
بسم الله الرحمن الرحيم
الحمد لله رب العالمين الرحمن الرحيم مالك يوم الدين
اياك نعبد واياك نستعين اهدنا الصراط المستقيم
صراط الذين انعمت عليهم غير المغضوب عليهم ولا الضالين

دانشگاه شاهرود

مرکز آموزش‌های الکترونیک

گروه مهندسی کامپیوتر (گرایش هوش مصنوعی)

فشرده‌سازی متن با استفاده از تکنیک‌های هوش مصنوعی

دانشجو : محبوبه سلیمانیان

استاد راهنما :

دکتر علی اکبر پویان

استاد مشاور :

دکتر هدی مشایخی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

شهریور ماه ۱۳۹۴

دانشگاه صنعتی شاهرود

مرکز آموزش‌های الکترونیک

گروه مهندسی کامپیوتر (هوش مصنوعی)

پایان نامه کارشناسی ارشد آقای/ خانم محبوبه سلیمانیان

تحت عنوان: فشرده سازی متن با استفاده از تکنیک های هوش مصنوعی

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم به

به پاس تعبیر عظیم و انسانی شان از کلمه ایثار و از خودگذشتگان

به پاس عاطفه سرشار و گرمای امیدبخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است

به پاس قلب‌های بزرگشان که فریاد رس است و سرگردانی و ترس در پناهشان به شجاعت می‌گراید

و به پاس محبت‌های بی دریغشان که هرگز فروکش نمی‌کند

این پایان نامه را تقدیم می کنم به:

چشمه‌های جوشان محبت

جلوه‌های مهر و عطوفت الهی

لبخندهای پر مهر زندگیم

پدر و مادر عزیزم

که در تمام مراحل زندگی، به من راه و رسم درست زیستن را آموختند.

تشکر و قدردانی

شکر شایان نثار ایزد منان که توفیق را رفیق راهم ساخت تا این پایان نامه را به پایان برسانم. سپاس خدایی را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت‌های او ندانند و کوشندگان، حق او را گزاردن نتوانند. و سلام و دورد بر محمّد و خاندان پاک او، طاهران معصوم، هم آنان که وجودمان وامدار وجودشان است؛

بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از زحمات بی‌شائبه‌ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم.

اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می‌کند و سلامت امانت‌هایی را که به دستش سپرده‌اند، تضمین؛ برحسب وظیفه و از باب " من لم یشکر المنعم من المخلوقین لم یشکر اللّه عزّ و جلّ" : از پدر و مادر عزیزم...این دو معلم بزرگوارم... که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت‌هایم گذشته‌اند و در تمام عرصه‌های زندگی یار و یاوری بی‌چشم داشت برای من بوده اند؛ از استاد با کمالات و شایسته؛ جناب آقای دکتر علی اکبر پویان که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛ از استاد صبور و بزرگوارم، سرکار خانم دکتر هدی مشایخی، که زحمت مشاوره این رساله را در حالی متقبل شدند که بدون مساعدت ایشان، این پروژه به نتیجه مطلوب نمی رسید ؛ کمال تشکر و قدردانی را دارم.

باشد که این خردترین، بخشی از زحمات آنان را سپاس گوید.

تعهد نامه

اینجانب ..... دانشجوی دوره کارشناسی ارشد رشته ..... دانشکده

..... دانشگاه صنعتی شاهرود نویسنده پایان نامه .....

.....تحت راهنمائی.....متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

رشد روز افزون اطلاعات دیجیتالی در سال‌های اخیر، موجب افزایش توجهات به فشرده-سازی متون شده است. اطلاعاتی از نوع متن که همه روزه شاهد ارسال و دریافت آن هستیم. نیاز به کاهش میزان داده‌ها و صرفه‌جویی در فضای ذخیره‌سازی، فشرده‌سازی را به امری مهم تبدیل نموده است. با افزایش روز افزون متون غیرانگلیسی و غیرلاتین، نیاز به رشد الگوریتم‌های فشرده‌سازی در زبان‌های دیگر نیز احساس می‌شود. این پایان‌نامه تلاشی در راستای ارائه تکنیکی جهت فشرده‌سازی متون فارسی است.

در این پژوهش هدف استفاده از قواعد و تکنیک‌های مدلسازی زبان می‌باشد. قواعدی که در الگوریتم‌های فشرده‌سازی معروف و پرکاربردی مانند زیپ مورد توجه قرار نگرفته است. در این تکنیک ما با استفاده از مدل آماری N-gram، احتمال قرار گرفتن دنباله‌ای از کلمات و کاراکترهای زبان را، بعد از دیگری با در نظر گرفتن پارامترهای تعداد تکرار و طول عبارت بررسی می‌کنیم. جهت ارزیابی و انتخاب مدلی با میزان کارایی بیشتر از معیار سرگشتگی که مستقل از سیستم و متناسب با احتمال‌های نسبت داده شده به عبارات (دنباله‌ای از کلمات و کاراکترها) می‌باشد، استفاده شده است. نتایج بدست آمده میزان فشرده‌سازی ۸۲٪ متن ورودی را با استفاده از الگوریتم پیشنهادی و در نظر گرفتن اطلاعات زبانی، در فایل فشرده بدست آمده از الگوریتم فشرده‌سازی زیپ نشان می‌دهد. در فصل‌های آتی مراحل مختلف تحلیل بر اساس مدل زبانی، مراحل ارزیابی و نتایج بدست آمده تشریح خواهد شد.

کلمات کلیدی؛ فشرده‌سازی، کدینگ اطلاعات، متون فارسی، مدل زبانی

فهرست مطالب

فهرست جداول

فهرست اشکال

فهرست نمودارها

فصل اول
کلیات تحقیق

۱-۱- مقدمه

امروزه جهت دسترسی سریع و آسان به اطلاعات، در هر زمان و مکان، میلیون‌ها شرکت و موسسه دولتی و غیردولتی در سرتاسر جهان اقدام به ضبط دیجیتالی اطلاعات خود می‌نمایند. میلیون‌ها بایت اطلاعات دیجیتالی که به طور روزمره توسط سازمان‌های الکترونیکی و دستگاههای دیجیتالی مختلف در قالب‌های متن، صوت، تصویر، ویدئو و .. در حال ذخیره‌سازی و انتقال هستند. یکی از جنبه‌های مهم ارتباطات، انتقال داده‌ها از سرویس‌دهنده به سرویس گیرنده می‌باشد.

هر منبع اطلاعاتی که یکسری اطلاعات را در دنیای بیرون منتشر می‌کند، دارای الفبا و زبان خاصی است که منحصر به آن منبع می‌باشد. این زبان، همان قواعد و ساختارهایی است که موجب می‌شود، سمبول‌های الفبای منبع در کنار هم قرار گیرند و معنا و مفاهیم خاصی را به دنیای اطراف خود منتشر کنند. اطلاعات استخراج شده از این منابع دارای اضافاتی است که در اصطلاح به آن، افزونگی می‌گویند. به بیان دیگر هر منبع مقداری داده بیشتر از آنچه که از آن انتظار می‌رود به بیرون ساطع می‌کند. این اضافات داده‌های مطلوبی نیست و سعی بر آن است بمنظور استفاده بهینه به داده‌های مطلوب تبدیل شوند.[1] روش‌ها و الگوریتم‌هایی که جهت کاهش افزونگی داده‌ها طراحی می‌شوند، الگوریتم‌های کدگذاری منابع یا فشرده‌سازی داده‌ها می‌گویند الگوریتم‌هایی که تا کنون در آن‌ها به استفاده از اطلاعات زبانی توجهی نشده است.

انتقال داده‌ها در عصر اینترنت وابستگی زیادی به زمان دارد. با استفاده از فشرده‌سازی، زمان لازم برای انتقال فایل‌ها کاهش می‌یابد. فشرده‌سازی به معنی کاهش سایز فیزیکی داده‌ها می‌باشد به‌گونه‌ای که فضا و حافظه ذخیره‌سازی کمتری اشغال نماید. بر این اساس انتقال فایل‌های فشرده‌سازی شده آسانتر می‌باشد و به همین دلیل زمان انتقال کاهش می

یابد. این امر مبین نیاز بشر به فشرده‌سازی اطلاعات، با توجه به دو مفهوم محدودیت زمان و مکان، می‌باشد. فشرده‌سازی داده‌ها به معنای حذف زوائد از کدهای تخصیص داده شده به سمبول‌های منبع داده‌هاست که نهایتاً هزینه‌ی ذخیره‌سازی و تبادل اطلاعات را کاهش می‌دهد. به عبارت دیگر فشرده‌سازی یک متن، فرآیندی برای کاهش حجم متن، با استفاده از یک برنامه‌ی کامپیوتری به منظور ایجاد و نگهداری یک متن با طول کوتاه‌تر می‌باشد. فشرده‌سازی داده‌ها یک نوع عمل کدینگ است که در آن داده‌های ورودی به طریقی کد می‌شوند که فضای کمتری را اشغال نموده و نیز بتواند دوباره در هر زمان دلخواه بازیابی شده و داده‌های اصلی را برای ما بازگردانند.[2] به عبارت دیگر فشرده‌سازی فرآیند کدینگ اطلاعات دیجیتال، بمنظور کاهش اندازه آن است.

الگوریتم‌های کدگذاری[1] منابع و فشرده‌سازی داده‌ها به دو دسته بی اتلاف و با اتلاف تقسیم می‌شوند[3]؛

- فشرده‌سازی بی‌اتلاف[2]: در این فشرده‌سازی کاهش بیت بر اساس شناسایی و حذف افزونگی آماری انجام می‌گیرد و هیچ اطلاعاتی از بین نمی‌رود.
- فشرده‌سازی اتلافی[3]: در این فشرده‌سازی کاهش بیت بر اساس شناسایی و حذف اطلاعات کم ارزش صورت می‌گیرد و این اطلاعات برگشت‌پذیر نخواهد بود.

ملاحظه می‌شود که در فشرده‌سازی بی‌اتلاف، هیچ‌گونه اطلاعاتی از داده‌های منبع از دست نمی‌رود اما در روش‌های فشرده‌سازی با اتلاف مقداری از اطلاعات مبداء جهت رسیدن به فشرده‌سازی بهتر از دست خواهد رفت. با توجه به حساس نبودن حواس بشر به بعضی تغییر و تحولات در مبداء داده‌ها، روش کدگذاری و فشرده‌سازی با اتلاف می‌تواند میزان فشرده-سازی بهتری نسبت به روش فشرده‌سازی بی‌اتلاف برسد.

[1] Coding
[2] Lossless
[3] Lossy

۱-۲- بیان مسئله

هدف اصلی در این پژوهش ارائه یک تکنیک فشرده‌سازی کارآمد در متون فارسی می‌باشد، نیاز شدید به کاهش تعداد بیت‌ها در ذخیره‌سازی اطلاعات و صرفه‌جویی در میزان مصرف حافظه، مقوله‌ی فشرده‌سازی را پر اهمیت کرده است[4].

