[1]» محاسبه می‌نماییم. این نوع فرکانس که با نماد *iaf* نمایش داده می‌شود، به صورت زیر قابل محاسبه است:

[1] Author Inverse Frequency (IAF)

$$iaf\ (g) = \log(\frac{M}{\sum_{j=1}^{M} I_{g \in G_{A_j}^{RLP}}}) \qquad (۲۸-۴)$$

که در آن M برابر با تعداد نویسندگان موجود در مجموعه‌متن بوده و عبارت $I_{g \in G_{A_j}^{RLP}}$ (مشابه رابطه‌ی ۳-۲۶) بدین صورت محاسبه می‌شود:

$$I_{g \in G_{A_j}^{RLP}} = \begin{cases} 1, & if\ g \in G_{A_j}^{RLP} \\ 0, & if\ g \notin G_{A_j}^{RLP} \end{cases} \qquad (۲۹-۴)$$

به عبارت دیگر $iaf\ (g)$ عبارتست از لگاریتم عکس نسبت زیر: (به همین جهت آن را فرکانس معکوس نویسنده می‌نامند)

نسبت (تعداد نویسندگانی که انگرام g در پروفایل آن‌ها وجود دارد) به (تعداد کل نویسندگان)

گام سوم: تشکیل بردارهای $P_U^{RLP-IAF}$ و $P_A^{RLP-IAF}$

بردارهای $P_U^{RLP-IAF}$ و $P_A^{RLP-IAF}$ شامل L' المان می‌باشند که هر یک از این المان‌ها متناظر با یک انگرام از مجموعه‌ی G است. در این گام جهت محاسبه‌ی این دو بردار از روندی مشابه آن‌چه در روش RLP برای محاسبه‌ی بردارهای P_U^{RLP} و P_A^{RLP} ارائه شد، استفاده می‌شود؛ با این تفاوت که در این روش هر المان از بردارهای P_U^{RLP} و P_A^{RLP} در مقدار iaf متناظر با آن المان ضرب می‌شود. به عبارت دیگر المان k-ام از بردار $P_U^{RLP-IAF}$ که با انگرام g_k متناظر است، بدین‌صورت محاسبه می‌شود:

$$P_U^{RLP-IAF}[k] = f_U^{RLP}(g_k) \times iaf\ (g_k) \qquad (۳۰-۴)$$

و به طور مشابه برای المان k-ام از بردار $P_A^{RLP-IAF}$ داریم:

$$P_A^{RLP-IAF}[k] = f_A^{RLP}(g_k) \times iaf\ (g_k) \qquad (۳۱-۴)$$

گام چهارم: محاسبه‌ی فاصله

در این گام، فاصله‌ی میان U و A با استفاده از معیاری شبیه به MCD محاسبه می‌شود. تنها تفاوت آنست که این‌بار به جای بردارهای P_U^{RLP} و P_A^{RLP}، از بردارهای $P_U^{RLP-IAF}$ و $P_A^{RLP-IAF}$ استفاده می‌شود. این معیار جدید که با نماد MCD-IAF نمایش داده می‌شود به صورت زیر تعریف می‌شود:

$$Dist^{RLP-IAF}(U,A) = \mathrm{MCDIAF}(U,A) = 1 - \frac{P_U^{RLP-IAF} \bullet P_A^{RLP-IAF}}{\left\|P_U^{RLP-IAF}\right\|^2 \times \left\|P_A^{RLP-IAF}\right\|^2} \qquad (۳۲-۴)$$

۴-۷ روش هشتم: روش O-SPI

روش O-SPI اولین‌بار در سال ۲۰۱۴ میلادی و در مقاله‌ی [۲۹] معرفی گردید. در این روش که در واقع از ایجاد یک تغییر جزئی در روش SPI ایجاد شده است، علاوه بر پارامترهای N و L، پارامتر جدید دیگری ارائه می‌شود: «آستانه‌ی فرکانس[1]». در ادامه این پارامتر را با نماد f_{th} نمایش می‌دهیم.

۴-۷-۱ فلسفه‌ی ارائه‌ی روش

اعمال روش SPI بر روی مجموعه‌متن‌های مختلف نشان داده است که انگرام‌های با تعداد تکرار کم (مثلاً انگرام‌هایی که فقط یک‌بار در متن آمده‌اند)، علاوه بر اینکه هیچ‌گونه اطلاعاتی در مورد سبک نوشتاری نویسنده‌ی متن در خود ندارند، گاهی به عنوان نویز عمل کرده و موجب کاهش دقت روش SPI می‌شوند. ایده‌ی اصلی روش O-SPI آنست که چنین انگرام‌هایی را از پروفایل متن حذف نماییم. این روش نوع جدیدی از پروفایل را ارائه داده، اما هم‌چنان معیار SPID را جهت محاسبه‌ی فاصله‌ی میان U و A به کار می‌برد.

۴-۷-۲ سیستم استخراج پروفایل

پروفایل ارائه‌شده در این روش، همانند پروفایل SP تنها دارای یک مجموعه می‌باشد:

$$PR_T^{O-SPI} = \left\{G_T^{O-SPI}\right\} \qquad (۳۳-۴)$$

که G_T^{O-SPI} عبارتست از: «مجموعه‌ی G_T^{CNG} پس از حذف انگرام‌هایی که دارای فرکانس کوچکتر یا مساوی با f_{th} می‌باشند». پس در روش O-SPI، به منظور استخراج پروفایل متن T باید دو گام را طی نمود:

۱) ابتدا می‌بایست پروفایل متن را با استفاده از روش CNG استخراج نموده و مجموعه‌های G_T^{CNG} و F_T^{CNG} را تشکیل داد.

[1] Frequency Threshold

۲) سپس می‌بایست تمامی انگرام‌هایی را که فرکانس متناظر آن‌ها مساوی با مقدار f_{th} و یا کمتر از آنست، از مجموعه‌ی G_T^{CNG} حذف نمود. مجموعه‌ی حاصل همان G_T^{O-SPI} خواهد بود.

این نوع پروفایل را که با حذف برخی انگرام‌ها از پروفایل ساده‌شده (SP) ایجاد می‌شود، اصطلاحاً «پروفایل ساده‌شده‌ی حذف‌شده[1]» نامیده و با نماد O-SP نمایش می‌دهیم.

در انتها مجدداً تأکید می‌نماییم که در تشکیل پروفایل O-SP، علاوه بر پارامترهای N و L، پارامتر f_{th} نیز نقش مهمی ایفا می‌نماید. البته این مقدار آستانه را معمولاً با استفاده از تعداد تکرار بیان می‌کنند، به عنوان مثال گفته می‌شود که تمامی انگرام‌هایی که تعداد تکرار آن‌ها ۳ و یا کمتر از آنست را باید حذف نمود. به عبارت دیگر در عمل به جای استفاده از f_{th}، از پارامتر دیگری به نام «تعداد تکرار آستانه[2]» استفاده می‌شود. این پارامتر که با نماد rc_{th} نمایش داده می‌شود، به صورت زیر با f_{th} در ارتباط است:

$$f_{th} = \frac{rc_{th}}{Q} \qquad (۳۴-۴)$$

که در آن Q تعداد کل نویسه‌های متن است. پژوهش‌های صورت‌گرفته بر روی مقادیر مختلف rc_{th} نشان داده است که معمولاً به ازای $rc_{th} \geq 4$ با افت شدید عملکرد روش O-SPI روبه‌رو می‌شویم. هم‌چنین بهترین پاسخ معمولاً به ازای $rc_{th} = 1$ حاصل می‌شود [۲۹].

۴-۷-۳ سیستم محاسبه‌ی فاصله

در این روش، برای محاسبه‌ی فاصله‌ی میان U و A معیار جدیدی با نام «فاصله‌ی اشتراک پروفایل ساده‌شده‌ی حذف‌شده[3]» ارائه می‌شود. قالب کلی این معیار (که آن را با نماد O-SPID نمایش می‌دهیم) با معیار SPID یکسان می‌باشد. به یاد داریم که معیار SPID برای محاسبه‌ی پروفایل‌های نوع SP به کار می‌رود. حال معیار O-SPID را می‌توان تعمیم‌یافته‌ی معیار SPID برای پروفایل O-SP دانست. این معیار به صورت زیر تعریف می‌شود:

$$Dist^{O-SPI}(U, A) = 1 - \frac{|OSPI(\mathrm{U}, \mathrm{A})|}{L} \qquad (۳۵-۴)$$

[1] Omitted Simplified Profile (O-SP)
[2] Threshold Repetition Count
[3] Omitted Simplified Profile Distance (OSPID)

که در آن:

$$OSPI(U,A) = G_U^{O-SPI} \bigcap G_A^{O-SPI} \qquad (۳۶-۴)$$

یعنی مبنای اصلی محاسبه‌ی فاصله‌ی میان U و A در معیار OSPID، در واقع تعداد انگرام‌های مشترک میان پروفایل‌های O-SP استخراج‌شده از U و A می‌باشد.

۴-۸ مقایسه و دسته‌بندی هشت روش معرفی‌شده

در این فصل با هشت مورد از مهم‌ترین روش‌های NDP آشنا شدیم. در انتهای فصل می‌خواهیم با ارائه‌ی یک دسته‌بندی از این روش‌ها، امکان درک بهتر آن‌ها را برای خواننده فراهم آوریم. برای آن‌که بتوانیم دسته‌بندی مناسبی از این هشت روش داشته باشیم، ابتدا دو سیستم اساسی این روش‌ها را مرور می‌کنیم: سیستم استخراج پروفایل و سیستم محاسبه‌ی فاصله. این مرور به طور خلاصه در جدول ۳-۴ آمده است.

جدول ۴-۱: مروری بر هشت روش معرفی‌شده در این فصل

ردیف	نام روش	نوع پروفایل	معیار فاصله
۱	CNG	پروفایل کلاسیک CP	فاصله‌ی نسبی RD
۲	CNG-SPI	پروفایل ساده‌شده SP	فاصله‌ی اشتراک پروفایل ساده‌شده SPID
۳	CNG-D1	پروفایل کلاسیک CP	معیار فاصله‌ی نسبی D1 RD1
۴	CNG-D2	پروفایل کلاسیک CP	معیار فاصله‌ی نسبی D2 RD2
۵	CNG-WPI	پروفایل ساده‌شده SP	فاصله‌ی اشتراک وزن‌دار پروفایل WPID
۶	RLP	پروفایل محلی باز-کانونی‌شده RLP	فاصله‌ی کسینوسی اصلاح‌شده MCD

۷	RLP-IAF	پروفایل محلی باز-کانونی‌شده RLP	فاصله‌ی کسینوسی اصلاح‌شده - فرکانس معکوس نویسنده MCD-IAF
۸	O-SPI	پروفایل ساده‌شده‌ی حذف‌شده O-SP	فاصله‌ی اشتراک پروفایل ساده‌شده‌ی حذف‌شده O-SPID

در ادامه جدول فوق را در دو سطح تحلیل می‌نماییم:

معیار فاصله:

همان‌طور که ملاحظه می‌شود، هیچ دو روشی دارای معیار فاصله‌ی یکسانی نمی‌باشند. لذا اگر بخواهیم روش‌ها را بر مبنای سیستم محاسبه‌ی فاصله دسته‌بندی نماییم، هشت دسته‌ی مختلف ایجاد می‌شود. واضح است که این نوع دسته‌بندی هیچ کمکی به درک بهتر روش‌ها نخواهد کرد.

نوع پروفایل:

هشت روش معرفی‌شده در مجموع چهار نوع پروفایل را مورداستفاده قرار می‌دهند: CP، SP، RLP و O-SP. در مورد این چهار نوع پروفایل می‌بایست به نکات زیر توجه نمود:

- پروفایل SP با حذف مجموعه‌ی فرکانس‌ها از پروفایل CP ایجاد می‌شود، لذا SP در واقع زیرمجموعه‌ای از CP می‌باشد.
- پروفایل O-SP با حذف برخی انگرام‌ها از پروفایل SP ایجاد می‌گردد، لذا O-SP را می‌توان زیرمجموعه‌ای از SP در نظر گرفت.
- پروفایل RLP روند کاملاً متفاوتی را نسبت به پروفایل CP ارائه می‌دهد، زیرا این پروفایل (بر خلاف CP که مبتنی بر فرکانس نرمال‌شده است)، مبتنی بر فرکانس باز-کانونی‌شده می‌باشد. پس پروفایل‌های RLP و CP دو نوع کاملاً متفاوت بوده و نمی‌توان یکی از آن‌ها را زیرمجموعه‌ی دیگری در نظر گرفت.