استفاده از تکنیک‌های قدرتمند فشرده‌سازی در ذخیره‌سازی اطلاعات بسیار مهم می‌باشد. همچنین در راستای رسیدن به اهدافی همچون صرفه‌جویی در فضای ذخیره‌سازی و استفاده هر چه بیشتر از فضای محدود ذخیره‌سازی استفاده از روش‌های کارآمد فشرده-سازی امری مهم به نظر می‌رسد. فشرده‌سازی می‌تواند با اتلاف یا بدون اتلاف باشد، در فشرده‌سازی متون از دست رفتن اطلاعات قابل قبول نیست، لذا از الگوریتم فشرده‌سازی بدون اتلاف استفاده می نماییم.

هدف از فشرده نمودن فایل‌ها کاهش ظرفیت فایل‌ها می‌باشد. همچنین در زمان استفاده از فایل می‌بایست مجدداً فایل به حالت اولیه برگردانده شود. در فرآیند فوق بیت‌هایی از فایل با استفاده از الگوریتم‌هایی خاص، از فایل حذف و زمینه کاهش ظرفیت فایل فراهم خواهد شد. در زمان استفاده از فایل با استفاده از الگوریتم فشرده‌سازی عملیات معکوس انجام و فایل به حالت اولیه خود برگردانده خواهد شد.

یک مدل آماری زبان، احتمال جملات را توصیف می‌کند. یعنی با استفاد از آن می‌توان با توجه به اطلاعات جاری، احتمال کلمات بعد را پیش‌بینی کرد. در این تحقیق افزایش میزان فشرده‌سازی اطلاعات متنی با استفاده از اطلاعات زبانی مد نظر می‌باشد (قواعدی که به کار بردن آن در الگوریتم‌های فشرده‌سازی معروف مورد توجه قرار نگرفته است). لذا به طراحی و تنظیم مدل‌های آماری برای کلمات و کاراکترها پرداخته شده است. در هر مسئله مدل-سازی در ابتدا می‌بایست دو عامل اصلی مشخص گردد؛ در ابتدا ساختار مناسب و جامع

بـرای مـدل و سـپس چگـونگی تنظـیم و محاسـبه پارامترهـای آن. در ایـن تحقیـق بـر اسـاس مطالعـات و بررسـی‌هـای همـه جانبـه و بـا توجـه بـه سـوابق، مـدل N-gram بـا توجـه بـه مقـاوم بـودن در برابـر نـویز بـرای کلمـات و کاراکترهـا انتخـاب، و بـرای تنظـیم و محاسـبه پارامترهـا تکنیک‌های هوش‌مصـنوعی مـورد اسـتفاده قـرار خواهـد گرفـت. جهـت انجـام فرآینـد یـادگیری بـه آماده‌سازی مجموعه‌ای از داده‌های آموزشی و سپس به بهینه‌سازی آن پرداخته شده است.

۱-۳- اهمیت و ضرورت تحقیق

کامپیوترهـا از زبـان مخـتص بـه خـود یعنـی زبـان بـاینری کـه رشـته‌ای از اعـداد ۰ و ۱ اسـت، استفاده می‌کننـد. اگـر چـه بـا گذشـت زمـان فضـای بیشـتری بـرای کـار بـا فایـل‌هـای بزرگتـر در اختیـــــار کـاربران گذاشـته شـده اسـت ولـی هنـوز مشـکل فضـای ذخیـره‌سـازی و تنگناهـای پهنای بانـد مطـرح اسـت. فشـرده‌سـازی تکنیکـی بـرای رسـیدن بـه کـاهش حجـم در داده‌هاسـت. پیچیـدگی‌هـای محاسـباتی بـالا مشـکلات عمـده در ذخیـره‌سـازی داده‌هـا، کمبـود حافظـه، زمـان زیاد صـرف شـده در انتقـال داده‌هـا از جملـه مشـکلات موجـود در زمـان کـار بـا داده‌هـای حجـیم است. از این رو تحقیـق در زمینـه‌ی فشـرده‌سـازی داده‌هـا امـری ضـروری و تعیـین‌کننـده بـه نظـر مـی‌رسـد. بـا توجـه بـه گسـترش حجـم اطلاعـات، ذخیـره‌سـازی داده‌هـا در حجـم محـدود، پیچیـدگی در ارسـال داده‌هـا و همـراه بـا آن نیـاز بـه اسـتفاده از آنهـا در زمـان مناسـب از اهمیـت ویژه‌ای برخوردار اسـت لـذا تحقیـق بـر روی فشـرده‌سـازی داده‌هـای متنـی امـری مهـم و ضـروری به نظر می‌رسد.

۱-۴- اهداف تحقیق

هدف از فشرده‌سـازی داده‌هـا، کـاهش تعـداد بیـت‌هـا و نمـایش آن بـا حتـی المکـان تعـداد بیـت-هـای کمتـر، در مقابـل ایجـاد اعوجـاج قابـل قبـول در دیتـای بازسـازی شـده اسـت بـا توجـه بـه

ضرورت‌های اشاره شده، هدف کلی از این تحقیق ایجاد یک سیستم فشرده‌سازی کارآمد و مناسب برای داده‌های متنی می‌باشد. در این راستا هدف اصلی از این تحقیق بصورت زیر بیان می‌گردد؛

"ارائه یک تکنیک که قابلیت فشرده‌سازی متون فارسی را دارا باشد"

الگوریتم‌های فشرده‌سازی بسیاری وجود دارند که بمنظور انجام فشرده‌سازی از افزونگی موجود در منابع استفاده می‌نمایند. این الگوریتم‌ها برای فشرده‌سازی متن کاربرد بسیار خوبی دارند. در این پایان‌نامه با توجه به عدم استفاده الگوریتم‌های فشرده‌سازی پرکاربرد از اطلاعات زبانی، قوانین مدلسازی زبان را برای فشرده‌سازی متون بکار می‌بریم و به مقایسه آن با سایر روش‌های بدون اتلاف می‌پردازیم.

۱-۵- فرضیه‌های تحقیق

- در این پژوهش، استفاده از فشرده‌سازی بدون اتلاف مد نظر خواهد بود.
- منبع داده ورودی اولیه که مورد استفاده قرار می‌گیرد، و تکنیک فشرده‌سازی مورد نظر بر روی آن ارزیابی خواهد شد، صرفاً از کاراکترهای فارسی و علائم نگارشی پرکاربرد تشکیل شده است.
- کدینگ استفاده شده در متن اولیه، استاندارد یونیکد خواهد بود.
- از قوانین مدلسازی زبان در ارائه تکنیک فشرده‌سازی استفاده خواهد شد.

۱-۶- ساختار مطالب

در فصل اول پس از بیان مقدمه و طرح مسئله، ضرورت تحقیق و اهداف تحقیق ارائه شده است. در فصل دوم به تشریح فشرده‌سازی پرداخته و روش‌های با اتلاف و بدون اتلاف مورد

بررسی قرار خواهد گرفت. در ادامه روش‌های فشرده‌سازی و فرضیات عنوان شده در این روش‌ها بیان خواهند شد. سپس به توضیح چند روش فشرده‌سازی در داده‌های متنی خواهیم پرداخت.

در فصل سوم به تشریح روش پیشنهادی پرداخته و بخش‌های مختلف آنرا مورد بررسی قرار می‌دهیم. فصل چهارم مربوط به پیاده‌سازی و ارزیابی نتایج بدست آمده خواهد بود. در این فصل به تجزیه و تحلیل داده‌ها و مقایسه الگوریتم پیشنهادی با سایر روش‌ها خواهیم پرداخت. در فصل پنجم به نتیجه‌گیری و ارائه پیشنهادات به منظور تقویت هر چه بیشتر الگوریتم پیشنهاد شده پرداخته می‌شود.

فصل دوم

ادبیات موضوعی تحقیق

۲-۱- مقدمه

فرآیند فشرده‌سازی اطلاعات، اعمال تغییراتی روی آنهاست که منجر به کاهش حجم ذخیره‌سازی شود و در عین حال اطلاعات اولیه را بتوان بی کم و کاست و یا با خطایی که از دیدگاه کاربر یا سامانه‌ی مصرف‌کننده‌ی اطلاعات قابل چشم‌پوشی باشد، بازیابی نمود. این کار با از میان برداشتن برخی داده‌های تکراری و یا قابل پیش‌بینی و یا اعمال تغییراتی بر روی اطلاعات که چنین داده‌هایی را تا حد ایجاد حداقل همبستگی در درون اطلاعات، مشخص و پالایش نماید، انجام می‌شود.

الگوریتم‌های فشرده‌سازی، طبیعتاً با نوع اطلاعاتی که ورودی فرآیند فشرده‌سازی هستند، بستگی مستقیم دارند. اطلاعات متنی که موضوع این پایان‌نامه هستند، به خاطر ویژگی‌های داده‌های متنی در زبان‌های مختلف، خود شاخه‌ای مجزا از دانش فشرده‌سازی را ایجاد کرده‌اند. نظریه‌ی اطلاعات، که به طور دقیقی مستقل از نوع اطلاعات، پایه‌ها و ابزارهای فشرده‌سازی را تبیین، طبقه‌بندی، تحلیل و تولید می‌نماید در زمینه‌ی فشرده‌سازی متن نیز از نقش بسزایی برخوردار است، اما در عین حال ویژگی‌های خاص متون، ماهیت فشرده-سازها را تا اندازه قابل توجهی به خود وابسته نموده است.

مسئله به بیان ساده این است که چگونه می‌توان به کمترین حجم داده‌ها رسید، به نحوی که هیچ اطلاعاتی از داده‌های ارسالی منبع از دست نرود.

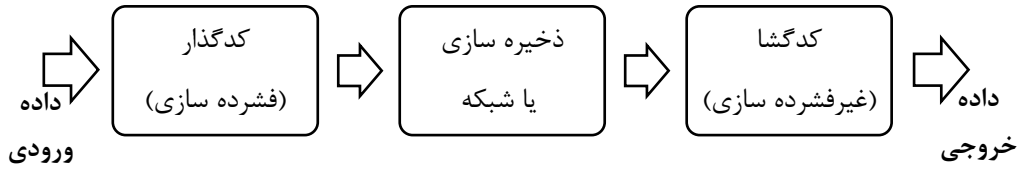
با نگاهی گذرا به متن فایل‌هایی که ما استفاده می‌کنیم دیده خواهد شد. تقریباً تمام فایل‌های متنی موجود از تعداد بسیار محدودی از لغات و عبارات تکراری تشکیل شده است که با قرار گرفتن در پشت سر یکدیگر عبارت و مفهوم‌های جدیدی را ساخته‌اند. کاری که روش‌های فشرده‌سازی انجام می‌دهند پیدا نمودن و حذف این تکرارهای زائد است.[2] این روش‌ها با پیدا نمودن لغات و عباراتی که بیش از دیگران تکرار می‌شوند و

جایگزین نمودن آنها با علائمی مناسب با طول کمتر داده‌ها را فشرده‌تر از آنچه که بوده است می‌نماید.

۲-۲- فشرده‌سازی

در علوم کامپیوتر و نظریه اطلاعات، فشرده‌سازی به معنای کدگذاری[1] داده‌ها به نحوی است که بیت کمتری را نسبت به داده‌ی اولیه آن داشته باشد. در واقع فشرده‌سازی فرآیند کدگذاری است که منجر به کاهش مؤثر تعداد کل بیت‌های لازم برای ارائه اطلاعات مشخصی می‌شود.

فشرده‌سازی به دو دسته با اتلاف و بی‌اتلاف تقسیم می‌شود.