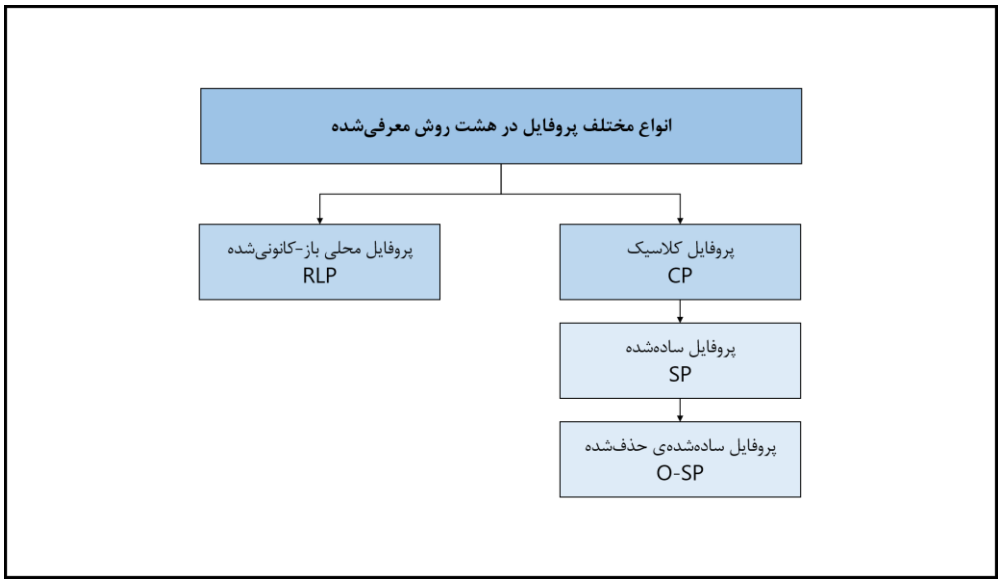
با توجه به مطالب فوق، می‌توان تمامی پروفایل‌هایی که تا کنون معرفی شده‌اند را به صورت شکل ۳-۱۳ تقسیم‌بندی نمود.

همانطور که در این شکل آمده است، در کل دو نوع پروفایل اصلی داریم و سایر پروفایل‌ها در واقع زیرمجموعه‌ای از این دو نوع پروفایل می‌باشند:

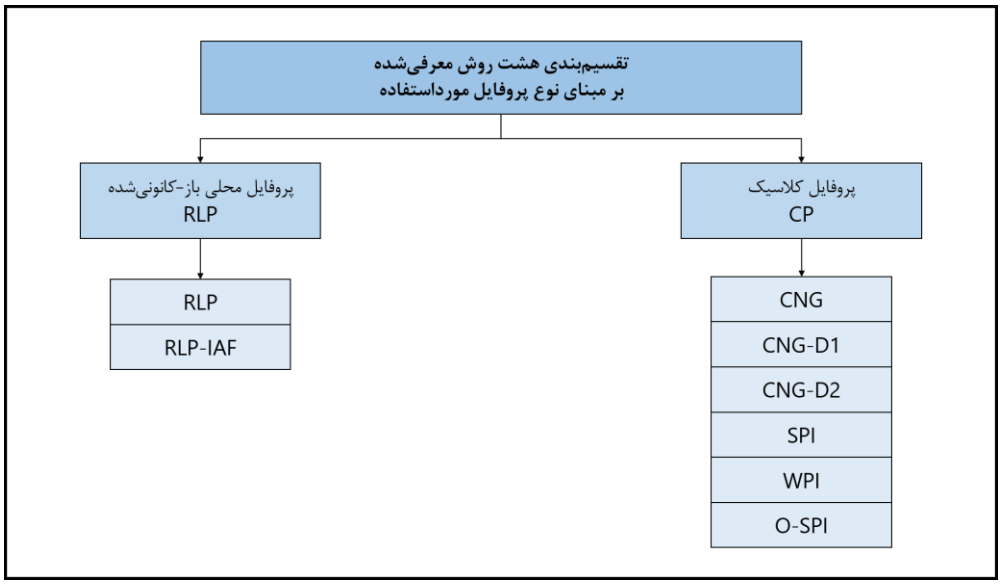
- پروفایل CP: که بر مبنای فرکانس نرمال‌شده تشکیل می‌شود.

- پروفایل RLP: که بر مبنای فرکانس باز-کانونی‌شده تشکیل می‌گردد.

اگر این دو نوع پروفایل را مبنای تقسیم‌بندی روش‌های معرفی‌شده قرار دهیم، روش‌ها در دو دسته جای می‌گیرند (شکل ۳-۱۴).

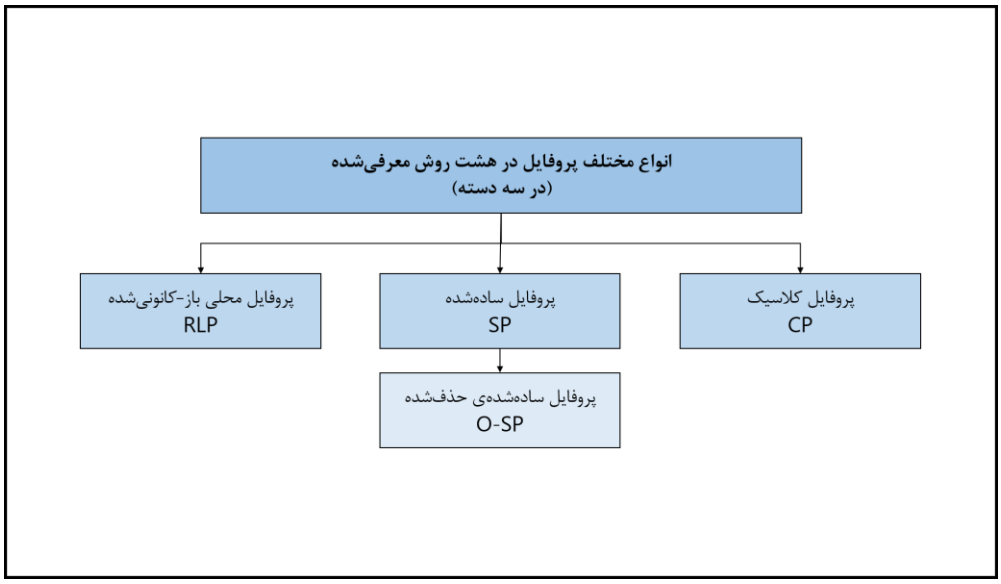


شکل ۴-۱: تقسیم‌بندی انواع پروفایل در روش‌های معرفی‌شده (به دو دسته‌ی اصلی)

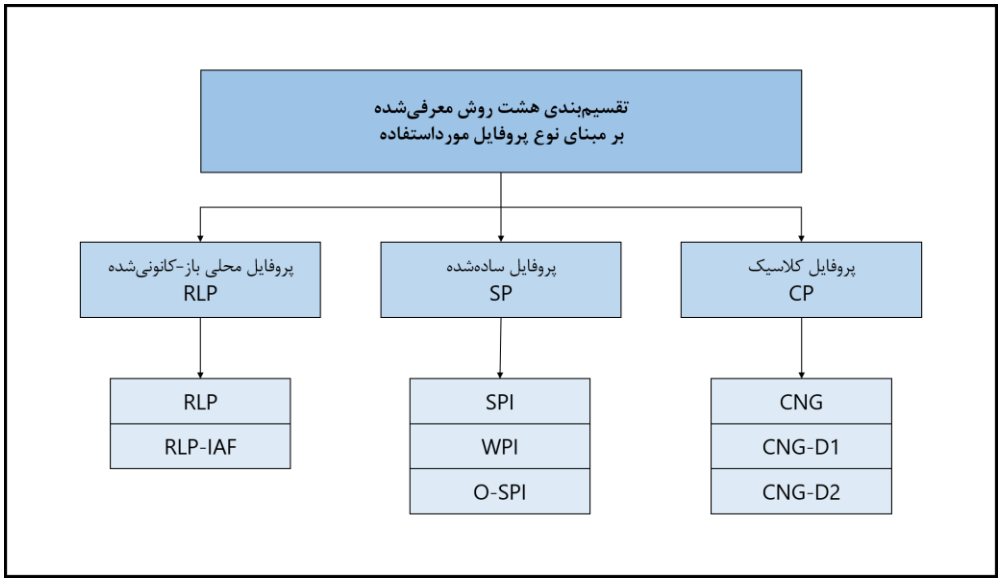


شکل ۴-۲: تقسیم‌بندی هشت روش معرفی‌شده در این فصل بر مبنای دو نوع اصلی پروفایل

با وجود اینکه این تقسیم‌بندی دید خوبی از روش‌های معرفی‌شده در اختیار ما قرار می‌دهد، اما به منظور کاراتر شدن این تقسیم بندی بهترست پروفایل SP (به همراه پروفایل O-SP) را به عنوان یک نوع جداگانه از پروفایل‌ها در نظر بگیریم. در نتیجه سه نوع پروفایل خواهیم داشت: CP، SP و RLP. لازم بذکرست در این حالت پروفایل O-SP به عنوان زیرمجموعه‌ای از پروفایل SP در نظر گرفته شده و در زمره‌ی انواع اصلی پروفایل‌ها به حساب نمی‌آید (شکل ۳-۱۵). بر مبنای این تقسیم‌بندی پروفایل‌ها، هشت روش معرفی‌شده را می‌توان همانند شکل ۳-۱۶ در سه دسته جای داد.



شکل ۴-۳: تقسیم‌بندی انواع پروفایل در روش‌های معرفی‌شده (به سه دسته)



شکل ۴-۴: تقسیم‌بندی هشت روش معرفی‌شده در این فصل بر مبنای سه دسته پروفایل

پیوست ۵:

مقایسه‌ی مجموعه‌متن‌های استفاده‌شده

پس از معرفی چهار مجموعه‌متن مورد استفاده در این پایان‌نامه، حال می‌خواهیم مقایسه‌ی مختصری از آن‌ها ارائه دهیم. این مقایسه بر مبنای چند ویژگی متفاوت صورت پذیرفته و ابتدا به صورت جداگانه بیان می‌شود. اما در انتهای بخش (قسمت ۴-۵-۷) تمامی این مقایسه‌ها به صورت یک‌جا جمع‌بندی می‌گردند.

۵-۱ بر مبنای رعایت ویژگی‌های مجموعه‌متن ایده‌آل

در بخش ۴-۴-۱ سه ویژگی برای یک مجموعه‌متن ایده‌آل در حوزه‌ی تخصیص نویسنده ذکر شد. این سه ویژگی را به طور خلاصه یادآوری می‌نماییم:

- ویژگی اول: هم‌دوره بودن نویسندگان موجود در مجموعه
- ویژگی دوم: هم-موضوع و هم-ژانر بودن متون انتخابی
- ویژگی سوم: مشابه بودن نویسندگان از نظر ویژگی‌هایی نظیر سن، تحصیلات و ...

حال می‌خواهیم بررسی کنیم که در هر کدام از چهار مجموعه‌ی معرفی‌شده، کدام ویژگی‌ها رعایت شده است. همان‌گونه که در جدول ۴-۱۷ آمده است:

- مجموعه‌های CPPT و RSK-40 ویژگی اول را رعایت نموده و از سایر ویژگی‌ها صرف نظر کرده‌اند.
- مجموعه‌ی BASU-2 علاوه بر رعایت ویژگی اول، تلاش نموده تا ویژگی‌های دوم و سوم را نیز در نظر بگیرد.
  - از آن‌جا که تمامی متون درباره با یک موضوع نوشته شده‌اند، این مجموعه دارای موضوع واحدی است؛ اما از نظر ژانر نیاز به بررسی متخصصین ادبیات دارد تا معلوم گردد آیا ژانر تمامی ۲۰ متن یکسان است یا خیر. گردآورندگان مجموعه، صحبتی از ژانر متون به میان نیاورده‌اند.
  - با توجه به اینکه تمامی نویسندگان، دانشجویان ترم اول مهندسی بوده اند، لذا این مجموعه تا حدودی در رعایت ویژگی سوم نیز موفق بوده است.

- مجموعه‌ی AAAT ویژگی‌های اول و دوم (تا حدودی) را رعایت نموده و از ویژگی سوم صرف نظر نموده است.

در انتها مطالب ارائه‌شده در این قسمت را جمع‌بندی نموده و مجموعه‌های معرفی‌شده بر مبنای نزدیک بودن به مجموعه‌متن ایده‌آل مرتب می‌کنیم:

- رتبه‌ی اول: BASU-2
- رتبه‌ی دوم: AAAT
- رتبه‌ی سوم: CPPT و RSK-40

جدول ۵-۱: مقایسه‌ی مجموعه‌متن‌ها بر مبنای رعایت کردن ویژگی‌های مجموعه‌ی ایده‌آل

نام مجموعه‌متن	ویژگی اول هم‌دوره بودن نویسندگان	ویژگی دوم موضوع و ژانر مشابه	ویژگی سوم ویژگی‌های مشابه نویسندگان
CPPT	رعایت شده	رعایت نشده	رعایت نشده
RSK-40	رعایت شده	رعایت نشده	رعایت نشده
BASU-2	رعایت شده	تا حدودی رعایت شده	تا حدودی رعایت شده
AAAT	رعایت شده	تا حدودی رعایت شده	رعایت نشده

۵-۲ بر مبنای زبان، ادبیات و دوره

در جدول ۴-۱۸ چهار مجموعه‌متن معرفی‌شده از نظر زبان (فارسی یا عربی)، ادبیات (کتابی یا محاوره‌ای) و دوره‌ی نگارش (معاصر یا کهن) مقایسه شده‌اند. از این جدول می‌توان نکات زیر را استخراج نمود:

- از میان مجموعه‌های معرفی‌شده، فقط مجموعه‌ی AAAT دارای زبان عربی بوده و سایر مجموعه‌ها به زبان فارسی می‌باشند.
- دو مجموعه‌ی CPPT و RSK-40 دارای متونی با ادبیات کتابی بوده و در دوره‌ی معاصر نوشته شده‌اند، اما مجموعه‌ی BASU هرچند همانند دو مورد قبلی در دوره‌ی معاصر نگاشته شده، اما بر خلاف آن‌ها دارای ادبیات محاوره‌ای است.
- مجموعه‌ی AAAT بر خلاف سایر مجموعه‌ها مربوط به دوره‌ی کهن می‌باشد.

جدول ۵-۲: مقایسه‌ی مجموعه‌متن‌های معرفی‌شده بر مبنای زبان، ادبیات و دوره

نام مجموعه	زبان	ادبیات	دوره
CPPT	فارسی	کتابی	معاصر
RSK-40	فارسی	کتابی	معاصر
BASU-2	فارسی	محاوره‌ای	معاصر
AAAT	عربی	کتابی	کهن

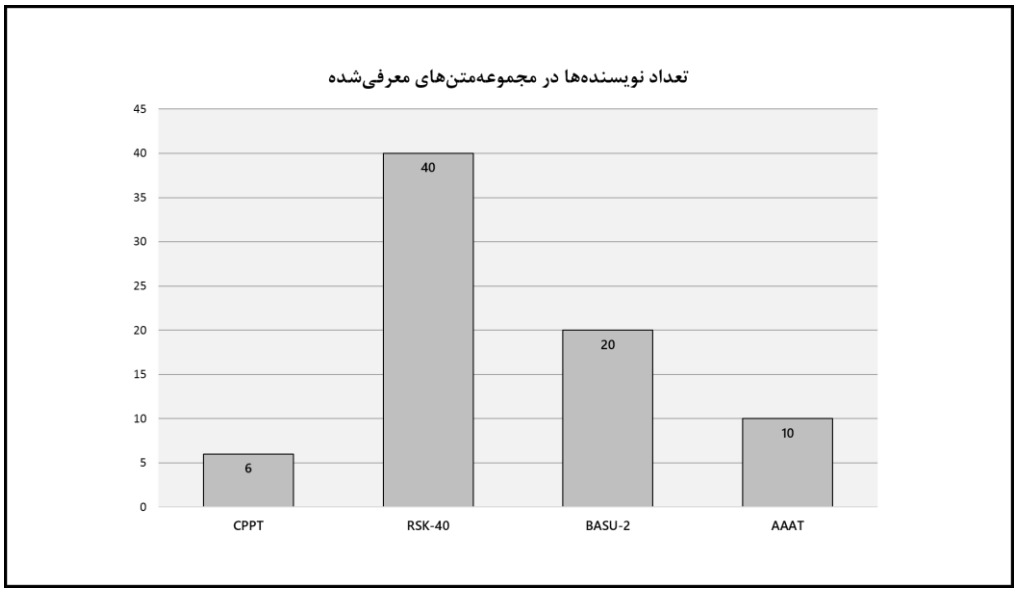
اهمیت این مقایسه زمانی آشکار می‌شود که بدانیم پژوهش‌های صورت‌گرفته در حوزه‌ی تخصیص نویسنده‌ی متون کتابی و محاوره‌ای، نشان داده است که تخصیص نویسنده در مجموعه‌های محاوره‌ای عموماً مشکل‌تر بوده و تقریباً در اکثر موارد دقت بدست‌آمده در این مجموعه‌ها، کمتر از مجموعه‌های کتابی است. لذا نمی‌بایست انتظار رسیدن به دقت بالا بر روی مجموعه‌متن BASU-2 داشت. فرهمندپور و همکاران [۴۹] در بهترین حالت دقتی معادل ۸۱/۶ درصد را گزارش نموده‌اند.

۵-۳ بر مبنای تعداد نویسنده

در این بخش می‌خواهیم چهار مجموعه‌ی معرفی‌شده را بر مبنای تعداد نویسندگان موجود در آن‌ها مقایسه نماییم. بدین منظور از جداول D ارائه‌شده برای هر مجموعه استفاده شده و نتایج حاصل در شکل ۴-۶ نمایش داده شده است.

در طراحی یک طبقه‌بند[1]، معمولاً هر چه تعداد طبقه‌ها بیش‌تر باشد، آن مسئله سخت‌تر است. تخصیص نویسنده را نیز می‌توان به صورت یک مسئله‌ی طبقه‌بندی[2] در نظر گرفته و هر نویسنده را معادل یک طبقه فرض نمود. در این صورت می‌توان نتیجه گرفت که هر چه تعداد نویسنده‌ها در یک مسئله‌ی تخصیص نویسنده بیش‌تر باشد، آن مسئله مشکل‌تر بوده و امید به دست یافتن به دقت‌های بالا در آن کمتر است.

[1] Classifier
[2] Classification



شکل ۵-۱: مقایسه‌ی مجموعه‌متن‌های معرفی‌شده بر مبنای تعداد نویسندگان موجود در مجموعه

از شکل ۴-۶ می‌توان موارد زیر را نتیجه گرفت:

- مجموعه‌متن RSK-40 با توجه به تعداد بالای نویسندگان موجود در آن، مسئله‌ی مشکلی به شمار رفته و انتظار دقت بالا در آن نمی‌رود. بالاترین دقت گزارش‌شده توسط رمضانی و همکاران [۵۱] بر روی این مجموعه، دقتی معادل ۶۴ درصد می‌باشد.
- مجموعه‌ی BASU-2 دارای تعداد نویسندگان تقریباً زیادی است و این امر موجب افزایش پیچیدگی آن می‌شود.
- مجموعه‌های CPPT و AAAT دارای تعداد نویسندگان متوسطی بوده و می‌توان آن‌ها را از این نظر مسائل نسبتاً آسانی در نظر گرفت.

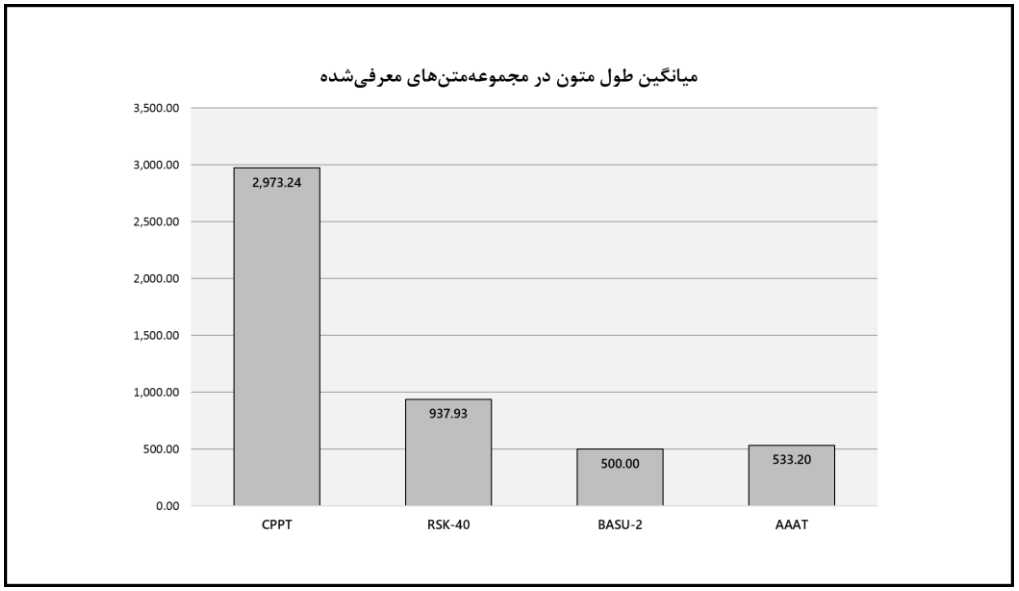
۵-۴ بر مبنای طول متون

در این بخش به بررسی طول متون در چهار مجموعه‌ی معرفی‌شده می‌پردازیم. سؤالی که ممکن است در ذهن ایجاد شود آن است که مقدار مناسب برای طول متون یک مجموعه‌متن چقدر است؟ به منظور پاسخ به این سؤال موارد زیر را در نظر می‌گیریم:

- در برخی مقالات (مثل [۹۱ و ۹۲]) یک قاعده‌ی سرانگشتی برای سنجش مناسب بودن طول متون یک مجموعه‌متن ارائه شده است: «کمینه‌ی طول مورد نیاز برای تخصیص نویسنده با دقت بالا، عبارتست از: حداقل ۲۵۰۰ کلمه در هر متن».
- البته تا کنون نتایج امیدوارکننده‌ای از کار بر روی متون کوتاه (حتی کوتاه‌تر از ۱۰۰۰ کلمه) ارائه شده است [۳۱].
- جالب‌تر آن‌که لایتون و همکاران بر روی متونی با طول ۱۴۰ نویسه و کمتر، به دقتی بالاتر از ۷۰ درصد رسیده اند [۹۳]!

پس در مجموع می‌توان چنین گفت که هرچند طول حداقل ۲۵۰۰ کلمه در هر متن برای رسیدن به دقت‌های بالا مناسب به نظر می‌رسد، اما این بدان معنا نیست که برای متونی با طول کمتر از آن، دسترسی به دقت مناسب غیر ممکن است. ولی در هر صورت می‌توان این گزاره را صحیح دانست: هر چه طول متون یک مجموعه‌متن بیش‌تر باشند، دقت بالاتری مورد انتظار است.

هرچند روش‌های NDP که در این پایان‌نامه بررسی می‌شوند، مبتنی بر نویسه هستند، اما (همانطور که در بخش ۴-۴-۲ بیان شد) از آن‌جا که در بررسی طول متون استفاده از معیار تعداد کلمات متداول‌تر از تعداد نویسه‌هاست، در این قسمت مقایسه‌ها را بر مبنای تعداد کلمات انجام می‌دهیم. اگر متوسط طول متون (تعداد کلمات) را مبنای مقایسه قرار دهیم، آن‌گاه چهار مجموعه‌ی معرفی‌شده، با استفاده از جداول C ارائه‌شده برای هر یک، به صورت شکل ۴-۷ مقایسه می‌شوند.



شکل ۵-۲: مقایسه‌ی مجموعه‌متن‌های معرفی‌شده بر مبنای متوسط طول متون موجود در مجموعه

از نمودار فوق می‌توان چنین نتایجی را استنباط نمود:

- مجموعه‌ی CPPT حتی با پذیرفتن معیار ۲۵۰۰ کلمه نیز، دارای طول مناسبی می‌باشد و لذا انتظار دقت بالا در آن می‌رود.
- مجموعه‌ی RSK-40 دارای طول متوسطی در حدود ۱۰۰۰ کلمه می‌باشد که چندان بالا نبوده و موجب افزایش پیچیدگی این مجموعه می‌شود.
- مجموعه‌های BASU-2 و AAAT به ترتیب دارای طول متوسط ۵۰۰ و ۵۳۳ کلمه می‌باشند که این مقدار برای رسیدن به دقت مناسب در تخصیص نویسنده، بسیار کم است. لذا نمی‌بایست انتظار دقت بالا در این مجموعه‌ها را داشت.

۵-۵ بر مبنای تعداد متون به ازای هر نویسنده

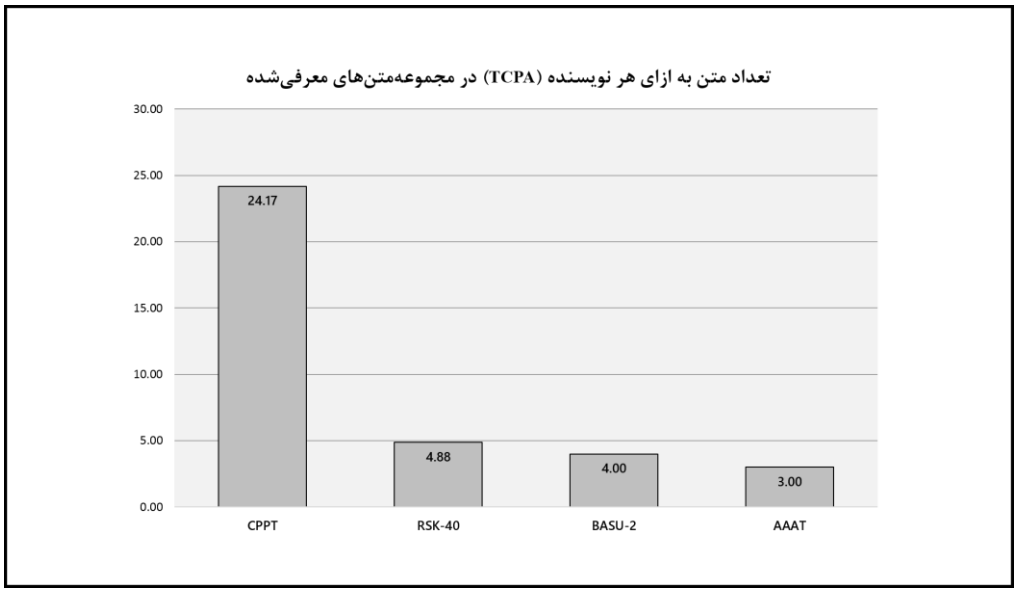
مجموعه‌متن‌های حوزه‌ی تخصیص نویسنده را می‌توان بر مبنای دیگری نیز مقایسه نمود: تعداد متون به ازای هر نویسنده[1]. این کمیت که آن را با نماد TCPA نشان می‌دهیم به صورت «تعداد کل متون تقسیم بر تعداد کل نویسندگان» تعریف می‌شود. هر چند این معیار معمولاً در الگوواره‌های مبتنی بر نمونه حائز اهمیت است، اما برای جامعیت بیش‌تر مقایسه‌های صورت گرفته در این فصل، مجموعه‌متن‌های معرفی‌شده را بر این مبنا نیز مقایسه می‌نماییم.

با استفاده از اطلاعات جداول D مجموعه‌ها، می‌توان نمودار شکل ۴-۸ را ترسیم نمود. هرچند تا کنون معیار خاصی برای تشخیص مناسب بودن یا نبودن TCPA در یک مجموعه ارائه نشده است، اما آنچه مسلم است هر چه مقدار این کمیت برای یک مجموعه‌متن بالاتر باشد، امید رسیدن به دقت‌های بالاتر برای در مجموعه افزایش می‌یابد.