شکل (۲-۱) ، نمایی از شمای کلی فشرده‌سازی

۲-۲-۱- فشرده‌سازی بی‌اتلاف[2]

ایده‌ای که در آن طی فرآیند فشرده‌سازی داده‌ها هیچ اطلاعات مفیدی از دست نرود و بازسازی دقیق آن طی فرآیند رمزگشایی مقدور باشد. در کدینگ بدون اتلاف مقادیر نمونه-ی اصلی دوباره دقیقاً بدست خواهد آمد و فشرده‌سازی بوسیله‌ی پیدا کردن زوائد آماری سیگنال انجام می‌شود. در این نوع فشرده‌سازی کیفیت خوب و نرخ فشرده‌سازی پایین خواهیم داشت، در نتیجه از آن در مواردی که دقت خیلی مهم است استفاده می‌شود. [5,6]

از جمله ویژگی‌های این گروه عبارت است از؛

[1] Coding
[2] Lossless Compression

- کاهش اندازه داده در اثر فشرده‌سازی
- عدم حذف هیچ قسمتی از بخش‌های داده در زمان فشرده‌سازی
- امکان بازیابی داده اصلی بطور کامل و با تمام جزئیات از روی داده فشرده شده
- میزان فشرده‌سازی پایین (در حدود ۳ تا ۴ برابر)
- مناسب برای فشرده‌سازی متن و اطلاعات باینری

در فشرده‌سازی بدون اتلاف، در هنگام فرآیند فشرده‌سازی یا در حالت خارج کردن از فشردگی، داده‌ها تغییر نکرده یا مفقود و ضایع نمی‌شوند. فرآیند خارج کردن از حالت فشردگی یک نسخه المثنی از شیء فشرده‌سازی ایجاد می‌کند از این روش فشرده‌سازی برای اسناد متنی، بانک‌های اطلاعاتی متنی و اشیا مرتبط با متن استفاده می‌شود.[7]

انواع روش‌های کدینگ (فشرده‌سازی) بدون اتلاف

- کدینگ مبتنی بر آنتروپی
- کدینگ مبتی بر دیکشنری (واژه‌نامه)

۲-۲-۱-۱- کدینگ مبتنی بر آنتروپی

فرض اساسی نظریه اطلاعات بر این است که می‌توان تولید اطلاعات را به صورت فرآیند احتمالی در نظر گرفت. آنتروپی بیانگر عدم قطعیت در منبع اطلاعاتی است. با افزایش آنتروپی، عدم قطعیت افزایش یافته و در نتیجه آن اطلاعات بیشتری به منبع منتسب می‌شود. شانون ثابت می‌کند که آنتروپی بیانگر متوسط اطلاعات موجود در هر کاراکتر (بر حسب بیت بر کاراکتر) می‌باشد و بنابراین حد بالای نرخ فشرده‌سازی یک متن (بر حسب بیت بر کاراکتر) برابر با آنتروپی آن می‌باشد. [8]

آنتروپی حد پایین میانگین تعداد بیت‌ها بمنظور کد کردن هر مجموعه نشانه S را مشخص می‌کند. آنتروپی η از یک منبع اطلاعاتی $S = \{s_1 , s_2 , \dots , s_n\}$ برابر است با:

$$\eta = H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (۱-۲)$$

p_i – احتمال رخداد s_i در S

$\log_2 \frac{1}{p_i}$ مشخص کننده میزان اطلاعات موجود در s_i است، بنابراین می‌تواند تعیین کننده تعداد بیت‌های مورد نیاز برای کد کردن s_i باشد.

در این دسته به ازای هر نشانه موجود در داده یک کلمه رمز یکتا در نظر گرفته می‌شود. بر اساس نظریه شانون بهینه‌ترین طول کد برای نشانه از رابطه زیر پیروی می‌کند که در آن p_i احتمال رخداد نشانه iام است.

$$\log_2 \frac{p_i}{2} \qquad (۲-۲)$$

از جمله روش‌های فشرده‌سازی مبتنی بر آنتروپی می‌توان به VLE، هافمن، حسابی و .. اشاره نمود.

۲-۲-۱-۲- کدینگ هافمن

روش هافمن در گذشته بیشترین توجه را به خود جلب نموده است. نزدیکی بسیار زیاد این روش به اصول نظریه‌ی اطلاعات، موجب شده است که تحقیقات گسترده‌ای برای بررسی توانایی‌ها و کاستی‌های آن صورت پذیرد. در علوم رایانه و تئوری اطلاعات الگوریتم هافمن یک الگوریتم کدگذاری برای فشرده‌سازی بی‌اتلاف اطلاعات است[9]. الگوریتم هافمن جزو خانواده الگوریتم‌هایی است که طول کد متغییری دارند. این به آن معناست که نمادهای مجزا (برای نمونه کاراکترهایی در یک فایل متنی) با رشته بیت‌هایی که طول‌های مختلفی دارند تعویض می‌شود. بنابراین به نمادهایی با احتمال تکرار بیشتر در یک فایل یک رشته

بیت کوتاه (کد کوچک) اختصاص داده می‌شود، در حالی‌که نمادهای دیگر که به ندرت دیده می‌شوند رشته بیت طولانی‌تری را می‌گیرند.[10]

این روش در ابتدا با ایجاد آماری اولیه از فایل ورودی، درختی به نام درخت هافمن ایجاد می‌نماید.کاربرد این درخت در تخصیص کد به کاراکترهای ورودی و خصوصیت آن در تخصیص کد کوتاه‌تر به کاراکتر با تکرار بیشتر به صورت بهینه است. سپس در مروری دیگر بر روی فایل هر کاراکتر با کد اختصاص داده شده به آن جایگزین می‌شود. در انتها نیز با ذخیره‌سازی کد هر کاراکتر استخراج داده‌ها به داده‌های ابتدایی ممکن خواهد شد. درخت هافمن تضمین می‌کند که کدهای تخصیص داده شده، کاراترین کدها باشند و لذا این روش در فشرده‌سازی استفاده می‌شود.

۲-۲-۱-۳- کدینگ مبتنی بر واژه‌نامه

یکی از روش‌های عمده فشرده‌سازی، روش لغت‌نامه‌ای است که کاربرد زیادی هم در برنامه‌های تجاری فشرده‌سازی دارد. [11] بسیاری از برنامه‌های فشرده‌سازی از مدل‌های متفاوت الگوریتم مبتنی بر واژه‌نامه ایجاد شده توسط "Lempel-Ziv"[12,13] ، بمنظور کاهش ظرفیت فایل‌ها، استفاده می-نمایند. منظور از واژه‌نامه در الگوریتم فوق، روش‌های کاتولوگ نمودن بخش‌هایی از داده است. در این دسته معمولاً مجموعه‌ای از زیر رشته‌ها در یک ساختمان داده که واژه‌نامه نام دارد ذخیره می‌شود و رمزکننده هرگاه با الگویی منطبق با یکی ار مدخل‌های واژه‌نامه پیدا کند، آن را با کد معادل در واژه-نامه جایگزین می‌کند. این واژه‌نامه برای رمزگذار و رمزگشا شناخته شده است و در طی عمل رمزگذاری و رمزگشایی تولید می‌شود.

سیستم استفاده شده برای سازماندهی واژه‌نامه متفاوت و در ساده‌ترین حالت می‌تواند شامل یک لیست عددی باشد. در این روش با مراجعه متن مورد نظر، کلمات تکراری را انتخاب و آنها را در لیست مرتب شده‌ای، ایندکس می‌نماییم. پس از ایجاد لیست، می‌توان در مواردی‌که از کلمات در متن اولیه

استفاده می‌شود، از اعداد نسبت داده شده و متناظر با آنها استفاده کرد. یکی از مسائل مهم و اساسی در این روش که نقطه اختلاف الگوریتم‌های مختلف است استراتژی تشکیل واژه‌نامه و نحوه اضافه و کم کردن لغات به آن است. البته آنچه معقول بنظر می‌رسد این است که کلماتی که تا به حال بیشتر تکرار شده‌اند در آینده نیز با احتمال بیشتری ظاهر خواهند شد و بنابراین کلمات با فرکانس تکرار بیشتر از اولویت بیشتری برای حضور در واژه‌نامه برخوردارند. [11] از جمله روش‌های فشرده‌سازی مبتنی بر واژه‌نامه می‌توان به RLE، LZW ، Deflate و .. اشاره نمود.

۲-۲-۱-۳-۱- Lzw[1]

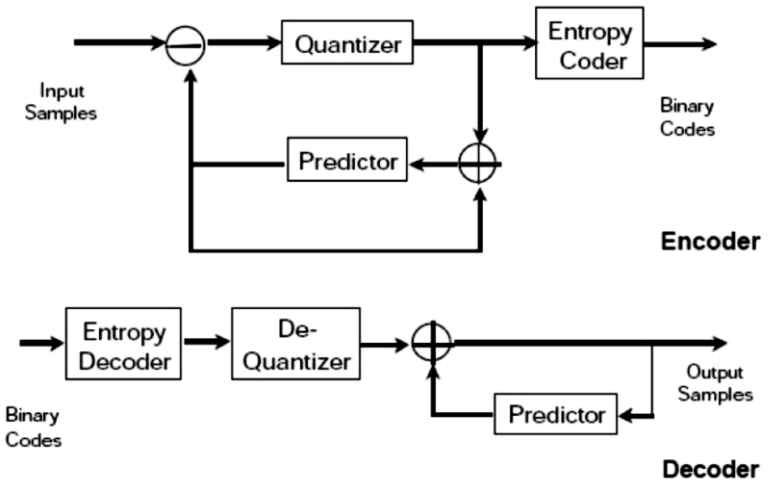
LZW یک الگوریتم همه منظوره است که قابل اعمال بر روی هر فایلی اعم از متنی یا دودویی می-باشد [4]. LZW الگوریتم مورد استفاده در بسیاری از نرم‌افزارهای عمومی فشرده‌سازی اطلاعات مانند pkzip و gzip می‌باشد. این الگوریتم بدین منظور طراحی شده که تعداد بیت‌هایی که به دیسک فرستاده می‌شود یا از دیسک خوانده می‌شود کمتر کند. همچنین از این الگوریتم در بسیاری از زمینه‌ها مانند برنامه‌های فشرده‌سازی GIF برای تصاویر استفاده می‌شود که به طور میانگین حجم تصویر را به یک سوم کاهش می‌دهد. الگوریتم LZW یک الگوریتم برگشت‌پذیر[2] است. بدین معنی که این الگوریتم هیچ اطلاعاتی را از دست نمی‌دهد و رمزگشا قادر خواهد بود اطلاعات اولیه را عیناً بازسازی نماید. قدرت فشرده‌سازی این الگوریتم از بسیاری از الگوریتم‌های فشرده‌سازی مشهور همچون هافمن بیشتر است. LZW از کلمه کدهایی با طول ثابت برای بیان کردن رشته‌ای از سمبول‌ها و کاراکترها با طول متغیر استفاده می‌کند[14,15]. در این الگوریتم رمزکننده و رمزگشا هنگام دریافت اطلاعات، واژه‌نامه[3] یکسانی را بصورت پویا تولید می‌کنند. از سایر روش‌های کدینگ که بیشتر در فشرده‌سازی منابع چندرسانه‌ای کاربرد دارند می‌توان کدینگ مبتنی بر پیش‌بینی و کدینگ تفاضلی را نام برد.