از نمودار فوق می‌توان چنین نتیجه گرفت:

- مجموعه‌ی CPPT نسبت به سایر مجموعه‌ها در وضعیت بسیار بهتری قرار داشته و انتظار دقت بالا از آن می‌رود.
- مجموعه‌های RSK-40، BASU-2 و AAAT تعداد متن کمتری نسبت به CPPT در خود دارند و لذا از این منظر مجموعه‌های پیچیده‌تری به نظر می‌رسند.

[1] Text Count Per Author (TCPA)



شکل ۵-۳: مقایسه‌ی مجموعه‌متن‌های معرفی‌شده بر مبنای تعداد متن به ازای هر نویسنده

۵-۶ بر مبنای تعداد کلمات به ازای هر نویسنده

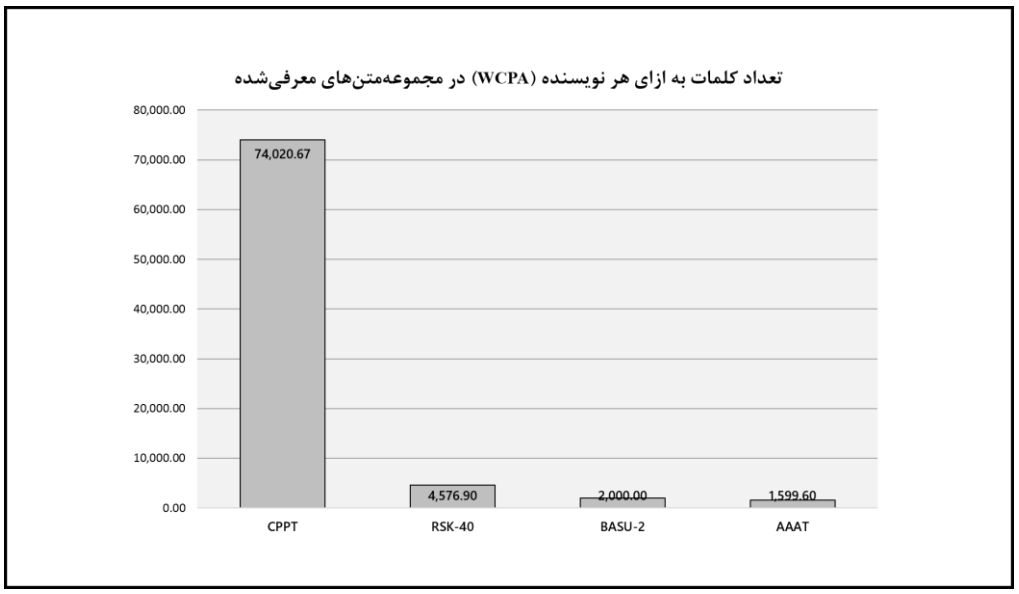
همان‌طور که بیان شد کمیت TCPA بیش‌تر زمانی دارای اهمیت است که بخواهیم از الگوواره‌های مبتنی بر نمونه استفاده نماییم. اما برای الگوواره‌های مبتنی بر پروفایل، کمیت دیگری مناسب به نظر می‌رسد: «تعداد کلمات به ازای هر نویسنده[1]». این کمیت اختصاراً با نماد WCPA نمایش داده شده و به صورت «تعداد کل کلمات موجود در مجموعه‌متن تقسیم بر تعداد کل نویسندگان» تعریف می‌شود. با استفاده از جداول C و D ارائه‌شده برای مجموعه‌متن‌های معرفی‌شده، می‌توان نمودار مقادیر WCPA را برای چهار مجموعه به صورت شکل ۴-۹ رسم نمود.

در الگوواره‌ی مبتنی بر پروفایل، به دنبال استخراج پروفایلی از یک نویسنده هستیم که به بهترین شکل ممکن سبک نوشتاری وی را نمایش دهد. به طور کلی هر چه مقدار WCPA برای یک مجموعه‌متن بالاتر باشد، احتمال استخراج چنین پروفایل‌هایی بالاتر رفته و لذا انتظار دست‌یابی به دقت‌های بالاتر افزایش می‌یابد. به همین جهت با استفاده از شکل ۴-۹ می‌توان نتیجه گرفت که:

- مجموعه‌ی CPPT دارای کمیت WCPA بسیار بالایی می‌باشد و لذا انتظار دقت بالا از آن می‌رود.

[1] Word Count Per Author (WCPA)

- مجموعه‌های BASU-2 و AAAT از نظر کمیت WCPA در وضعیت بسیار پایینی قرار داشته و نمی‌بایست انتظار دقت بالا از آن‌ها داشته باشیم.
- مجموعه‌ی RSK-40 در وضعیتی بین دو حالت فوق قرار داشته و لذا از این منظر، دقتی متوسط از آن انتظار می‌رود.



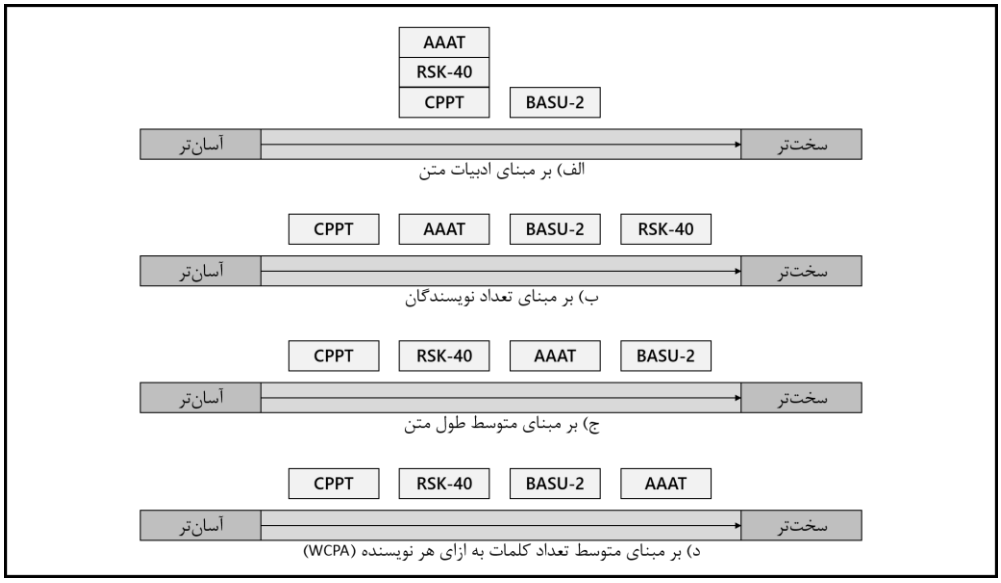
شکل ۵-۴: مقایسه‌ی مجموعه‌متن‌های معرفی‌شده بر مبنای تعداد کلمات به ازای هر نویسنده

۵-۷ جمع‌بندی مقایسه‌ها

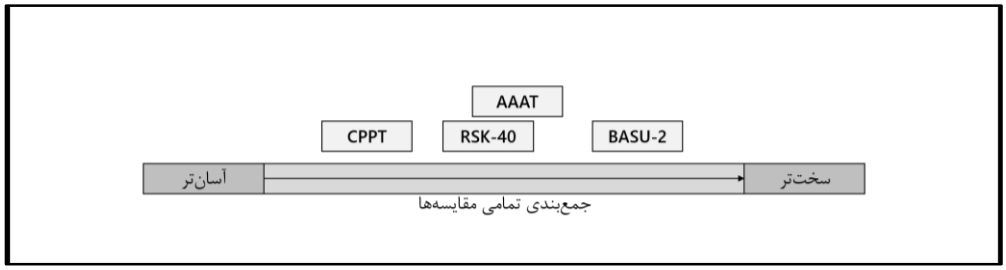
در قسمت‌های ۴-۵-۱ تا ۴-۵-۶ هر بار مجموعه‌متن‌های ارائه‌شده را از یک منظر خاص مورد بررسی و مقایسه قرار دادیم. به منظور رسیدن به یک دید جامع نسبت به این مجموعه‌ها، می‌بایست مقایسه‌های صورت‌پذیرفته در قسمت‌های قبلی را جمع‌بندی نماییم. هدف ما از این جمع‌بندی تعیین سطح دشواری نسبی مجموعه‌متن‌هاست، لذا در این جمع‌بندی فقط نتایج حاصل از مقایسه‌هایی استفاده می‌شود که در تعیین سطح دشواری مؤثر باشند. به همین جهت نتایج مقایسه‌ی رعایت ویژگی‌های مجموعه‌متن ایده‌آل (به دلیل عدم در بر داشتن اطلاعاتی در مورد سطح دشواری) و مقایسه‌ی تعداد متن موجود به ازای هر نویسنده (به دلیل عدم تناسب با روش‌های استفاده‌کننده از الگوواره‌ی مبتنی بر پروفایل) در این جمع‌بندی لحاظ نمی‌شوند. این جمع‌بندی به صورت کِیفی در شکل ۴-۱۰ نمایش داده شده است. با توجه به شکل ۴-۱۰، چهار مجموعه‌متن ارائه‌شده را می‌توان به صورت تقریبی بر اساس دشواری مرتب نمود:

- مجموعه‌ی BASU-2 دشوارترین مجموعه‌ی موجود است، زیرا در تمامی مقایسه‌ها در زمره‌ی دو مجموعه‌ی سخت‌تر قرار دارد.
- مجموعه‌های RSK-40 و AAAT مشترکاً در رده‌ی دوم این مقایسه قرار می‌گیرند. هر چند مجموعه‌ی AAAT کمی دشوارتر به نظر می‌رسد.
- مجموعه‌ی CPPT آسان‌ترین مجموعه‌ی موجود می‌باشد، زیرا در تمامی مقایسه‌ها عنوان ساده‌ترین مجموعه را داراست.

این رتبه‌بندی به صورت کِیفی در شکل ۴-۱۱ نمایش داده شده است.



شکل ۵-۵: مقایسه‌ی کِیفی سطح دشواری مجموعه‌متن‌های معرفی‌شده بر چهار مبنای مختلف



شکل ۵-۶: مقایسه‌ی نهایی سطح دشواری مجموعه‌متن‌های معرفی‌شده به صورت کِیفی

پیوست ۶:

پرسش‌های بنیادین پایان‌نامه و پاسخ به آن‌ها

۶-۱ پرسش‌های بنیادین

از فصل دوم به یاد داریم که تا کنون تمامی پژوهش‌های صورت‌گرفته در حوزه‌ی تخصیص نویسنده‌ی زبان فارسی منحصر به روش‌های مبتنی بر سیستم‌های NLP (پردازش زبان‌های طبیعی) بوده است. اما در این پایان‌نامه روش‌های NDP در دستور کار قرار گرفتند که در واقع روش‌هایی کاملاً مستقل از سیستم‌های NLP می‌باشند. لذا نوآوری اصلی این پایان‌نامه استفاده از روش‌های NDP در حل مسائل تخصیص نویسنده‌ی زبان فارسی می‌باشد. از آن‌جا که هیچ پیشینه‌ای در استفاده از روش‌های NDP در زبان فارسی وجود ندارد، لذا در این پایان‌نامه سعی شد پژوهش‌های صورت‌گرفته به نوعی باشند که بتوان با استفاده از نتایج آن‌ها، یک دید کلی از عملکرد روش‌های NDP در زبان فارسی ارائه داد. بدین منظور دو مورد اساسی در دستور کار قرار گرفت:

۱) در این پایان‌نامه به جای مطالعه‌ی فقط یک روش خاص و تلاش برای ارتقای عملکرد آن، سعی شد گستره‌ی وسیعی از روش‌های NDP (هشت روش بیان‌شده در فصل سوم) مطالعه گردیده و مورد آزمایش قرار گیرند تا بتوان با استفاده از نتایج بدست‌آمده، یک دید کلی در مورد عملکرد این نوع روش‌ها در زبان فارسی ارائه داد.

۲) علاوه بر مطالعه‌ی گستره‌ی وسیعی از روش‌های NDP، تلاش شد مجموعه‌متن‌های مورد آزمایش نیز دارای گستره‌ی تقریباً وسیعی باشند تا نتایج بدست‌آمده دارای جامعیت بیش‌تری باشند. بدین‌منظور از سه مجموعه‌متن دارای زبان فارسی استفاده شد: CPPT، RSK-40 و BASU-2. این مجموعه‌ها گستره‌ی وسیعی از متون فارسی را شامل می‌شوند؛ زیرا:
   - در آن‌ها هم متون کتابی وجود دارد (CPPT و RSK-40) و هم متون محاوره‌ای (BASU-2).
   - علاوه بر متون بلند (CPPT)، شامل متون کوتاه نیز می‌باشند (RSK-40 و BASU-2).

- تعداد نویسندگان مجموعه‌ها نیز دارای گستره‌ی وسیعی است: ۶ نویسنده در CPPT، ۲۰ نویسنده در BASU-2 و ۴۰ نویسنده در RSK-40.

دید کلی حاصل از آزمایش‌های این فصل می‌تواند مبنای کار پژوهش‌های آینده در حوزه‌ی تخصیص نویسنده‌ی زبان فارسی قرار گرفته و اطلاعات مفیدی را به پژوهش‌گران این حوزه ارائه دهد.

در این پایان‌نامه، علاوه بر نوآوری اصلی (استفاده از روش‌های NDP در حل مسائل تخصیص نویسنده)، نوآوری دیگری نیز صورت پذیرفته و دو روش جدید برای حل مسائل تخصیص نویسنده پیشنهاد شده است. لذا موضوع دیگری که بیان آن در این فصل ضروری می‌باشد، بررسی عملکرد دو روش پیشنهادی و مقایسه‌ی آن‌ها با هشت روش مطالعه شده می‌باشد.