[1] Lempel–Ziv–Welch
[2] Reversible
[3] dictionary

۲-۲-۱-۳-۲- کدینگ مبتنی بر پیش‌بینی

در تصاویر مقدار پیکسل حاضر معمولاً تغییرات سریعی نسبت به پیکسل‌های مجاور ندارد. بنابراین می‌تواند از روی نمونه‌های قبلی پیش‌بینی شود. بنابراین، به جای کد کردن پیکسل‌های اصلی، می‌توان پیکسل حاضر را از روی قبلی‌ها پیش‌بینی کرد و تنها خطای پیش‌بینی را مشخص کرد. کدینگ پیش‌بینی کننده می‌تواند با تلفات یا بدون تلفات باشد. در کدینگ پیش‌بینی کننده با تلفات، خطای پیش‌بینی ابتدا کوانتیزه شده و سپس کد می‌شود. دلیل این کار این است که پیش‌بینی به شکلی که در ادامه خواهیم گفت به کاهش نرخ بیت کمک می‌کند. خطای پیش‌بینی توزیع غیریکنواخت خواهد داشت که عمدتاً در نزدیکی صفر متمرکز است و آنتروپی کمتری نسبت به نمونه‌های اصلی دارد که معمولاً توزیع یکنواخت دارند. با کدینگ آنتروپی مقادیر خطا می‌توانند با بیت‌های کمتری نسبت به مقادیر نمونه‌های تصویر اصلی کد شوند.

شکل (۲-۲) بلوک دیاگرام کلی برای کدگذار و کدگشای یک سیستم کدینگ با تلفات پیش‌بینی کننده را نشان می‌دهد. در یک کدکننده بدون تلفات، مرحله کوانتیزاسیون انجام نمی‌شود. شکل (۲-۳) چگونگی استفاده‌ی کدینگ پیش‌بینی کننده برای یک تصویر را نشان می‌دهد.



شکل (۲-۲)، بلوک دیاگرام کلی برای کدگذار و کدگشای یک سیستم کدینگ
با تلفات پیش‌بینی کننده [16]

A	B	C	D
E	F	G	H
I	J	K	L

$$\hat{f}_k = a\, f_F + b\, f_G + c\, f_H + d\, f_J$$

شکل (۲-۳)، کدینگ پیش‌بینی کننده [16]

در کدک‌های ویدیویی کلاسیک که بر مبنای کدینگ پیش‌بینی کننده بنا شده‌اند، برای استفاده از همبستگی زمانی بین فریم‌ها، عملیات جستجوی حرکت بر روی فریم قبلی یا بعدی انجام می‌گیرد. برای انجام این عمل، کدکننده باید توان محاسباتی بالایی داشته باشد. اما در طرف دیگر و در کدگشا عملیات دیکدینگ به نسبت ساده انجام می‌پذیرد. بنابراین این کدک‌ها در کاربردهایی مانند چندبخشی که در آن یک کدکننده و چندین کدگشا وجود دارد، به کار می‌روند. این الگوریتم‌ها به دلیل این ویژگی‌ها گزینه مناسبی برای مسیر رو به پایین در سیستم مورد بررسی می‌باشند. مسیر رو به پایین مسیر ارسال اطلاعات از سمت ایستگاه مرکزی به سمت گره‌هایی است که توانایی تصمیم‌گیری در مورد شرایط محیط را دارند. در طرف دیگر و در مسیر رو به بالا تعداد زیادی گره فرستنده و فقط یک گره گیرنده که همان ایستگاه مرکزی است، وجود دارد. گره‌های فرستنده یا همان کدکننده‌ها دارای محدودیت توان محاسباتی و انرژی هستند. از طرف دیگر کدگشا که در ایستگاه مرکزی قرار دارد، دچار این محدودیت‌ها نیست. با توجه به این شرایط، کدک‌های فعلی ویدیو را نمی‌توان در مسیر رو به بالا مورد استفاده قرار داد.

۲-۲-۲- فشرده‌سازی با اتلاف[1]

در کدینگ با اتلاف داده ورودی تا حدودی تغییر می‌کند و به نرخ فشرده‌سازی بالاتری دست می‌یابیم. فشرده‌سازی اتلافی شامل الگوریتمی می‌باشد که به جای اینکه داده‌ها را دقیقاً همان‌گونه که هست ارائه دهد، سعی در خلاصه سازی آنها دارد.

به این روش، کد کردن مفهومی نیز گفته می‌شود و در واقع عمل فشرده‌سازی با دور ریختن اطلاعات با اهمیت کمتر انجام می‌شود. الگوریتم‌های فشرده‌سازی از این دست بسیار پیچیده هستند و سعی بر این دارند که ضمن دستیابی به حداکثر فشرده‌سازی، کاهش کیفیت را به حداقل برسانند.

از جمله ویژگی‌های این گروه عبارت است از؛

- حذف اطلاعات کم اهمیت و یا بی‌اهمیت
- میزان فشرده‌سازی بالا
- فایل فشرده شده بطور کامل قابل بازیابی نبوده و ویرایش فایل بسیار مشکل است.

هنگامیکه از فشرده‌سازی همراه با اتلاف استفاده می‌شود، بخشی از اطلاعات از بین می‌رود. از این روش زمانی استفاده می‌گردد که دقت و صحت داده‌ها چندان ضروری نباشد. فشرده‌سازی همراه با اتلاف متداولترین روش فشرده‌سازی می‌باشد که از این نوع فشرده‌سازی در ارتباط با مستندات تصویری و اشیا صوتی و تصویری استفاده می‌شود.

انواع تبدیلات مورد استفاده در این فشرده‌سازی را می‌توان به شرح ذیل نام برد؛

- تبدیل کسینوس گسسته [2]
- تبدیل ویولت گسسته[1]

[1] Lossly Compression
[2] DCT

۲-۲-۲-۱- تبدیل کسینوس گسسته

DCT برای سیگنال‌های تصویر محبوب است زیرا به خوبی با خصوصیات آماری سیگنال‌های تصویر موجود منطبق است. بردارهای پایه یک بعدی N-نقطه‌ای DCT بدین صورت تعریف می‌شوند؛

$$h_k(n) = \propto (k)\ \cos\left(\frac{(2n+1)k\pi}{2N}\right), with\ \alpha(k) = \begin{cases} \sqrt{\frac{1}{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & k = 1,2,\dots,N-1 \end{cases} \quad (۲-۳)$$

دلیل مناسب بودن DCT برای فشرده‌سازی تصویر این است که یک بلوک تصویر معمولاً می‌تواند توسط ضرایب DCT بسیار فرکانس پایین نشان داده شود. این کار به این خاطر است که مقادیر شدت در هر تصویر معمولاً به صورت هموار تغییر می‌کنند و مولفه‌های فرکانس خیلی بالا تنها در نزدیکی لبه‌ها وجود دارند.

از جمله ویژگی‌های این تبدیلات عبارت است از؛

- فقط قسمت حقیقی دارد
- محاسبه آن ساده‌تر است
- در چند رسانه‌ای بیشتر از آن استفاده می‌شود.

چالش عملیات غیرفشرده‌سازی (کدگشایی[۲]) است که عملیاتی پیچیده و زمان‌بر است.

۲-۳- نرخ فشرده‌سازی

میزان فشرده‌سازی اطلاعات، به الگوریتم استفاده شده توسط برنامه فشرده‌سازی نیز بستگی دارد. بدیهی است استفاده از یک الگوریتم با کارآیی بالا، نتایج مثبتی را در رابطه با فشرده‌سازی به ارمغان

[1] DWT
[2] Decoding

خواهد آورد. میزان فشرده‌سازی بر اساس معیاری تحت عنوان نرخ فشرده‌سازی[1] ارزیابی می‌شود. نرخ فشرده‌سازی یکی از پارامترهای مهم در بررسی کارآیی روش‌های فشرده‌سازی می‌باشد. رایج‌ترین تعریف نرخ فشرده‌سازی بصورت زیر است[7].

$$Compression\ ratio = B_0/B_1 \quad (۴-۲)$$

که در آن B_0 تعداد بیت‌ها قبل فشرده‌سازی و B_1 تعداد بیت‌ها بعد از فشرده‌سازی است.

تعریف دیگری که بعنوان نرخ فشرده‌سازی بکار می‌رود، بیانگر میزان صرفه‌جویی در فضای لازم برای ذخیره داده ها می‌باشد و با فرمول زیر محاسبه می‌گردد[17].

$$Compression\ ratio\ \% = (1 - B_1/B_0) * 100 \quad (۵-۲)$$

۲-۴- مزایا و معایب فشرده‌سازی

۲-۴-۱- مزایا

- کاهش میزان فضای فیزیکی مورد نیاز جهت نگهداری اطلاعات و در نتیجه آن صرفه‌جویی در حجم منابع ذخیره‌سازی
- بهینه‌سازی پهنای باند با کاهش پهنای باند لازم جهت ارسال اطلاعات

۲-۴-۲- معایب

- در روش‌های با اتلاف بخشی از اطلاعات منبع از بین خواهد رفت.
- با افزایش پیچیدگی در روش‌های پیاده‌سازی، سربارهای محاسباتی در انجام پردازش جهت رمزگشایی اطلاعات بوجود خواهد آمد که گاهاً نیازمند سخت‌افزار مخصوص است.

[1] Compression ratio

۲-۵- کدگذاری

کامپیوترها فقط اعداد باینری[1] یا دودویی (۱و۰) را می فهمند، بنابراین تمام اطلاعات در هنگام ورود به کامپیوترها باید بصورت داده‌های دودویی نشان داده شوند. برای این منظور اطلاعات کدگذاری می-شوند. یعنی اطلاعات ورودی همچون حروف و یا علائم و ... بصورت یکسری اعداد صفر یا یک در می-آیند.

هر داده‌ای که در کامپیوتر وجود دارد حتی تصاویر یا فیلم‌ها، به صورت یک سری از ارقام در مبنای ۲ ذخیره می‌شود. این داده‌ها در اصل از رقم‌های صفر و یک تشکیل شده و به صورت خاصی ذخیره شده‌اند. ما می‌توانیم اطلاعاتی مانند یک تصویر را طبق الگوریتم‌هایی ذخیره کرده و آن را مشاهده یا ویرایش کنیم. تمامی حروف و اعدادی که نمایش داده می‌شوند نیز به همین شکل هستند. اختصاص دادن یک عدد خاص یا در نظر گرفتن قالب خاص برای کاراکترها را که به صورت استانداردهای جهانی در می‌آیند و (مانند استاندارد اسکی[2] و یا یونیکد[3]) در سیستم‌های کامپیوتری از آنها استفاده می‌شود، کدگذاری می‌گوییم.

Unicode و ASCII[4] هر دو از استانداردهایی هستند که برای رمزگذاری[5] متن‌ها استفاده می‌شوند. استفاده از این استانداردها در سراسر دنیا یک امر بسیار مهم است، در یک استاندارد یا Code ، هر سمبول[6] یا کاراکتر بدون توجه به نوع زبان برنامه‌نویسی که از آن استفاده می‌شود برای خودش یک عدد منحصر به فرد دریافت می‌کند. وقتی صحبت از استاندارد می‌شود یعنی چه برنامه‌نویس‌های شخصی و چه شرکت‌ها و سازمان‌های بزرگی که در زمینه برنامه‌نویسی فعالیت می‌کنند باید از این

[1] Binary
[2] ASCII
[3] Unicode
[4] American Standard Code for Information Interchange
[5] Encoding
[6] SymbolF

استانداردها تبعیت کنند و در اینجاست که دو استاندارد مهم ASCII و Unicode بسیار پرکاربرد هستند و بیشترین استفاده را در استانداردهای کدگذاری به خودشان اختصاص می‌دهند.