هم‌چنین با توجه به اشتراک زیاد و شباهت بالای الفبای زبان‌های فارسی و عربی، در این پایان‌نامه یک مجموعه‌متن دارای زبان عربی نیز مورد آزمایش قرار گرفته است تا بتوان عملکرد روش‌های NDP را در زبان عربی نیز ارزیابی نمود. البته این ارزیابی دارای جامعیت نبوده و نمی‌توان آن را به کل متون زبان عربی تعمیم داد، زیرا فقط از بررسی یک مجموعه‌متن حاصل شده است.

با توجه به موارد فوق، اکنون پرسش‌های بنیادینی را ارائه می‌نماییم که می‌بایست در این فصل پاسخ داده شوند:

الف) عملکرد روش‌های NDP در زبان‌های فارسی و عربی

پرسش ۱: آیا روش‌های NDP قادر به حل مسائل تخصیص نویسنده‌ی زبان فارسی می‌باشند؟

پرسش ۲: آیا روش‌های NDP قادر به حل مسائل تخصیص نویسنده‌ی زبان عربی می‌باشند؟

ب) عملکرد روش‌های پیشنهادی

پرسش ۳: آیا استفاده از اندیس انگرام‌ها (به جای فرکانس انگرام‌ها) می‌تواند در حل مسائل تخصیص نویسنده مفید باشد؟ به عبارت دیگر آیا روش CNG-WIS قادر است مسائل تخصیص نویسنده را با دقت معقولی حل نماید؟ به منظور پاسخ دادن به این سؤال می‌بایست دقت این روش را به صورت جداگانه (مستقل از دقت سایر روش‌ها) موردبررسی قرار دهیم.

پرسش ۴: عملکرد نسبی روش CNG-WIS نسبت به سایر روش‌ها چگونه است؟ ارائه‌ی پاسخ این سؤال مستلزم مقایسه‌ی دقت این روش با سایر روش‌هاست.

پرسش ۵: آیا استفاده از انگرام‌های پراکنده (به جای انگرام‌های متداول) می‌تواند در حل مسائل تخصیص نویسنده مفید باشد؟ به عبارت دیگر آیا روش VNG قادر است مسائل تخصیص نویسنده را با دقت معقولی

حل نماید؟ به منظور پاسخ دادن به این سؤال می‌بایست دقت این روش را به صورت جداگانه موردبررسی قرار دهیم.

پرسش ۶: عملکرد نسبی روش VNG نسبت به سایر روش‌ها چگونه است؟ ارائه‌ی پاسخ این سؤال مستلزم مقایسه‌ی دقت این روش با سایر روش‌هاست.

۶-۲ پاسخ به پرسش‌های بنیادین

در این بخش، به پرسش‌های بنیادینی که در ابتدای فصل مطرح شد، پاسخ می‌دهیم:

الف) عملکرد روش‌های NDP در زبان‌های فارسی و عربی

پرسش ۱: آیا روش‌های NDP قادر به حل مسائل تخصیص نویسنده‌ی زبان فارسی می‌باشند؟

بلی، نتایج بدست‌آمده در این فصل حاکی از قدرت بالای روش‌های NDP در حل چنین مسائلی می‌باشد. نکته‌ی جالب آنکه در دو آزمایش اول، دقتی معادل ۱۰۰ درصد حاصل شده است. هم‌چنین در آزمایش سوم (که امکان مقایسه‌ی روش‌های NDP و روش‌های مبتنی بر سیستم‌های NLP وجود داشت)، دیدیم که روش‌های NDP عملکرد بهتری نسبت به روش‌های مبتنی بر سیستم‌های NLP از خود ارائه می‌دهند.

پرسش ۲: آیا روش‌های NDP قادر به حل مسائل تخصیص نویسنده‌ی زبان عربی می‌باشند؟

با توجه به اینکه در این فصل فقط یک مجموعه‌متن عربی مورد آزمایش قرار گرفت، لذا نمی‌توان یک حکم کلی ارائه داد؛ اما مقایسه‌ی دقت‌های بدست‌آمده در آزمایش چهارم با دقت‌های ارائه‌شده توسط گردآورندگان مجموعه، نشان می‌دهد که روش‌های NDP بر روی زبان عربی نیز عملکرد بهتری نسبت به روش‌های مبتنی بر سیستم‌های NLP دارند.

ب) عملکرد روش‌های پیشنهادی

پرسش ۳: آیا استفاده از اندیس انگرام‌ها (به جای فرکانس انگرام‌ها) می‌تواند در حل مسائل تخصیص نویسنده مفید باشد؟ به عبارت دیگر آیا روش CNG-WIS قادر است مسائل تخصیص نویسنده را با دقت معقولی حل نماید؟

پاسخ این سؤال مثبت است، زیرا این روش در دو آزمایش اول دقت‌های بسیار بالایی را از خود ارائه داده است: در آزمایش اول: 97.93% و در آزمایش دوم: 100%

پس با استفاده از نتایج این آزمایش‌ها، اصل ایده‌ی «مفید بودن اندیس انگرام‌ها در حل مسائل تخصیص نویسنده» تأیید می‌شود.

پرسش ۴: عملکرد نسبی روش CNG-WIS نسبت به سایر روش‌ها چگونه است؟

عملکرد نسبی این روش نسبت به هشت روش مطالعه‌شده در مسائل زبان فارسی رضایت‌بخش می‌باشد، زیرا روش CNG-WIS در تمامی آزمایش‌ها رتبه‌ی معقولی را کسب نموده است:

- در آزمایش اول در رتبه‌ی ششم قرار گرفته است: متوسط
- در آزمایش دوم در رتبه‌ی دوم قرار گرفته است: خوب
- در آزمایش سوم در رتبه‌ی چهارم قرار گرفته است: متوسط

اما این روش در مجموعه‌متن دارای زبان عربی رتبه‌ی هشتم را به خود اختصاص داده و لذا عملکرد نسبی آن در زبان عربی نسبت به عملکردش در زبان فارسی ضعیف‌تر می‌باشد.

پرسش ۵: آیا استفاده از انگرام‌های پراکنده (به جای انگرام‌های متداول) می‌تواند در حل مسائل تخصیص نویسنده مفید باشد؟ به عبارت دیگر آیا روش VNG قادر است مسائل تخصیص نویسنده را با دقت معقولی حل نماید؟

پاسخ این سؤال نیز مثبت است، زیرا این روش در آزمایش‌های اول و دوم دقت‌های بالایی را از خود ارائه داده است: در آزمایش اول: 97.93% و در آزمایش دوم: 99.13%

پس با استفاده از نتایج این آزمایش‌ها، اصل ایده‌ی «مفید بودن انگرام‌های پراکنده در حل مسائل تخصیص نویسنده» تأیید می‌شود.

پرسش ۶: عملکرد نسبی روش VNG نسبت به سایر روش‌ها چگونه است؟

عملکرد نسبی این روش نسبت به هشت روش مطالعه‌شده در مسائل زبان فارسی چندان رضایت‌بخش نمی‌باشد، زیرا روش VNG در هیچ کدام از سه آزمایش اول نتوانسته است در رده‌های ابتدایی جدول قرار گیرد:

- در آزمایش اول در رتبه‌ی هفتم قرار گرفته است: ضعیف
- در آزمایش دوم در رتبه‌ی ششم قرار گرفته است: متوسط
- در آزمایش سوم در رتبه‌ی هفتم قرار گرفته است: ضعیف

اما این روش در زبان عربی بسیار عالی عمل نموده و دارای بهترین عملکرد نسبت به ۹ روش دیگر می‌باشد. کشف دلیل تفاوت عملکرد این روش در زبان‌های فارسی و عربی، نیاز به اطلاعات مختصصان حوزه‌ی زبان‌شناسی دارد که آن را به پژوهش‌های آتی واگذار می‌نماییم.

پیوست ۷:

نحوه‌ی ارائه‌ی یک آزمایش تخصیص نویسنده

در این بخش می‌خواهیم روند استانداردی برای ارائه‌ی یک آزمایش تعریف نماییم، که در ادامه‌ی فصل هنگام ارائه‌ی هر چهار آزمایش مورداستفاده قرار گیرد. به طور کلی هنگام معرفی یک آزمایش می‌بایست به سه سؤال اساسی پاسخ داده شود:

- آزمایشِ انجام‌گرفته دارای چه مشخصاتی است؟
- نتایج حاصل از آزمایش چه بوده است؟
- چگونه می‌توان نتایج بدست‌آمده را تحلیل نمود؟

در ادامه، روند استانداردی برای پاسخ به هر کدام از سه سؤال فوق بیان می‌شود: پاسخ سؤال اول در قسمت ۵-۲-۱، پاسخ دومین سؤال در ۵-۲-۲ و در نهایت پاسخ آخرین سؤال در ۵-۲-۳ ارائه می‌شود.

۷-۱ مشخصات یک آزمایش

در این قسمت می‌خواهیم به این سؤال پاسخ دهیم که «هنگام معرفی یک آزمایش، می‌بایست چه مشخصه‌های از آن بیان شوند؟» باید توجه داشت که هدف از بیان این مشخصه‌ها در واقع افزایش درک خواننده از آزمایشِ انجام‌شده است، لذا ارائه‌ی مشخصه‌هایی که درک چندانی به خواننده القا نمی‌کنند، ضروری به نظر نمی‌رسد. به عنوان مثال (در آزمایش‌های حوزه‌ی تخصیص نویسنده) ذکر مواردی همچون زبان برنامه‌نویسی مورداستفاده و مشخصات فنی رایانه‌ای که پردازش‌ها بر روی آن صورت گرفته است، کمک چندانی به درک بهتر خواننده از آزمایش صورت‌گرفته نمی‌کند؛ لذا از بیان چنین مواردی صرف نظر می‌کنیم. البته در پیوست توضیحاتی در مورد نرم‌افزار تولیدشده جهت انجام آزمایش‌های این پایان‌نامه ارائه شده است.

به طور کلی هنگام ارائه‌ی یک آزمایش بهترست مشخصه‌های زیر بیان شوند، تا خواننده درک بهتری از آن آزمایش بدست آورد:

مشخصه‌ی اول: مجموعه‌متن

اولین مشخصه‌ای که می‌بایست برای هر آزمایش بیان شود، مجموعه‌متن مورداستفاده در آن آزمایش است. همان‌طور که در مقدمه‌ی فصل گفته شد، در این پایان‌نامه از چهار مجموعه‌متن مختلف استفاده شده است: CPPT، RSK-40، BASU-2 و AAAT. پس در هر آزمایش می‌بایست تعیین شود که کدام یک از این چهار مجموعه‌متن مورداستفاده قرار گرفته است. هرچند این مجموعه‌متن‌ها در فصل چهارم به طور کامل معرفی شده و ویژگی‌های هر یک از آن‌ها به طور جداگانه موردبررسی قرار گرفته است، اما در این فصل نیز هنگام بیان اولین مشخصه‌ی هر آزمایش (یعنی مجموعه‌متن)، اطلاعاتی از قبیل تعداد نویسنده‌ها و تعداد متون موجود در مجموعه‌متن را ارائه می‌دهیم.

مشخصه‌ی دوم: نحوه‌ی تشکیل مجموعه‌های آموزش و آزمایش

از فصل سوم به یاد داریم که یک سیستم تخصیص نویسنده، متن دیده‌نشده را (که نویسنده‌ی آن برای سیستم نامعلوم است) بر مبنای متون موجود در مجموعه‌متن (که نویسنده‌ی آن‌ها برای سیستم معلوم می‌باشند)، به یکی از نویسندگان حاضر در مجموعه تخصیص می‌دهد. البته ممکن است به جای یک متن، چندین متن دیده‌نشده جهت تخصیص نویسنده به سیستم اعمال شوند، که در این صورت سیستم هر یک از متون دیده‌نشده را به صورت جدا پردازش نموده و هر کدام را به یک نویسنده تخصیص می‌دهد.

در حوزه‌ی تخصیص نویسنده، گاهی متون موجود در مجموعه‌متن را (که عملیات تخصیص بر مبنای آن‌ها صورت می‌پذیرد) اصطلاحاً «متون آموزشی[1]» یا «مجموعه‌ی آموزش[2]» و متون دیده‌نشده را اصطلاحاً «متون آزمایشی[3]» یا «مجموعه‌ی آزمایش[4]» می‌نامند. لازم بذکرست که معمولاً در یک مجموعه‌متن، مجموعه‌های آموزش و آزمایش از ابتدا معلوم نبوده و پیش از شروع فرایند آزمایش تشکیل می‌شوند. روند کلی تشکیل این دو مجموعه بدین صورت است:

- گام اول: تعدادی از متون موجود در مجموعه‌متن اصلی را به عنوان متون دیده‌نشده انتخاب نموده و آن‌ها را در مجموعه‌ی آزمایش قرار می‌دهیم.

[1] Training Texts
[2] Train Set or Train Corpora
[3] Testing Texts
[4] Test Set or Test Corpora

- گام دوم: متون انتخاب شده در گام اول را از مجموعه‌متن اصلی حذف می‌نماییم تا مجموعه‌ی آموزش حاصل شود.

سؤالی که در این‌جا مطرح می‌شود آنست که متون دیده‌نشده را می‌بایست بر چه مبنایی انتخاب نماییم؟ تشریح روند انتخاب متون دیده‌نشده (و یا به طور کلی روند تشکیل مجموعه‌های آموزش و آزمایش)، مشخصه‌ی دومی است که بیان آن هنگام ارائه‌ی هر آزمایش ضروری می‌باشد.

مشخصه‌ی سوم: روش مورداستفاده جهت تخصیص نویسنده

سومین مشخصه‌ی لازم هنگام ارائه‌ی یک آزمایش، روش مورداستفاده جهت تخصیص نویسنده می‌باشد. یعنی می‌بایست تعیین نماییم که در اجرای آزمایش از کدام یک از روش‌های NDP استفاده شده است. همان‌طور که در مقدمه بیان شد، در هر کدام از چهار آزمایش ارائه‌شده در این فصل، تمامی ۱۰ روش مطالعه‌شده در فصول قبلی مورداستفاده قرار گرفته‌اند، تا امکان مقایسه‌ی عملکرد آن‌ها فراهم گردد. در نتیجه، مشخصه‌ی سوم در تمامی آزمایش‌های این فصل یکسان بوده و لذا نیازی به تکرار مجدد آن نمی‌باشد و هنگام ارائه‌ی یک آزمایش سخنی از آن به میان نمی‌آوریم.

مشخصه‌ی چهارم: مجموعه‌ی پارامترهای مورداستفاده

از فصول گذشته به یاد داریم که هر روش NDP دارای دو پارامتر اساسی است: N (طول انگرام) و L (طول پروفایل). بدیهی است که مقدار این دو پارامتر در دقت روش تأثیرگذار خواهند بود. مقادیری از N و L را که بالاترین دقت به ازای آن‌ها بدست می‌آید، «پارامترهای بهینه» و یا «جفت بهینه» می‌نامیم. البته ممکن است بالاترین دقت به ازای چند جفت از پارامترها رخ دهد، که در این صورت روش موردنظر به جای یک جفت بهینه، چند جفت بهینه خواهد داشت؛ لذا بهترست آن‌ها را به صورت یک مجموعه نمایش دهیم: «مجموعه‌ی جفت‌های بهینه[1]». در این مجموعه ممکن است یک یا چند جفت قرار گیرند.

از آن‌جا که معیار خاصی برای تعیین مجموعه‌ی جفت‌های بهینه پیش از شروع آزمایش وجود ندارد، لذا معمولاً در آزمایش‌های تخصیص نویسنده از فرایند «جستجوی مُشبک[2]» استفاده می‌شود. در این فرایند به جای استفاده از یک مقدار خاص برای N و L، مجموعه‌ای از این پارامترها مورداستفاده قرار گرفته و آزمایش به ازای هر جفت از این پارامترها یک‌بار اجرا می‌شود. به عبارت دیگر روند کلی جستجوی مشبک بدین‌صورت است:

[1] Optimal-Pairs Set
[2] Grid-Search

۱) ابتدا برای هر کدام پارامترهای N و L، مجموعه‌ای از مقادیر مختلف انتخاب می‌نماییم. به عنوان مثال:

$$N = \{3, 4\}$$
$$L = \{100, 200, 300\}$$

۲) سپس هر جفت از پارامترهای موجود در این دو مجموعه را یک بار می‌آزماییم. یعنی ابتدا آزمایش را به ازای جفت ($N = 3, L = 100$) اجرا نموده، بعد از آن به ازای جفت ($N = 3, L = 200$) انجام داده و این روند را تا آن‌جا ادامه می‌دهیم که تمامی جفت‌های ممکن (در این مثال ۶ جفت)، هر کدام یک‌بار آزموده شوند.

۳) در نهایت با مقایسه‌ی دقت‌های بدست‌آمده به ازای هر جفت، اعضای مجموعه‌ی جفت‌های بهینه آشکار می‌شوند. بدین صورت که ابتدا مقدار بیشینه‌ی دقت را تعیین نموده ($Acc_{\max}$) و سپس هر جفتی از N و L را که منجر به تولید $Acc_{\max}$ شده باشد، درون مجموعه‌ی جفت‌های بهینه قرار می‌دهیم.

در تمامی آزمایش‌های این فصل، از مجموعه‌های زیر به عنوان مقادیر N و L استفاده شده است:

$$N = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$
$$L = \{50, 100, 200, 300, ..., 1000, 1100, ..., 2000, 2100, ..., 2900, 3000\} \quad (۱-۷)$$

این مجموعه‌ها گستره‌ی وسیعی از مقادیر N و L را شامل می‌شوند، بدین صورت که شامل ۱۰ مقدار مختلف برای N و ۳۱ مقدار مختلف برای L می‌باشند. لذا در مجموع، در هر آزمایش این فصل، هر یک از ده روش موجود به ازای ۳۱۰ جفت پارامتر مختلف آزموده می‌شوند. در نتیجه مشخصه‌ی چهارم نیز (همانند مشخصه‌ی سوم) در تمامی آزمایش‌ها یکسان بوده و نیازی به تکرار آن هنگام ارائه‌ی هر آزمایش نمی‌باشد.

جمع‌بندی

در این بخش چهار مشخصه برای هر آزمایش ارائه گردید که از میان آن‌ها، دو مشخصه برای تمامی آزمایش‌های این فصل یکسان می‌باشند: مشخصه‌ی سوم (روش مورداستفاده جهت تخصیص نویسنده) و مشخصه‌ی چهارم (مجموعه‌ی پارامترهای مورداستفاده). در نتیجه بیان مجدد این دو مشخصه در هر آزمایش ضرورتی نداشته و لذا در ادامه‌ی فصل هنگام معرفی یک آزمایش، فقط دو مشخصه‌ی اول آن بیان خواهد شد: مشخصه‌ی اول (مجموعه‌متن) و مشخصه‌ی دوم (نحوه‌ی تشکیل مجموعه‌های آموزش و آزمایش).

۷-۲ ارائه‌ی نتایج یک آزمایش

در این قسمت می‌خواهیم با پاسخ به این سؤال که «نتایج یک آزمایش می‌بایست چگونه بیان شوند؟»، روندی استاندارد برای ارائه‌ی نتایج یک آزمایش پیشنهاد دهیم. از بخش قبل به یاد داریم که در هر آزمایش، هر کدام از روش‌های NDP به ازای ۳۱۰ جفت پارامتر مختلف آزموده می‌شوند (جستجوی مشبک). لذا نتایج عملکرد هر روش را می‌توان به صورت جدولی همانند شکل ۵-۱ نمایش داد. چنین جدولی را «جدول جستجوی مشبک[1]» می‌نامند.

3000	...	1500	...	50	L/N
Acc(1,3000)	...	Acc(1,1500)	...	Acc(1,50)	**1**
Acc(2,3000)	...	Acc(2,1500)	...	Acc(2,50)	**2**
Acc(3,3000)	...	Acc(3,1500)	...	Acc(3,50)	**3**
Acc(4,3000)	...	Acc(4,1500)	...	Acc(4,50)	**4**
Acc(5,3000)	...	Acc(5,1500)	...	Acc(5,50)	**5**
Acc(6,3000)	...	Acc(6,1500)	...	Acc(6,50)	**6**
Acc(7,3000)	...	Acc(7,1500)	...	Acc(7,50)	**7**
Acc(8,3000)	...	Acc(8,1500)	...	Acc(8,50)	**8**
Acc(9,3000)	...	Acc(9,1500)	...	Acc(9,50)	**9**
Acc(10,3000)	...	Acc(10,1500)	...	Acc(10,50)	**10**

شکل ۷-۱: نمای کلی از جدول جستجوی مشبک برای آزمایش‌های این فصل

لازم بذکرست که دقت یک روش به ازای هر جفت از N و L به صورت زیر محاسبه می‌شود:

$$Acc(N,L) = 100 \times \frac{C_{true}(N,L)}{C_{unseen}} \qquad (۲-۷)$$

که در آن:

- C_{unseen} برابراست با تعداد کل متون دیده‌نشده

[1] Grid-Search Table (GST)

- $C_{true}(N,L)$ برابرست با تعداد تعداد متون دیده‌نشده‌ای که سیستم طراحی‌شده با پارامترهای N و L، نویسنده‌ی آن‌ها را به درستی تخصیص داده است.

هم‌چنین، مجموعه‌ی جفت‌های بهینه را نیز می‌توان به صورت زیر بیان نمود:

$$\left\{(N,L)_{optimal}\right\}=\left\{(N,L)\mid Acc(N,L)=Acc_{\max}\right\} \qquad (۱-۸)$$

از آنجا که در هر کدام از آزمایش‌های این فصل، ۱۰ روش NDP بر روی یک مجموعه‌متن آزمایش می‌شوند، لذا برای هر آزمایش ۱۰ جدول جستجوی مشبک خواهیم داشت. با توجه به تعداد بالای این جداول، ارائه‌ی آن‌ها در متن پایان‌نامه موجب سردرگمی خواننده خواهد شد؛ پس بهترست فقط نتایج نهایی حاصل از جستجوی مشبک را بیان نماییم، یعنی مقدار بیشینه‌ی دقت ($Acc_{\max}$) و مجموعه‌ی جفت‌های بهینه ($\left\{(N,L)_{optimal}\right\}$)

به عبارت دیگر نتایج هر کدام از آزمایش‌های این فصل را می‌توان در قالب جدولی دارای ۱۰ سطر و ۳ ستون نمایش داد. هر سطر از جدول متناظر با یک روش بوده و ستون‌های مذکور بدین شرح می‌باشند: ستون اول: نام روش، ستون دوم: بالاترین دقت بدست‌آمده در روش (بر حسب درصد) و ستون سوم: مجموعه‌ی جفت‌های بهینه‌ی مربوط به روش. یک نمای کلی از این جدول (که در ادامه از آن با عنوان «جدول نتایج» یاد خواهیم کرد)، در شکل ۵-۲ نمایش داده شده است.

نام روش	بالاترین دقت	پارامترهای بهینه
CNG		
CNG-SPI		
CNG-D1		
CNG-D2		
CNG-WPI		
RLP		
RLP-IAF		
O-SPI		
CNG-WIS		
VNG		

شکل ۷-۲: نمای کلی از جدول نتایج مربوط به هر آزمایش

۳-۷ تحلیل نتایج یک آزمایش

همانطور که در بخش قبل بیان شد، نتایج حاصل از هر آزمایش را می‌توان به صورت جدولی همانند شکل ۵-۲ (جدول نتایج) نمایش داد. در این قسمت می‌خواهیم بررسی کنیم که چگونه می‌توان نتایج موجود در این جدول را تحلیل نمود؟ هدف از این بررسی، ارائه‌ی یک روند استاندارد جهت تحلیل تمامی آزمایش‌های این فصل می‌باشد. سه مورد از مهم‌ترین تحلیل‌های جدول نتایج به شرح زیر می‌باشند:

تحلیل اول: رتبه‌بندی روش‌ها

اولین و مهم‌ترین تحلیلی که می‌توان از جدول نتایج ارائه داد، تعیین رتبه‌ی هر کدام از ۱۰ روش استفاده‌شده می‌باشد. در این قسمت دو نوع رتبه‌بندی پیشنهاد می‌شود:

۱) رتبه‌بندی نوع اول: بر مبنای بالاترین دقت

در این نوع رتبه‌بندی می‌بایست جدول نتایج را بر مبنای بالاترین دقت عملکرد هر روش (ستون دوم) مرتب نموده و به هر روش یک رتبه اختصاص دهیم، بدین‌صورت که روش یا روش‌های دارای بالاترین دقت در رتبه‌ی اول جای گرفته، روش یا روش‌های دارای دومین مقدار بیشینه در رتبه‌ی دوم قرار داده شده و این روند به ازای تمامی روش‌ها ادامه می‌دهیم. در این حالت ممکن است دو روش دارای رتبه‌ی یکسانی باشند.

۲) رتبه‌بندی نوع دوم: بر مبنای بالاترین دقت و تعداد جفت‌های بهینه

این نوع رتبه‌بندی همانند نوع اول است، با این تفاوت که اگر دقت دو روش با هم برابر شود، روشی را که تعداد اعضای مجموعه‌ی جفت‌های بهینه‌ی متناظر با آن بیش‌تر باشد، در رتبه‌ی بهتری قرار می‌دهیم.

به عنوان مثال فرض کنیم نتایج حاصل از جستجوی مشبک برای روش‌های A و B روی یک مجموعه‌متن فرضی به صورت زیر باشد:

$$A: \quad Acc_{\max} = 95\% \ , \ \left\{(\mathrm{N},\mathrm{L})_{optimal}\right\} = \left\{(3,500),(3,700),(4,600)\right\}$$

$$B: \quad Acc_{\max} = 95\% \ , \ \left\{(\mathrm{N},\mathrm{L})_{optimal}\right\} = \left\{(4,400),(4,700),(5,500),(5,700),(6,800)\right\}$$

با این فرض در رتبه‌بندی نوع اول به روش‌های A و B رتبه‌ی یکسانی اختصاص خواهد یافت، اما در رتبه‌بندی نوع دوم، روش B در مکان بهتری قرار خواهد گرفت. دلیل این برتری آنست که تعداد جفت‌های بهینه‌ی روش B بیش‌تر بوده و لذا احتمال دست‌یابی به دقت بیشینه در آن بیش‌تر از روش A است.

تحلیل دوم: بررسی عملکرد روش‌های پیشنهادی

دومین تحلیلی که می‌توان از جدول نتایج ارائه داد، بررسی و ارزیابی عملکرد روش‌های پیشنهادی (روش CNG-WIS و روش VNG) می‌باشد.