۲-۵-۱- کد ASCII

کد ASCII از اجداد پیشین یونیکد است. کدهای ASCII دارای محدودیت‌های فراوانی بودند. همان-طورکه عنوان شد هر کاراکتری که در کامپیوتر نشان داده می‌شود از یک سری صفر و یک (یا بیت) تشکیل شده است که مجموعه این بیت‌ها، یک کاراکتر را مشخص می‌کند و این صفر و یک‌ها را (که معمولاً در مبنای ۱۶ و یا احیاناً در مبنای ۱۰نمایش داده می‌شوند) کد کاراکتر گویند.

در سیستم کدگذاری ASCII برای نمایش کاراکترها از یک بایت استفاده می‌شود و لذا می‌توان نتیجه گرفت که با سیستم کدگذاری ASCII روی هم رفته می‌توان ۲ به توان ۸ یعنی ۲۵۶ کاراکتر را کدگذاری کرد، یعنی تنها یک قالب یک بایتی که شامل حروف A تا Z و a تا z ، اعداد، کاراکترهای کنترلی و برخی نمادها)کاراکترهایی برای نمادهای خاص مثل $ و @ و...(.

از توضیحات ارائه شده می‌توان به این نتیجه رسید، در سیستم ASCII جایی برای کاراکترهای دیگر زبان‌های دنیا مانند روسی، اسپانیایی، چینی، فارسی و غیره وجود ندارد و این کشورها برای نمایش کاراکترهای مربوط به خود از روش‌های متفاوتی مانند نگاشت (Mapping) و یا ماسک (Mask) که همیشه دارای مشکلات فراوانی بود، استفاده می‌کردند. کدگذاری ASCII از ۷ یا ۸ بیت استفاده می-کند و حروف را به لاتین نمایش می‌دهد و می‌تواند تا ۲۵۶ کاراکتر را نمایش دهد. این کدگذاری علائم ریاضی و علمی را پشتیبانی نمی‌کند.

۲-۵-۲- یونیکد

یونیکد یک استاندارد کدگذاری برای حروف الفباست، کاملترین استاندارد بین‌المللی موجود که نیازهای مربوط به تبادل اطلاعات چند زبانه را مرتفع نموده است. [18] یونیکد در بسیاری از

سیستم‌عامل‌ها، همه‌ی مرورگرهای اخیر، و بسیاری از محصولات دیگر پشتیبانی می‌شود. یونیکد مستقل از محیط، مستقل از برنامه، و مستقل از زبان به همه کاراکترها اعداد یکتایی اختصاص می-دهد. در این استاندارد، از کدگذاری ۱۶ بیتی استفاده می‌شود (برای هر کاراکتر از ۲ بایت استفاده می‌شود) که برای بیش از ۶۵۰۰۰ کاراکتر جا فراهم می‌کند.

به طور کلی، برخی از مشخصات یونیکد عبارت است از:

- یکسان‌سازی کاراکترهای مشترک در چند زبان مختلف؛ به عنوان مثال، به علت یکسانی «ع» در زبان عربی و فارسی، برای آن یک کد در نظر گرفته می‌شود.
- در استاندارد یونیکد، کاراکترهای فارسی در بلوک خط عربی قرار دارند. این بلوک برای در برگرفتن نویسه‌های زبان‌هایی که در آنها از خط عربی استفاده می‌شود مثل فارسی، اردو، پشتو، سندی، و کردی گسترش یافته است.
- در یونیکد، با وجود یکی‌سازی کدهای حروف مشترک، برای حروف فارسی که بار معنایی یا نمایشی متفاوت با حروف عربی دارند، کاراکترهای جداگانه در نظر گرفته شده است. یعنی کلیهٔ حروف خاص فارسی (پ، چ، ژ،گ) و نیز « ک » و « ی » فارسی که با حروف مشابه در عربی تفاوت نمایشی دارند، مکان جداگانه‌ای به خود اختصاص داده‌اند.
- یونیکد به هر کاراکتر مستقل از محیط، مستقل از برنامه، و مستقل از زبان یک عدد یکتا اختصاص می‌دهد.

از مهم‌ترین مزایایی که یونیکد برای زبان فارسی دارد، می‌توان موارد زیر را نام برد: [19]

- در نسخه استاندارد هر نرم‌افزاری که از این استاندارد پشتیبانی کند، می‌توان فارسی نوشت یا متون فارسی را خواند. و دیگر نیازی به تأمین نسخه‌های خاص فارسی یا عربی نیست.

- برای خواندن متون فارسی که توسط شرکت خاصی نوشته شده‌اند، نیازی به داشتن فونت خاص آن شرکت نداریم و هر متن فارسی که با استاندارد یونیکد، کدگذاری شده باشد، با هر فونت یونیکدی قابل مشاهده است.

با توجه به مزایای مطرح شده از استاندارد یونیکد در فشرده‌سازی متون فارسی استفاده می‌شود. بصورت کلی می‌توانیم تفاوت های ASCII و Unicode را به صورت زیر عنوان کنیم :

- ASCII اجازه استفاده از ۱۲۸ کاراکتر را می دهد اما Unicode تعداد کاراکتر بسیار زیادی را پشتیبانی می‌کند.
- ASCII و UTF-8[1] بیشترین استفاده را در پروتکل WWW دارند اما ASCII در حال جایگزین شدن با UTF-8 است.
- ترتیب کدگذاری ASCII با ترتیب حروف الفبایی است.
- ترتیب کدگذاری Unicode بر اساس اعداد منحصر به فرد است.
- ASCII بیشتر برای Encoding کاراکترهای زبان انگلیسی مورد استفاده قرار می‌گیرد.
- Unicode برای Encoding تقریباً تمامی کاراکترهای زبان‌های مختلف مورد استفاده قرار می-گیرد.
- Unicode برای نمایش از ساختار ۸ ، ۱۶ و ۳۲ بیتی استفاده می‌کند.
- ASCII برای نمایش از ساختار و فرمول ۷ بیتی استفاده می‌کند.
- در نهایت بیشترین تفاوت ASCII و Unicode در نمایش صفحات وب است.

۲-۶- مدل‌سازی زبان

زبان طبیعی[1]، به زبانی می‌گویند که بین انسان‌ها رایج است و انسان‌ها می‌توانند از آن برای ارتباط با یکدیگر به صورت‌های نوشتن، حرف زدن، خواندن و ... استفاده کنند. وقتی گفته می‌شود مدل‌سازی

[1] Universal Character Set Transformation Format 8 bit

زبان طبیعی، یعنی روابط و قواعد زبان را به طور هدفمند، یافته و ساده شود تا به ابزاری رسید که بتوان از آن برای بررسی و یا استفاده زبان و یا حتی تولید آن استفاده کرد.

مدل‌سازی زبان، تلاشی در جهت تسخیر قواعد طبیعی به منظور بهبود کارایی کاربردهای مختلف زبان طبیعی است. مدل‌های زبانی برای کاربردهای مختلفی از فن‌آوری زبان از جمله بازشناسی گفتار، ترجمه ماشینی، طبقه‌بندی متون، بازشناخت نوری کاراکترها، بازشناسی دست نوشته و تصحیح هجا و .. بکار گرفته شده‌اند. مدل‌های زبانی که به منظور بازشناسی گفتار و دیگر فناوری‌های زبانی به کار برده می‌شوند، برای اولین بار در سال ۱۹۸۰ مطرح شدند. از آن زمان تا کنون تلاش‌های فراوانی برای اصلاح و توسعه این مدل‌ها به جهت کاربرد در سیستم‌های پیشرفته امروزی صورت گرفته است. بسیاری از این مدل‌ها، از یک پس زمینه ریاضی برخوردار هستند، مانند گراف، احتمالات و... مدل N-gram یکی از این مدل‌هاست. مدل‌های آماری زبان توزیع احتمال واحدهای زبانی مختلفی مانند آواها، کلمات و جملات یک متن را محاسبه می‌نمایند.

۲-۷- مدل زبانی

مدل زبانی نحوهٔ رخداد توالی کلمات در زبان را مدل‌سازی می‌کند. مدل زبانی به دو گروه آماری و ساختاری تقسیم می‌شود.

مدل زبانی آماری به یک دنباله از کلمات زبان مانند $W = w_1 w_2 \dots w_m$ یک احتمال P(W) نسبت می‌دهد. و مدل زبانی ساختاری با استفاده از یک سری قواعد زبانی نحوه توالی لغات را مشخص می‌کند.

$$P(W) = P(w1, w2, w3, w4, w5 \dots wm) \qquad (۶-۲)$$

[1] Natural Language

یک مدل آماری زبان احتمال جملات را توصیف می‌کند. یعنی با استفاده از آن می‌توان با توجه به اطلاعات جاری، احتمال کلمات بعد را پیش‌بینی کرد. هر چند مدل‌های دیگری از جمله مدل‌های توانی[1] [20] یا گرامرهای مستقل از متن[2] [21] برای مدل‌سازی زبان‌های طبیعی پیشنهاد شده‌اند، اما در عمل مدل‌های آماری N-gram از سایر روش‌ها عملی‌تر و موثرتر بوده‌اند. زیرا علاوه بر سادگی پیاده‌سازی نسبت به نویز هم مقاوم هستند. [22]

از جمله کاربردهای مدل زبانی پیش‌بینی کلمات[3] ، بازشناسی گفتار[4]، درک زبان طبیعی[5]، ترجمهٔ ماشینی[6] و بازشناسی نویسه‌های نوری[7] را نام برد.