تحلیل سوم: بررسی جفت‌های بهینه

سومین تحلیلی که می‌تواند بر روی جدول نتایج صورت پذیرد، بررسی مقادیر جفت‌های بهینه در روش‌های مختلف می‌باشد. معمولاً پارامتر N در هر زبان دارای یک بازه‌ی معمول است که اکثر روش‌ها به ازای مقدار N موجود در آن بازه، بهترین عملکرد خود را نشان می‌دهند. به عنوان مثال در زبان انگلیسی بازه‌ی معمول برای N بین ۳ تا ۵ می‌باشد [۳۱]، یعنی:

$$\text{Normal-Range in English:} \quad 3 \le N \le 5 \tag{۲-۸}$$

و در زبان یونانی داریم [۳۱]:

$$\text{Normal-Range in Greek:} \quad 5 \le N \le 7 \tag{۳-۸}$$

سؤالی که شاید در ذهن مخاطب ایجاد شده باشد آنست که «آیا در یک زبان خاص، می‌توان برای پارامتر L نیز یک بازه‌ی معمول را معرفی نمود؟» پاسخ این سؤال مثبت است، اما معمولاً بازه‌ی معرفی‌شده برای L، بر خلاف بازه‌ی معمول N شامل گستره‌ی وسیعی از مقادیر بوده و لذا چندان به کار نمی‌آید. به عنوان مثال در زبان انگلیسی داریم [۳۱]:

$$\text{Normal-Range in English:} \quad 1,000 \le L \le 5,000 \tag{۴-۸}$$

متأسفانه تا کنون هیچ پیشنهادی برای بازه‌ی معمول پارامترهای N و L در زبان فارسی ارائه نشده است. لذا یکی از تحلیل‌های مفیدی که می‌توان بر روی جدول نتایج صورت داد، بررسی مقادیر عددی جفت‌های بهینه‌ی هر روش می‌باشد. هدف از این بررسی، در واقع تلاش برای تعیین یک بازه‌ی معمول برای پارامترهای N و L می‌باشد. با بررسی این پارامترها بر روی نتایج سه مجموعه‌متن دارای زبان فارسی (CPPT، RSK-40 و BASU-2) می‌توان گستره‌ای را به عنوان بازه‌ی معمول N و L پیشنهاد داد؛ اما با توجه به اینکه برای زبان عربی فقط یک مجموعه‌متن در اختیار داریم، ارائه‌ی بازه‌ی معمول برای زبان عربی بر مبنای آن یک مجموعه‌متن چندان معقول به نظر نمی‌رسد.

پیوست ۸:

توضیحات تکمیلی در زمینه‌ی مطالعات موردی

۸-۱ نظیره‌های گلستان

«نظیره‌نویسی» یکی از شاخه‌های ادبیات فارسی است که می‌توان آن را به عنوان چالشی برای حوزه‌ی تخصیص نویسنده در نظر گرفت. «نظیره» به اثری گفته می‌شود که به تقلید از یک اثر مشهور و مقبول، پدید آمده باشد. با این توصیف، اگر بتوانیم مجموعه‌متنی شامل یک اثر و نظیره‌های آن گردآوری نماییم، آن‌گاه می‌توان قدرت روش‌های NDP را در حل یک مسئله‌ی دشوار ادبی نیز آزمود.

یکی از معروف‌ترین نمونه‌های نظیره‌نویسی، نظیره‌هایی‌ست که بر کتاب «گلستان» نوشته شده است. این کتاب که اثر ماندگاری از سعدی شیرازی می‌باشد، «همواره رایج‌ترین کتاب نثر فارسی در سراسر جهان بوده و بیش از هر کتاب دیگری از آن نسخه برداشته و آن را به زبان‌های دیگر ترجمه کرده‌اند. [۹۵]» در مورد نظیره‌های گلستان و چرایی پیدایش آن‌ها، تا کنون اظهارنظرهای متفاوتی ارائه شده است؛ اما به نظر می‌رسد در این زمینه اشارات آقای حمید یزدان‌پرست جامع و راهگشا باشد: «زیبایی بی‌نظیر گلستان و شوق و شوری که در همگان (از پیر و جوان و خاص و عام) برانگیخت و آن را به عنوان گل سرسبد ادبیات پارسی بر تارک متون نظم و نثر نشاند و به مکتب‌خانه‌ها رفت و متن درس فارسی‌خوانان شد، بسیاری را بر آن داشت که بخت‌آزمایی و هنرنمایی کنند و در میدانی که سعدی توسن سخن را به جولان درآورده، خودی نشان دهند و آوازه‌ای کسب کنند. [۹۶]» برخی از پژوهش‌گران تعداد نظیره‌های گلستان را ۴۵ اثر عنوان کرده‌اند [۹۷]. از جمله‌ی آن‌ها می‌توان به این آثار اشاره نمود: نگارستان، خارستان، بهارستان، بلبلستان، شکرستان، سُنبلستان، پریشان، گلستان (اثر محمد شریف، معاصر ناصرالدین‌شاه)، نمکدان، نخلستان، مُلستان، هزاردستان.

کتاب گلستان شامل حکایاتی در هشت باب است که قالب ظاهری هر حکایت متشکل از چند خط نثر به همراه چند بیت شعر می‌باشد. مقلدان سعدی سعی کرده‌اند هم از نظر محتوی و هم از نظر شیوه‌ی نگارش کتابی مانند گلستان بنویسند.

۸-۲ غزلیات سبک هندی

در فصل پنجم دیدیم که روش‌های NDP بر روی متون نثر فارسی عملکرد بسیار مناسبی دارند. نتایج بدست‌آمده در بخش قبلی نیز بیان‌گر قدرت بالای این روش‌ها در حکایات فارسی (ترکیبی از نظم و نثر) می‌باشد. حال می‌خواهیم به سراغ متون نظم رفته و بررسی کنیم که آیا روش‌های NDP توانایی حل مسائل تخصیص نویسنده بر روی اشعار را نیز دارا هستند یا خیر؟

مهم‌ترین پژوهشی که تا کنون بر روی تخصیص نویسنده‌ی اشعار فارسی صورت گرفته توسط شاهمیری و همکاران در سال ۱۳۸۶ انجام شده است [۴۸]. همانطور که در فصل دوم نیز بیان شد، آن‌ها مجموعه‌متنی متشکل از اشعار چهار شاعر بزرگ ایرانی (فردوسی، خیام، حافظ و مولوی) گردآوری نموده و با استفاده از روش‌های مبتنی بر سیستم‌های NLP (پردازش زبان‌های طبیعی)، به دقتی معادل 95.9% دست یافته‌اند. به نظر نگارنده‌ی این پایان‌نامه نقدی بر پژوهش شاهمیری و همکاران وارد است: شعرای انتخاب شده دارای سبک‌های کاملاً متفاوتی بوده و لذا تشخیص اشعار آن‌ها چندان دشوار به نظر نمی‌رسد. به منظور رفع این نقیصه، تصمیم بر آن شد که مجموعه‌متنی شامل اشعار هم‌سبک گردآوری شده و سپس روش‌های NDP بر روی آن آزموده شوند. بدین‌منظور می‌بایست ابتدا سبک‌های شعر فارسی را بشناسیم:

ملک‌الشعرای بهار (در کتاب سبک‌شناسی) برای شعر فارسی شش نوع سبک و دوره قائل شده است [۹۸]:

۱) سبک خراسانی یا ترکستانی (آغاز شعر فارسی تا قرن ۶)
۲) سبک عراقی (از قرن ۶ تا قرن ۱۰)
۳) سبک هندی (از قرن ۱۰ تا قرن ۱۳)
۴) دوره‌ی بازگشت (بازگشت به سبک‌های گذشته) (تمام طول قرن ۱۳)
۵) دوره‌ی مشروطه
۶) دوره‌ی معاصر

از بین این سبک‌ها و دوره‌ها، سبک هندی را (به جهت شباهت زیاد اشعار آن) برای گردآوری مجموعه‌متن انتخاب نمودیم. از شعرای معروف سبک هندی می‌توان به بزرگانی چون صائب تبریزی، بیدل دهلوی، محتشم کاشانی، کلیم کاشانی، حزین لاهیجی، عرفی شیرازی، غالب دهلوی و طالب آملی اشاره کرد.

منابع

[1] K. Dave, *Study of feature selection algorithms for text-categorization,* University of Nevada, Las Vegas: Master of science thesis, December 2011.

[2] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary, "Twitter Trending Topic Classification," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, December 2011.

[3] R. Liu, M. Jiang and Z. Tie, "Automatic Genre Classification by Using Co-training," in *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 14-16 Aug. 2009.

[4] Y.-B. Lee and S. H. Myaeng, "Automatic identification of text genres and their roles in subject-based categorization," in *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, 5-8 Jan. 2004.

[5] E. Stamatatos, G. Kokkinakis and N. Fakotakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics, 26(4):471[495].*, vol. 26, no. 4, pp. 471-495, Dec. 2000.

[6] A. Kennedy and D. Inkpen, "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters," *Computational Intelligence,* vol. 22, no. 2, pp. 110-125, 2006.

[7] X. Zhang, S. Wang, M. Xu and Y. Yin, "Chinese text sentiment classification based on granule network," in *IEEE International Conference on Granular Computing*, Nanchang, 17-19 Aug. 2009.

[8] H. Xia, M. Tao and Y. Wang, "Sentiment text classification of customers reviews on the Web based on SVM," in *Sixth International Conference on Natural Computation*, Yantai, China, 10-12 Aug. 2010.

[9] V. Siva RamaKrishna Reddy, D. Somayajulu and A. Dani, "Sentiment Classification of text reviews using novel feature selection with reduced over-fitting," in *International Conference for Internet Technology and Secured Transactions (ICITST)*, London, 8-11 Nov. 2010.

[10] J. N. Khasnabish, M. Sodhi, J. Deshmukh and G. Srinivasaraghavan, "Detecting Programming Language from Source Code Using Bayesian Learning Techniques," *Machine Learning and Data Mining in Pattern Recognition Lecture Notes in Computer Science*, vol. 8556, pp. 513-522, 2014.

[11] M. Koppel, S. Argamon and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401-412, 2002.

[12] D. D. Pham, G. B. Tran and S. B. Pham, "Author Profiling for Vietnamese Blogs," in *International Conference on Asian Language Processing*, Singapore, 7-9 Dec. 2009.

[13] "PAN 2015," [Online]. Available: http://pan.webis.de/. [Accessed 7 2 2015].

[14] S. M. Z. Eissen, B. Stein and M. Kulig, "Plagiarism detection without reference collections," in *in Advances in Data Analysis*, Springer, 2007, pp. 359-366.

[15] M. Potthast, B. Stein, A. Barrón-Cedeño and P. Rosso, "An evaluation framework for plagiarism detection," in *In Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, Aug. 2010.

[16] N. Potha and E. Stamatatos, "A Profile-based Method for Authorship Verification," in *In Proceedings of the 8th Hellenic Conference on Artificial Intelligence*, Ioannina, Greece, May 2014.

[17] M. Koppel, J. Schler, S. Argamon and Y. Winter, "The "Fundamental Problem" of Authorship Attribution," *English Studies*, vol. 93, no. 3, p. 284–291, 2012.

[18] M. Koppel and Y. Winter, "Determining if Two Documents are by the Same Author," *Journal of the American Society for Information Science and Technology*, vol. 65, no. 1, p. 178–187, 2014.

[19] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski and B. S. Howald, "Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method," in *3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI)*, Athens, Greece, June 7–9, 2006.

[20] R. Layton, P. Watters and R. Dazeley, "Recentred local profiles for authorship attribution," *Natural Language Engineering*, vol. 18, no. 3, pp. 293 - 312, 2012.

[21] M. F. Tennyson and F. J. Mitropoulos, "Choosing a profile length in the SCAP method of source code authorship attribution," in *SOUTHEASTCON 2014*, Lexington, KY, 13-16 Mar. 2014.

[22] M. Koppel, J. Schler and S. Argamon, "Authorship Attribution in the Wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83-94, Mar. 2011.

[23] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538-556, Mar. 2009 .

[24] E. Stamatatos, G. Kokkinakis and N. Fakotakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics*, vol. 26, no. 4, p. 471–495, Mar. 2000.

[25] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," *Computers and the Humanities,* vol. 35, no. 2, pp. 193-214, 2001.

[26] P. C. Shlomo Argamon, S. Dhawle, S. Raj, H. Navendu and G. S. Levitan, "Stylistic text classification using functional lexical features," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 802-822, 2007.

[27] M. Halliday, Introduction to functional grammar (2nd ed.), London: Routledge, 1994.

[28] S. Argamon, M. Šarić and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington DC, USA, 2003.

[29] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems,* vol. 20, no. 5, pp. 67-75, Sep. 2005.

[30] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg and S. Levitan, "Stylistic text classification using functional lexical features," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 802-822, Apr. 2007.

[31] J. Burrows, "Not unless you ask nicely: The interpretative nexus between analysis and information," *Literary and Linguistic Computing*, vol. 7, no. 2, p. 91–109, 1992.

[32] J. Schler, E. Bonchek-dokow and M. Collins, "Measuring differentiability: Unmasking pseudonymous authors," *Journal of Machine Learning Research*, vol. 8, pp. 1261-1276, 2007.

[33] E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles," *International Journal on Artificial Intelligence Tools*, vol. 15, no. 5, pp. 823-838, 2006.

[34] I. Androutsopoulos, J. Koutsias, K. V. Cb and C. D. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, Athens, Greece, 2000.

[35] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization," in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[36] H. J. Escalante, M. Montes-y-Gómez and T. Solorio, "A Weighted Profile Intersection Measure for Profile-Based Authorship Attribution," in *Proceedings of 10th Mexican*

International Conference on Artificial Intelligence (MICAI 2011), Puebla, Mexico, 26 Nov. - 4 Dec. 2011.