۲-۸- شمارش کلمات (قانون زیف[8])

یکی از مهمترین موارد استفاده آن در نظریه فشرده‌سازی است. در فشرده‌سازی متون، دانستن قاعده دقیق بسامد واژه‌ها بر طول رمزی که مورد استفاده قرار می‌گیرد، تاثیر دارد و با این آگاهی می‌توان به حد مطلوبی از فشردگی کلمات و عبارات دست یافت. زیف با مطالعهٔ فراوانی واژه‌هایی که در هر متن به کار می‌رود به مصادیقی برای اصل کمترین کوشش دست یافت. وی مشاهده کرد که بین طول واژه و تعداد دفعاتی که واژه‌ها در هر متن به کار می‌روند، رابطهٔ معکوس ثابتی وجود دارد. طبق قانون زیف چنانچه متنی، با هر طولی، برگزیده شود، فراوانی هر واژه موجود در داخل همان متن شمارش شود، این فراوانی‌ها از زیاد به کم مرتب شوند و رتبهٔ هر واژه در فراوانی همان معکوس فراوانی‌ها خواهد بود.

$$f(w) = \frac{C}{z(w)^a} \qquad (۷-۲)$$

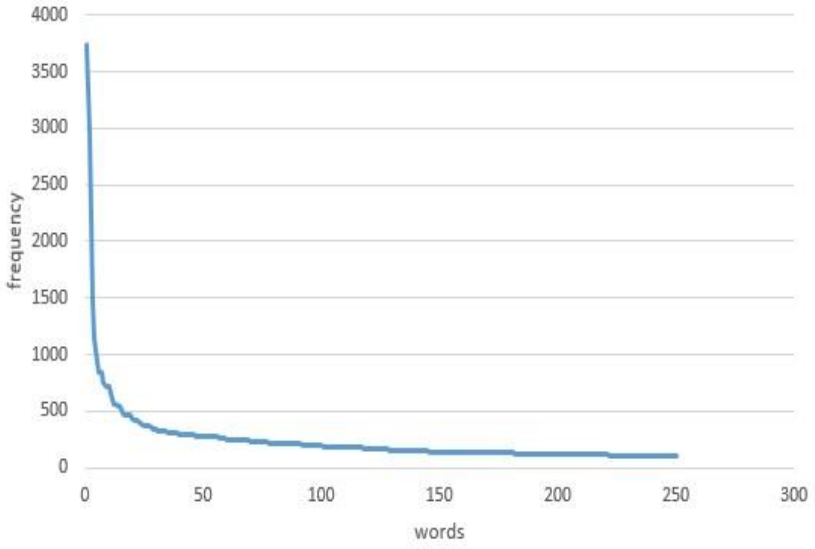
f(w): فراوانی کلمه w

[1] Exponential Models
[2] Context Free Grammars
[3] Prediction Words
[4] Speech Recognition
[5] Natural Language Perception
[6] Machine Translation
[7] Optical Character Recognition
[8] Zipf's Law

z(w): رتبه کلمه w

a و C: مقادیر ثابت (پارامترهای مدل)

فرکانس یا احتمال وقوع کلمات یک زبان بسیار متفاوت است. این احتمالات برای زبان‌های طبیعی مطابق با توزیع zipf برای ۲۵۰ کلمه اول پرکاربرد زبان فارسی با استفاده از متون آموزشی ما در نمودار زیر نشان داده شده است.



نمودار (۲-۱) ، نمودار فراوانی ۲۵۰ کلمه اول

با شمارش دفعات وقوع یک کلمه در مجموعه‌ی متون آموزشی و تقسیم آن بر کل تعداد کلمات می‌توان یک تخمین ML[1] برای احتمال وقوع هر کلمه بدست آورد. این ایده را می‌توان به زوج کلمات (و ترکیب‌های طولانی‌تر) تعمیم داد با این هدف که بعد از دیدن یک یا چند کلمه بتوان کلمه بعدی را حدس زد. در مدل‌های آماری زبان هدف پیش‌بینی کلمه بعدی یا به بیان دیگر محاسبه احتمال دنباله کلمات است.

در این پژوهش مدل زبانی یا به بیان دیگر مدل پیش‌بینی کلمه‌ای که برای متون فارسی استفاده کرده‌ایم یک مدل آماری پر استفاده به نام N-gram است که در آن احتمال کلمه N ام با استفاده از

[1] Maximum Likelihood

N-1 کلمه قبلی تخمین زده می‌شود. موفقیت این مدل‌ها اولین بار در آزمایشگاه‌های بازشناسی گفتار IBM به اثبات رسید و پس از آن در بسیاری از زمینه‌ها مورد توجه و استفاده محققین قرار گرفته است.

۲-۹- مدل N-gram

مدل N-gram یکی از این مدل‌های آماری زبان است. در این مدل، از آمار کلاسیک و احتمال بهره گرفته شده است. مدل‌سازی آماری زبان N-gram در حوزه‌های بسیاری مثل تشخیص گفتار، شناسایی زبان، ترجمه ماشینی، به رسمیت شناختن کاراکتر و طبقه‌بندی موضوع مورد استفاده قرار گرفته است.

فرض کنید که یکسری اشیاء یا نشانه‌ها و یا هر چیز دیگری داشته باشیم. هر کدام از این‌ها را به صورت یک راس در گراف تصور کنید که می‌تواند به راس دیگری یال جهت‌دار داشته باشد. این یال جهت‌دار، نشان دهنده یک نوع رابطه است که با توجه به مورد دلخواه ما می‌تواند معانی متفاوتی داشته باشد. مثلاً در مورد متن، می‌تواند توالی دو کاراکتر باشد (اگر کاراکتری بعد از کاراکتر دیگر بیاید، یک یال از اولی به دومی وجود دارد). [23] به یک توالی n تایی از این راس‌ها، N-gram می-گوییم (توالی 3-gram[1]، 2-gram[2] و ...). در این مدل، یک مجموعه داده‌های آماری بسیار بزرگ نیاز داریم که هر کدام مجموعه‌ای از این نشانه‌ها به همراه روابط بین آنهاست. برای مثال، در مورد یک زبان خاص، یک سری متن به آن زبان می‌باشد. حال، روابطی در این مدل تعریف می‌شود که می‌توان با استفاده از آن، درستی یک توالی خاص از این نشانه‌ها را بررسی کرد [24,25].

در رشته‌ها و زمینه‌های محاسباتی، زبان‌شناسی و احتمالات، یک n گرم به عنوان یک زنجیره‌ی متصل و مربوط به هم از آیتم‌های n از زنجیره معلوم و مفروضی از متن و گفتار می‌باشد. آیتم‌های مورد بحث

[1] Trigram
[2] Bigram

می‌توانند آواها، هجاها، حروف الفبا، کلمات یا جفت‌های مبنایی مورد استفاده و کاربرد باشند. N گرم-ها از یک متن با مجموعه‌ای از کاراکترها جمع‌آوری می‌شوند. یک n گرم دارای سایز و اندازه‌ی ۱، تحت عنوان یونی‌گرم، با اندازه و سایز ۲ بایگرام (یا دیگرم)، با اندازه و سایز ۳ تری گرم، با اندازه و سایز ۴ فورگرام و اندازه ۵ یا بیشتر به راحتی n گرم نامیده می‌شود.

در واقع مدل N-gram ساده‌ترین و پرکاربردترین مدل زبانی آماری است که احتمال رخداد پس از دنباله‌ای از n-1 کلمه را محاسبه می‌نماید.

در حالت کلی احتمال دنباله لغات $W = w_1 w_2 \dots w_m$ برابر است با؛

$$P(W) = P(w_1 w_2 \dots w_m) = \prod_{i=1}^{m} P(w_i \mid w_1 \dots w_{i-1})$$

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \dots P(w_m|w_1 \dots w_{m-1}) \qquad (۸-۲)$$

در این فرمول در صورتی‌که بزرگ باشد، محاسبهٔ احتمال فوق بسیار مشکل و در عمل غیرممکن است.

۲-۱۰- ویژگی مدل‌های N-gram

از ویژگی‌های قابل انتظار مدل‌های N-gram این است که دقت (کارایی) مدل با افزایش مقدار N افزایش می‌یابد. علیرغم این واقعیت، عملا در بیشتر کاربردها از مدلهای Bigram یا حداکثر Trigram استفاده می‌شود؛ زیرا مدل‌های مرتبه بالاتر از ۳ برای آموزشِ مناسب، احتیاج به مجموعه متون بسیار بزرگی دارند و در غیر این صورت نمی‌توان تخمین‌های مناسبی برای احتمالات به دست آورد. علاوه بر این، مدل‌های مرتبه بالاتر از ۳ بسیار بزرگ هستند و ذخیره و استفاده از آنها احتیاج به حافظه زیادی دارد. مرتبه حافظه مورد نیاز یک مدل N-gram برابر است با تعداد ترکیب‌های مختلف N کلمه‌ای که دارای حد بالای V^N است.

ویژگی دیگر مدل‌های N-gram وابستگی زیاد آنها به متون (به خصوص نوع و اندازه) آموزشی است. یکی از روش‌های معمول برای مشاهده عملکرد کیفی یک مدل N-gram، تولید رشته‌های تصادفی کلمات با استفاده از مدل است . به این ترتیب که اولین کلمه بر اساس احتمال Unigram انتخاب می-شود و سپس دومین کلمه با مدل Bigram و پس از آن کلمات بعدی می‌توانند با مدل Bigram یا Trigram تولید شوند.

نظر به اینکه احتمالات در یک مدل آماری همچون N-gram از یک مجموعه آموزشی استخراج می-شوند، این مجموعه را باید با دقت انتخاب کرد. اگر مجموعه آموزشی تنها مربوط به یک زمینه (موضوع) خاص باشد، احتمالات نتیجه شده نمی‌توانند برای جملات جدید به شکلی مناسب تعمیم یابند. از طرفی اگر مجموعه متون آموزشی بسیار عمومی و کلی باشد، ممکن است احتمالات نتیجه شده برای آن کاربرد خاص مناسب نباشند.

برای آموزش و سپس محاسبه کارایی یک مدل، همانند سایر مسائل یادگیری محاسباتی، مجموعه متون موجود را به دو مجموعه مجز ای آموزشی و آزمایشی تقسیم می‌کنیم. مدل را بر اساس مجموعه آموزشی می‌سازیم و سپس کارایی آن را با استفاده از معیاری به نام سرگشتگی[1]، - که معیاری مشابه با آنتروپی است- روی مجموعه آزمایشی محاسبه می‌کنیم. البته در برخی موارد به بیش از یک مجموعه آزمایشی احتیاج داریم. مثلا فرض کنید که چند مدل مختلف را در اختیار داریم و هدف این است که بهترین را انتخاب و سپس کارایی آن را محاسبه کنیم. البته باید توجه کرد که برای مقایسه کارایی این مدل‌ها لازم است از تست‌های آماری استفاده شود تا معلوم شود تفاوت بین دو مدل تا چه حد قابل توجه است.

۲-۱۱- هموارسازی

[1] Perplexity

مساله اصلی مدل‌های استاندارد N-gram که مورد بررسی قرار می‌گیرند این است که تخمین‌های احتمال آنها با استفاده از یک مجموعه متون آموزشی محدود بدست می‌آیند که در هر صورت نمی-توانند شامل تمامی ترکیب‌های ممکن زبان باشند و این امر ماتریس‌های خروجی بدست آمده از مدل-های N-gram را به ماتریس‌های خلوت تبدیل می‌نماید. در این ماتریس بسیاری از دنباله‌های مجاز و با معنی زبان وجود دارند که می‌بایست احتمال غیرصفر داشته باشد. حتی در صورتی‌که مشکل مقدار صفر وجود نداشته، وجود مقدار فراوانی‌های بسیار کوچک باز هم تخمین‌های بسیار ضعیفی را نتیجه خواهد داد. برای حل این مشکلات از تکنیک‌های هموارسازی[1] استفاده می‌شود که در این تکنیک‌ها به احتمالات صفر (و پایین) مقدار غیرصفر (و بیشتری) نسبت داده می‌شود. در اجرای این تکنیک به همین نسبت از احتمالات زیاد کاسته می‌شود (تا جمع احتمالات یک باقی بماند). تکنیک هموارسازی به رفع مشکل احتمال‌های صفر با تخمین احتمال برای رخدادهای دیده نشده کمک می‌کند. روش-های هموارسازی سعی می‌کنند احتمال رخدادهای دیده نشده را به نحوی تخمین بزنند.