[37] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," in *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[38] V. Keselj, F. Peng, N. Cercone and C. Thomas, "N-gram-based author profiles for authorship attribution," *Proceedings of the Pacific Association for Computational Linguistics*, pp. 255-264, 2003.

[39] M. Koppel and J. Schler, "Exploiting stylisticidiosyncrasies for authorship attribution," in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[40] R. Layton, S. McCombie and P. Watters, "Authorship Attribution of IRC Messages Using Inverse Author Frequency," in *Proceedings of the 2012 Third Cybercrime and Trustworthy Computing Workshop*, Washington DC, USA, 2012.

[41] D. Madigan, A. Genkin, D. D. Lewis, E. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, L. Ye and D. D. L. Consulting, "Author identification on the large scale," in *Proceedings of the Meeting of the Classification Society of North America*, 2005.

[42] T. Matsuura and Y. Kanada, "Extraction of authors' characteristics from Japanese modern sentences via n-gram distribution," in *Proceedings of the 3rd International Conference on Discovery Science*, Kyoto, Japan, 4–6 Dec. 2000.

[43] E. Stamatatos, "Author identification using imbalanced and limited training texts," in *Proceedings of the 4th International Workshop on Text-based Information Retrieval*, Regensburg, 3-7 Sept. 2007.

[44] B. Stein and S. M. z. Eissen, "Intrinsic plagiarism analysis with meta learning," in *Proceedings of the SIGIR Workshopon Plagiarism Analysis, Authorship Attribution, and NearDuplicate Detection*, Amsterdam, 2007.

[45] Y. Zhao and J. Zobel, "Effective and scalable authorship attribution using function words," in *Proceedings of Second Asia Information Retrieval Symposium*, Jeju Island, Korea, 13-15 Oct. 2005.

[۴۶] ا. شاهمیری, ر. دژکام و س. شیری, "شناسایی اشعار شاهنامه‌ی فردوسی به کمک شبکه عصبی مصنوعی," در *یازدهمین کنفرانس بین‌المللی کامپیوتر انجمن کامپیوتر ایران، پژوهشگاه دانش‌های بنیادی، پژوهشکده علوم کامپیوتر*, تهران، ایران, ۴ تا ۶ بهمن ۱۳۸۴.

[۴۷] ا. شاهمیری و م. م. بروجردی, "تعیین شاعر به کمک روش‌های یادگیری ماشین," در *سومین کنفرانس فناوری اطلاعات و دانش*, مشهد، ایران, ۶ تا ۸ آذر ۱۳۸۶.

[۴۸] ا. شاهمیری, س. ش. قیداری و ر. دژکام, "شناسایی اشعار شاهنامه‌ی فردوسی به کمک شبکه عصبی مصنوعی," *نشریه علمی پژوهشی انجمن کامپیوتر ایران*, جلد ۴, شماره ۳, 1385 ,26-17 .pp.

[۴۹] ز. فرهمندپور, ه. نیک‌مهر, م. منصوری‌زاده و ا. ط. قمصری, "مطالعه‌ی مقایسه‌ای روش‌های مبتنی بر یادگیری ماشین در تشخیص نویسنده‌ی فارسی‌زبان بر اساس سبک نوشتاری," در *نخستین کنفرانس بین المللی پردازش خط و زبان فارسی*, دانشگاه سمنان, ۱۵ و ۱۶ شهریور ۱۳۹۱.

[۵۰] ز. فرهمندپور, ه. نیک‌مهر, م. منصوری‌زاده و ا. ط. قمصری, "یک سیستم نوین هوشمند تشخیص هویت نویسنده‌ی فارسی‌زبان بر اساس سبک نوشتاری," *نشریه‌ی علمی-ترویجی محاسبات نرم*, جلد ۲, -26 .pp 35, پاییز و زمستان ۱۳۹۱.

[51] R. Ramezani, N. Sheydaei and M. Kahani, "Evaluating the Effects of Textual Features on Authorship Attribution Accuracy," in *3rd International Conference on Computer and Knowledge Engineering (ICCKE 2013)*, Mashhad, Iran, Oct. 31 & Nov. 1 2013.

[52] D. M.R., S. M. and A. M., "Farsi lexical analysis and stop word list," *Library Hi Tech*, vol. 27, no. 3, p. 435 – 449, 2009.

[53] H. J. Escalante, M. Montes-y-Gómez and T. Solorio, "A Weighted Profile Intersection Measure for Profile-Based Authorship Attribution," *Advances in Artificial Intelligence Lecture Notes in Computer Science*, vol. 7094, pp. 232-243, 2011.

[54] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis and D. Tambouratzis, "Discriminating the registers and styles in the Modern Greek language – Part 2: Extending the feature vector to optimize author discrimination," *Literary and Linguistic Computing*, vol. 19, no. 2, pp. 221-242, 2004.

[55] C. E. Chaski, "Who's at the keyboard? Authorship attribution in digital evidence investigations," *International Journal of Digital Evidence*, vol. 4, no. 1, 2005.

[56] O. d. Vel, A. Anderson, M. Corney and G. Mohay, "Mining e-mail content for author identification forensics," *ACM SIGMOD Record*, vol. 30, no. 4, pp. 55-64, Dec. 2001.

[57] J. Diederich, J. Kindermann, E. Leopold and G. Paass, "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, no. 1, pp. 109-123, 2003.

[58] G.-f. Teng, M.-s. Lai, J.-B. Ma and Y. Li, "E-mail authorship mining based on SVM for computer forensic.," in *Proceedings of the International Conference on Machine Learning and Cybernetics*, Aug. 2004.

[59] J. Li, R. Zheng and H. Chen, "From fingerprint to writeprint," *Communications of the ACM*, vol. 49, no. 4, pp. 76-82, 2006.

[60] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation," in *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering*, Stroudsburg, PA, USA, 2006.

[61] Ö. Uzuner and B. Katz, "A comparative study of language models for book and author recognition," in *Proceedings ofthe 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea, Oct. 11-13, 2005.

[62] Y. Zhao and J. Zobel, "Effective and scalable authorship attribution using function words," in *Proceedings of the 2nd Asia Information Retrieval Symposium*, Berlin, 2005.

[63] R. Zheng, J. Li, H. Chen and Z. Huang, "A framework for authorship identification of online messages: Writing style features and classification techniques," *Journal of the American Society of Information Science and Technology,* vol. 57, no. 3, pp. 378-393, Feb. 2006.

[64] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, Mar. 2002.

[65] R. Matthews and T. Merriam, "Neural computation in stylometry: An application to the works of Shakespeare and Fletcher," *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 203-209, 1993.

[66] T. Merriam and R. Matthews, "Neural compuation in stylometry II: An application to the works of Shakespeare and Marlowe," *Literary and Linguistic Computing,* vol. 9, no. 1, pp. 1-6, 1994.

[67] F. J. Tweedie, S. Singh and D. I. Holmes, "Neural network applications in stylometry: The Federalist Papers," *Computers and the Humanities,* vol. 30, no. 1, pp. 1-10, 1996.

[68] R. Zheng, J. Li, H. Chen and Z. Huang, "A framework for authorship identification of online messages: Writing style features and classification techniques," *Journal of the American Society of Information Science and Technology*, vol. 57, no. 3, pp. 378-393, 2006.

[69] F. Khosmood and R. Levinson, "Toward unification of source attribution processes and techniques," in *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, Dalian, China, 13-16 Aug. 2006.

[70] D. Holmes and R. Forsyth, "The Federelist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111-127, 1995.

[71] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary and Linguistic Computing*, vol. 22, no. 3, pp. 251-270, 2007.

[72] B. Kjell, "Discrimination of authorship using visualization," *Information Processing and Management*, vol. 30, no. 1, pp. 141-150, 1994.

[73] R. Forsyth and D. Holmes, "Feature-finding for text classification," *Literary and Linguistic Computing*, vol. 11, no. 4, pp. 163-174, 1996.

[74] F. Peng, D. Schuurmans, S. Wang and V. Keselj, "Language independent authorship attribution using character level language models," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2003.

[75] V. Keselj, F. Peng, N. Cercone and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the Pacific Association for Computational Linguistics*, 2003.

[76] E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles," *International Journal on Artificial Intelligence Tools*, vol. 15, no. 5, pp. 823-838, 2006.

[77] R. Baayen, H. van Halteren, A. Neijt and F. Tweedie, "An experiment in authorship attribution," in *Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*, 2002.

[78] M. Koppel and J. Schler, "Exploiting stylisticidiosyncrasies for authorship attribution," in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[79] M. Gamon, "Linguistic correlates of style: Authorship classification with deep linguistic analysis features," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[80] G. Hirst, "Bigrams of syntactic labels for authorship discrimination of short texts," *Literary and Linguistic Computing*, vol. 22, no. 4, p. 405–417, 2007.

[81] H. v. Halteren, " Author verification by linguistic profiling: An exploration of the parameter space," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 1, pp. 1-17, 2007.

[82] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and H. R., "Indexing by latent semantic analysis," *Journal of the American Society for Information Science,* vol. 41, no. 6, pp. 391-407, 1990.

[83] P. McCarthy, G. Lewis, D. Dufty and D. McNamara, "Analyzing writing styles with coh-metrix," in *Proceedings of the Florida Artificial Intelligence Research Society International Conference,* 2006.

[84] Y. Marton, N. Wu and L. Hellerstein, "On compression-based text classification," in *Proceedings of the European Conference on Information Retrieval*, 2005.

[85] F. Peng, D. Shuurmans and S. Wang, "Augmenting naive Bayes classifiers with statistical language models," *Information Retrieval Journal,* vol. 7, no. 1, pp. 317-345, 2004.

[86] W. R. Bennett, Scientific and engineering problem-solving with the computer, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1976.

[87] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Automatic authorship attribution," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 1999.

[88] E. Stamatatos, G. Kokkinakis and N. Fakotakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics,* vol. 26, no. 4, pp. 471-495, 2000.

[89] V. Keselj and N. Cercone, "CNG Method with Weighted Voting," in *Ad-hoc Authorship Attribution Competition (AAAC), June 2004. Part of ALLC/ACH 2004 conference, extended abstract,*, 2004.

[90] S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier," in *International Conference on Communications and Information Technology (ICCIT)*, 2012.

[91] F. Mosteller and D. L. Wallace, Inference and disputed authorship: The Federalist, Addison-Wesley, 1964.

[92] M. Eder, "Does size matter? : autorship attribution, short samples, big problem," in *Digital humanities 2010 conference*, London, 2010.

[93] S. Ouamour and H. Sayoud, "Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features," in *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),*, 10-12 Oct. 2013.

[94] R. Layton, P. Watters and R. Dazeley, "Authorship Attribution for Twitter in 140 Characters or Less," in *Second Cybercrime and Trustworthy Computing Workshop (CTC)*, 19-20 Jul. 2010.

[۹۵] س. نفیسی, گلستان سعدی, تهران: چاپ مروی, ۱۳۷۰.

[۹۶] ح. یزدان‌پرست, "پیشگفتار," در *شرح و ساده‌نویسی گلستان سعدی*, تهران, انتشارات اطلاعات, ۱۳۸۹.

[۹۷] ا. منزوی, "تتبع در گلستان سعدی," *مجله‌ی وحید*, جلد ۱۱۳, 1352 ,pp. 167-178.

[۹۸] م. بهار, سبک شناسی (تاریخ تطور نثر فارسی), تهران: نشر امیرکبیر, ۱۳۸۹.

Abstract

Authorship Attribution is a subfield of Text-Processing and aims to determine identity of a text's writer. In the other word, main objective of this subfield is to design a system which can "attribute" an unlabeled text (text with unknown author) to one of candidate authors. To design such a system we must access some labeled texts (texts with known author) for each candidate author.

All previous research on authorship attribution of Persian texts were belonged to NLP-based methods; but the main objective of this thesis is to study performance of NDP methods in solving Persian author attribution problems. These methods are designed based on "N-grams" and are completely independent of NLP systems.

In this thesis most important NDP methods are studied and then two novel methods are proposed based on them. The first proposed method (CNG-WIS) uses "Indices" of n-grams rather than their "Frequencies". The second proposed method (VNG) uses "Variant N-grams" rather than "Most-Frequent N-grams".

In order to evaluate studied methods and also compare proposed method with them, we use four different corpuses (i.e. databases). One of them is gathered by author and contains 145 text from 6 Persian contemporary authors. Results indicates that in addition to studied NDP methods, the two proposed methods are powerful in solving authorship attribution problems in Persian and Arabic texts.

At the end, two especial problem in Persian Literature are studied: Golestan's Rivals and Indian-Styled Sonnets. For this purpose, two additional corpuses are gathered by the author: GBP (contains 75 Hekayats from 3 authors) and SBH (contains 90 sonnets from 3 poets). Results indicates that in addition to Persian prose texts, NDP methods have good performance on Hekayat (mixture of prose and poem) and poems.

Keywords: authorship attribution, n-gram, profile-based methods, style marker, CPPT corpus, GBP corpus, SBH corpus.

Shahrood University

Faculty of Electrical and Robotics Engineering

Dissertation Submitted in Partial Fulfillment of the Requirements for

the Degree of Master of Science (M.Sc.) in Electronic Engineering

Subject-Analysis and Author-Attribution in Farsi-Arabic Documents Using Structural Information of the Text

Ali Shahnama

Supervisors:

Dr. Alireza Ahmadyfard

Dr. Hossein Marvi

Advisor:

Dr. Morteza Zahedi

February 2015