۲-۱۱-۱- هموارسازی جمع با یک

اولین و ساده‌ترین روشی که برای هموارسازی روش جمع با یک است که مفاهیم کلی و اساسی هموارسازی را در بر دارد و بررسی آن به درک و توصیف الگوریتم‌های پیچیده‌تر کمک می‌کند.

$$P_{Add-1}(w_n|w_{n-N+1}^{n-1}) = \frac{c(w_{n-N+1}^{n-1}w_n) + 1}{c(w_{n-N+1}^{n-1}) + V} \qquad (۹-۲)$$

۲-۱۱-۲- هموارسازی تخفیف[2]

در این روش به منظور هموارسازی، از شمارش‌های غیرصفر کاسته شده و بر روی شمارش‌های صفر توزیع می‌گردد. کاستن از شماره‌های غیرصفر با استفاده از روش‌های تخفیف صورت می‌گیرد.

(۱۰-۲)

[1] Smoothing
[2] Discounting

$$r^* = r.d_r$$

در این محاسبه r شمارش اولیه ، r^* شمارش تخفیف داده شده و d_r ضریب تخفیف می‌باشد.

از مهم‌ترین روش‌های تخفیف می‌توان Good-Turing ، Interpolation Absolute Discounting و Kneser-Ney را نام برد.

۲-۱۱-۳- هموارسازی عقب‌گرد[1]

در این روش از هموارسازی در صورتی‌که برای یک N-gram مقداری وجود نداشته باشد (یا مقدار آن به اندازه کافی معتبر نباشد)، سراغ N-gram هایی با درجات پایین‌تر می‌رویم.

$$\hat{P}(w_i \mid w_{i-1}) = \begin{cases} \alpha(w_{i-1})\hat{P}(w_i) & \text{if } C(w_{i-1}w_i) = 0 \\ d_{C(w_{i-1}w_i)}.P(w_i \mid w_{i-1}) & \text{if } 1 \le C(w_{i-1}w_i) \le k \\ P(w_i \mid w_{i-1}) & \text{if } C(w_{i-1}w_i) > k \end{cases}$$

$$\alpha(w_{i-1}) = \frac{\hat{\beta}(w_{i-1})}{\sum_{w_i : N(w_{i-1}w_i)=0} \hat{P}(w_i)} \qquad (۱۱-۲)$$

[1] Back Off

فصل سوم
روش اجرای تحقیق

۳-۱- کلیات

فشرده‌سازی در بازیابی اطلاعات و در صنعت صفحات فشرده و نظایر آن موجب صرفه‌جویی در فضا و منابع مالی است. لذا این امر در فناوری چندرسانه‌ای از عوامل اصلی محسوب می‌شود. در روش‌های متداول فشرده‌سازی اطلاعات می‌بایست، هر کاراکتر به معادل باینری آن تبدیل شود. حال اگر بخواهیم این تبدیل را بر اساس کد اسکی انجام دهیم هر کاراکتر به ۸ بیت و اگر بخواهیم بر اساس استاندارد یونیکد انجام دهیم هر کاراکتر به ۱۶ بیت فضا نیاز خواهد داشت.

در این پژوهش از استاندارد یونیکد استفاده نموده و فشرده‌سازی متن را با توجه به فرآیند کاهش تعداد بیت و اعمال قوانین مدلسازی زبان خواهیم داشت. اما با توجه تنوع کدگذاری‌های موجود برای متون فارسی ابتدا عمل پیش‌پردازش را بر روی داده‌های آموزشی انجام خواهیم داد.

۳-۲- پیش‌پردازش فایل‌های متنی

فایل‌هایی متنی آموزشی جمع‌آوری شده از اینترنت، به فرمت HTML هستند. در این فایل‌ها، کاراکترهای فارسی با سیستم‌های کدگذاری متفاوتی (UTF8 , Unicode, Cp1256) نشان داده می‌

شوند. در اولین مرحله پیش‌پردازش را انجام می‌دهیم. طی این عمل، فایل ورودی، با هر سیستم کدگذاری را به یونی‌کد تبدیل می‌کنیم . برای این کار یک برنامه جاوا نوشته شده است.

گاهی ممکن است کلمات یکسان با مجموعه‌ای از کاراکترهای غیر یکسان نمایش داده شوند. به عنوان مثال کلمه «یک» می‌تواند چهار شکل نمایش مختلف داشته باشد، که از ترکیب دو کاراکتر «ی» و «ک» فارسی و عربی به دست می‌آیند. بنابراین، در اولین مرحله پیش‌پرازش جهت نرمال‌سازی، هر حرفی را که در استاندارد یونی‌کد با دو یا چند کاراکتر مختلف نمایش داده می‌شود، فقط با یکی از اشکال فارسی نشان می‌دهیم.

مرحله بعدی، حذف اعراب کلمات است، با وجود اینکه گاهی تفاوت دو کلمه فقط با اعراب مشخص می‌شود، اما چون در اکثریت متونی که از اینترنت جمع‌آوری شده‌اند، کلمات اعراب‌گذاری نشده‌اند، مجبور هستیم اعراب کلمات را به طور کلی حذف کنیم. به بیان دیگر، کلمات یا باید همواره دارای اعراب باشند، یا همواره دارای اعراب نباشند و چون کلمات موجود در متون آموزشی همواره دارای اعراب نیستند، پس تنها راه برای نرمال‌سازی این است که کلمات هیچ اعرابی نداشته باشند.

۳-۳- مکانیزم فشرده‌سازی اطلاعات (فشرده‌سازی با استفاده از مدل-های زبانی)

در این پژوهش، برای دستیابی به فشرده‌سازی متن، از مدل زبانی[1] استفاده می‌شود. به طور کلی مدل زبانی، بیانگر قواعد و روابط حاکم بر زبان است. [26]. در سال ۲۰۰۷ روشی توسط کتس [27] جهت تشخیص سیگنال‌های مخدوش ارائه شده است. بر اساس این روش برای یافتن قواعد و روابط زبان، بدون نیاز به تحلیل سیگنال صوتی مخدوش شده، می توان کلمه صحیح را با احتمال بسیار بالا تشخیص داد. روش کار بدین صورت است که

[1] Language model

با تحلیل حجم بسیار بالایی از سیگنال‌های صحیح مشخص می‌شود که بعنوان مثال در اکثر موارد بعد از کلمه "بریم" ، از کلمه "خانه" استفاده شده است. حال اگر سیگنال مفقود شده بعد از کلمه "بریم" باشد، به احتمال زیاد کلمه صحیح "خانه" بوده است. می‌توان همین عمل را برای تمام کلمات موجود در متن انجام داد، و نتایج بدست آمده را در بانک-اطلاعاتی قرار داد. در این روش با از بین رفتن یک کلمه در یک پیام صوتی می‌توان احتمالات بسیار نزدیک به واقعیت را در نظر گرفت.

با توجه مطالب عنوان شده، ما در تکنیک فشرده‌سازی مورد استفاده در پژوهش پیش رو، در مرحله اول احتمال قرار گرفتن هر یک ار کلمات را پس از سایر کلمات و در مرحله دوم احتمال قرار گرفتن هر یک از کاراکترهای زبان، بعد از یک کاراکتر دیگر را بررسی می‌کنیم. با این کار مشخص می شود که بعد از یک کلمه یا کاراکتر خاص چه کلمات یا کاراکترهایی بیشتر احتمال قرار گرفتن دارند.

۳-۳-۱- مرحله اول: اعمال مدل‌های زبانی کلمات

در این مرحله از مدل‌های زبانی یونیگرام، بایگرام، ترایگرام و فورگرام استفاده می‌کنیم. پایگاه دادگان مورد استفاده جهت انجام بررسی‌های لازم، متون فارسی منابع ذکر شده در جدول (۴-۱)، می‌باشد. آموزش‌ها در مدل زبانی یونیگرام بصورت یک کلمه‌ای، در مدل زبانی بایگرام به صورت دوتایی، در مدل زبانی ترایگرام به صورت سه تایی و در مدل زبانی فورگرام به صورت چهار تایی انجام می‌شود.

۳-۳-۱-۱- مدل زبانی یونیگرام

در مدل یونیگرام با تحلیل متون مختلف، کلماتی که بیشترین کاربرد در زبان فارسی را دارند مشخص می شوند. در این مدل زبانی (یونیگرام) متون آموزشی بصورت کلمه به کلمه،

بررسی می‌شوند و تعداد تکرار برای هر کلمه در متون آموزشی داده شده بدست آمده و در یک جدول به نام "فراوانی توالی یک کلمه‌ای" ذخیره می‌شود. مدل یونیگرام ساده‌ترین حالت در مدل‌های N-gram بوده که وابستگی با کلمات قبل را در نظر نمی‌گیرد.

$$P(W) = P(w_1 w_2 \dots w_m)$$
$$\cong \prod_{i=1}^{m} P\left(w_i \mid w_{i-(k)} \dots w_{i-1}\right)$$
$$\cong \prod_{i=1}^{m} P(w_i) = P(w_1)\, P(w_2) P(w_3) \dots P(w_m)$$

$$P_{monogram}(w_i) = \frac{count\ (w_i)}{count\ (All\ Words)} = \frac{N(w_i)}{N(total)} \qquad (۱-۳)$$

تعدادی از فراوانی‌های بدست آمده برای مدل زبانی یونیگرام کلمات در این مرحله، به عنوان نمونه در جدول (۳-۱) نشان داده شده است.

جدول (۳ -۱) ، فراوانی توالی یک کلمه‌ای ۲۰ مورد اول

ردیف	دنباله یک کلمه‌ای	تعداد تکرار
۱۱	خود	۶۵۴
۱۲	باید	۵۶۸
۱۳	می‌شود	۵۶۵
۱۴	حضرت	۵۵۴
۱۵	نیز	۵۴۵
۱۶	اقتصادی	۴۸۰
۱۷	مردم	۴۷۱
۱۸	اما	۴۶۷
۱۹	گزارش	۴۶۱
۲۰	دولت	۴۴۲

ردیف	دنباله یک کلمه‌ای	تعداد تکرار
۱	این	۳۷۴۲
۲	است	۲۹۶۰
۳	برای	۱۴۶۹
۴	سال	۱۱۴۴
۵	کرد	۹۳۷
۶	ایران	۸۵۲
۷	شده	۸۳۹
۸	کشور	۷۵۶
۹	گفت	۷۲۴
۱۰	بود	۷۱۸

۳-۳-۱-۲- مدل زبانی بایگرام

مرحلـه بعـد از یـونیگرام، بـایگرام اسـت. در ایـن مرحلـه بـا بررسـی متـون فارسـی، احتمـال قـرار گرفتن هر یک از کلمـات، پـس از دیگـری، (دو کلمـه متـوالی) محاسـبه مـی‌شـود. ایـن مـدل یـک کلمه قبل از کلمه مورد بررسی را در نظر می‌گیرد.

$$P(w_1 w_2 \dots w_m)$$

$$\cong \prod_{i=1}^{m} P\left(w_i \mid w_{i-(k)} \dots w_{i-1}\right)$$

$$\cong \prod_{i=1}^{m} P(w_i|w_{i-1}) = P(w_1)\, P(w_2|w_1) P(w_3|w_2) \dots P(w_m|w_{m-1})$$

$$P(w_i|w_1 w_2 \dots w_{i-1}) \approx P(w_i|w_{i-1})$$

$$P_{bigram}(w_j/w_i) = \frac{N(w_i w_j)}{N(w_i)} \qquad (۲-۳)$$

در مدل زبانی دوم (بـایگرام) متـون آموزشـی بصـورت دو کلمـه دو کلمـه مـورد بررسـی قـرار مـی-گیرنـد. در ایـن حالـت تعـداد تکـرار بـرای هـر دو کلمـه متـوالی در جـدول "فراوانـی تـوالی دو کلمه‌ای" ذخیـره مـی‌شـود. در جـدول زیـر نمونـه‌ای از خروجـی‌هـای بدسـت آمـده از ایـن مرحلـه نشان داده شده است.

جدول (۳ -۲) ، فراوانی توالی دو کلمه‌ای ۲۰ مورد اول

ردیف	دنباله دو کلمه‌ای	تعداد تکرار
۱	است که	۶۷۱
۲	در این	۵۹۰
۳	که در	۴۹۰
۴	و در	۳۹۲
۵	را به	۳۷۷
۶	است و	۳۵۱
۷	را در	۲۹۳
۸	از این	۲۷۸
۹	در سال	۲۴۸
۱۰	اشاره به	۲۴۳

ردیف	دنباله دو کلمه‌ای	تعداد تکرار
۱۱	با اشاره	۲۳۹
۱۲	خود را	۲۳۷
۱۳	و به	۲۳۰
۱۴	یکی از	۲۲۹
۱۵	که به	۲۱۳
۱۶	زهرا (س)	۱۹۸
۱۷	به این	۱۹۳
۱۸	که از	۱۹۳
۱۹	و با	۱۹۲
۲۰	و از	۱۸۱

۳-۳-۱-۳- مدل زبانی ترایگرام[1]

مرحله سوم تحلیل بر اساس مدل زبانی[2]، ترایگرام می‌باشد. در این مرحله احتمال دنباله‌های سه تایی از کلمات مورد بررسی قرار می‌گیرد. در این مدل دو کلمه قبل از هر کلمه در نظر گرفته می‌شود.

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_2w_3) \dots P(w_m|w_{m-2}w_{m-1})$$

$$P_{trigram}(w_k/w_i\, w_j) = \frac{N(w_iw_jw_k)}{N(w_iw_j)} \qquad (۳-۳)$$

در این مدل محاسبات لازم برای مدل زبانی ترایگرام به صورت سه کلمه سه کلمه صورت می‌پذیرد. خروجی بدست آمده در جدولی به نام "فراوانی توالی سه کلمه‌ای" ذخیره می‌شود. نمونه‌ای از اعداد بدست آمده در جدول فراوانی توالی سه کلمه‌ای به صورت زیر می‌باشد.

جدول (۳ -۳) ، فراوانی توالی سه کلمه‌ای ۲۰ مورد اول

[1] Trigram
[2] Language model

ردیف	دنباله سه کلمه‌ای	تعداد تکرار
۱۱	این است که	۷۷
۱۲	در حال حاضر	۶۸
۱۳	اقتصادی باشگاه خبرنگاران	۶۲
۱۴	اشاره به اینکه	۶۰
۱۵	دانشجویان ایران ایسنا	۵۹
۱۶	رهبر معظم انقلاب	۵۶
۱۷	مجلس شورای اسلامی	۵۵
۱۸	که در این	۵۴
۱۹	تحقق شعار سال	۵۰
۲۰	در این زمینه	۴۸

ردیف	دنباله سه کلمه‌ای	تعداد تکرار
۱	با اشاره به	۲۳۷
۲	در گفتگو با	۱۶۰
۳	حضرت زهرا (س)	۱۴۹
۴	مقام معظم رهبری	۱۳۶
۵	با توجه به	۱۲۱
۶	خبرگزاری دانشجویان ایران	۱۰۴
۷	با بیان اینکه	۱۰۲
۸	خبرنگار اقتصادی باشگاه	۹۸
۹	گفتگو با خبرنگار	۹۸
۱۰	است که در	۸۲

۳-۳-۱-۴- مدل زبانی فورگرام

در مرحله چهارم تحلیل بر اساس مدل زبانی[1] (فورگرام)، احتمال دنباله‌های چهار تایی از کلمات مورد بررسی قرار می‌گیرد. در این مدل سه کلمه قبل از هر کلمه در نظر گرفته می‌شود.

$$P_{forgram}(w_l/w_i\, w_j w_k) = \frac{N(w_i w_j w_k w_l)}{N(w_i w_j w_k)} \qquad (۴-۳)$$

محاسبات لازم برای مدل زبانی فورگرام به صورت چهار کلمه چهار کلمه صورت می‌پذیرد و خروجی بدست آمده در جدولی به نام "فراوانی توالی چهار کلمه‌ای" ذخیره می‌شود. به عنوان نمونه جدول "فراوانی توالی چهار کلمه‌ای" برای مدل زبانی فورگرام در جدول (۳-۴) نشان داده شده است.

جدول (۳ -۴) ، فراوانی توالی چهار کلمه‌ای ۲۰ مورد اول

[1] Language model

ردیف	دنباله چهار کلمه‌ای	تعداد تکرار
۱۱	گزارش گروه دریافت خبر	۳۶
۱۲	مواد سوختی و فاسد	۳۴
۱۳	گزارش گروه اقتصادی باشگاه	۳۴
۱۴	به گزارش گروه اقتصادی	۳۳
۱۵	حاملین مواد سوختی و	۳۱
۱۶	حجت الاسلام و المسلمین	۳۱
۱۷	مقام معظم رهبری در	۲۹
۱۸	اقتصادی باشگاه خبرنگاران به	۲۸
۱۹	انواع تریلر و کامیون	۲۸
۲۰	باشگاه خبرنگاران به نقل	۲۸

ردیف	دنباله چهار کلمه‌ای	تعداد تکرار
۱	در گفتگو با خبرنگار	۹۸
۲	با اشاره به اینکه	۶۰
۳	خبرگزاری دانشجویان ایران ایسنا	۵۹
۴	خبرنگار اقتصادی باشگاه خبرنگاران	۵۱
۵	گزارش خبرنگار اقتصادی باشگاه	۵۰
۶	با خبرنگار اقتصادی باشگاه	۴۶
۷	گزارش خبرگزاری دانشجویان ایران	۴۶
۸	گفتگو با خبرنگار اقتصادی	۴۶
۹	به گزارش خبرنگار اقتصادی	۳۹
۱۰	گزارش خبرگزاری خبر آنلاین	۳۹

امکان توسعه این مدل‌ها با تعداد بیشتر نیز وجود دارد و و مقدار معمول آن بین 2-gram تا 5-gram می‌باشد.

۳-۳-۲- مرحله دوم: اعمال مدل‌های زبانی کاراکترها

در مرحله دوم، متون آموزشی برای بدست آوردن فراوانی‌های دنباله‌های چهار، سه، دو و یک کاراکتری مورد بررسی قرار گرفته و نتایج بدست آمده به ترتیب در جداولی با نام‌های "فراوانی توالی چهار کاراکتری"، "فراوانی توالی سه کاراکتری"، "فراوانی توالی دو کاراکتری" و "فراوانی توالی یک کاراکتری" ذخیره می‌شوند.

۳-۳-۲-۱- مدل زبانی یونیگرام

با تحلیل متون آموزشی بر اساس این مدل، کاراکترهایی که بیشتر از سایر موارد در زبان فارسی کاربرد دارند مشخص می‌شوند. فراوانی‌های بدست آمده در جدولی به نام "فراوانی توالی یک کاراکتری" ذخیره می‌شوند. جدول (۳-۵) فراوانی تعداد تکرار این کاراکترها در متون آزمایشی را نشان می‌دهد.

جدول (۳-۵)، تعداد تکرار کاراکترها در متون آزمایشی

ردیف	کاراکتر	تعداد تکرار	احتمال(٪)
۱	space	۲۳۱۲۳۶	۱۹.۵۲
۲	ا	۱۳۷۱۸۶	۱۱.۵۸
۳	ی	۹۰۴۴۹	۷.۶۳
۴	ر	۸۱۸۲۳	۶.۹۱
۵	د	۶۴۰۴۲	۵.۴۱
۶	ن	۶۲۳۳۹	۵.۲۶
۷	ه	۵۶۲۷۲	۴.۷۵
۸	م	۵۴۴۷۳	۴.۶
۹	و	۵۳۶۷۵	۴.۵۳
۱۰	ت	۴۵۵۷۶	۳.۸۵
۱۱	ب	۴۰۲۵۳	۳.۴
۱۲	س	۲۹۲۹۳	۲.۴۷
ردیف	**کاراکتر**	**تعداد تکرار**	**احتمال(٪)**
۱۳	ل	۲۵۸۷۳	۲.۱۸
۱۴	ک	۲۳۹۲۲	۲.۰۲
۱۵	ش	۲۳۴۶۶	۱.۹۸
۱۶	ز	۲۰۹۶۰	۱.۷۷
۱۷	ف	۱۳۹۰۹	۱.۱۷
۱۸	گ	۱۳۰۱۸	۱.۱
۱۹	ع	۱۱۳۶۸	۰.۹۶
۲۰	خ	۱۰۷۵۳	۰.۹۱
۲۱	ق	۱۰۳۱۲	۰.۸۷
۲۲	ج	۹۷۹۶	۰.۸۳
۲۳	ح	۹۰۸۳	۰.۷۷
۲۴	.	۷۵۷۶	۰.۶۴
۲۵	،	۶۸۲۴	۰.۵۸
۲۶	پ	۵۶۸۵	۰.۴۸
۲۷	ص	۵۶۰۲	۰.۴۷
۲۸	آ	۵۵۷۹	۰.۴۷

۲۹	ط	۴۷۲۰	۰.۴
۳۰	چ	۲۷۹۵	۰.۲۴
۳۱	:	۲۶۳۳	۰.۲۲
۳۲	ض	۲۵۸۶	۰.۲۲
۳۳	ذ	۱۹۴۱	۰.۱۶
۳۴	ظ	۱۸۱۴	۰.۱۵
۳۵	غ	۱۶۵۸	۰.۱۴
۳۶	ث	۱۳۶۸	۰.۱۲
۳۷	ئ	۱۲۵۴	۰.۱۱
۳۸	۱	۱۱۹۸	۰.۱
۳۹	۰	۹۴۰	۰.۰۸
۴۰	۲	۸۵۶	۰.۰۷
۴۱	ژ	۸۰۸	۰.۰۷
۴۲	۳	۷۵۴	۰.۰۶
۴۳	-	۷۱۵	۰.۰۶
۴۴	»	۶۹۱	۰.۰۶
۴۵	«	۶۹۱	۰.۰۶
۴۶	(	۶۵۹	۰.۰۶
۴۷	)	۶۵۹	۰.۰۶
۴۸	۹	۶۱۹	۰.۰۵
۴۹	۵	۶۰۳	۰.۰۵
۵۰	۴	۵۸۳	۰.۰۵
۵۱	“	۵۳۲	۰.۰۴
۵۲	؛	۳۷۳	۰.۰۳
۵۳	۶	۳۵۶	۰.۰۳
۵۴	۸	۳۲۳	۰.۰۳
۵۵	۷	۲۹۱	۰.۰۲
۵۶	?	۲۲۸	۰.۰۲
۵۷	/	۱۸۸	۰.۰۲
۵۸	❊	۱۶۰	۰.۰۱
۵۹	+	۸۸	۰.۰۱

۶۲	]	۶۱	۰.۰۱
۶۳	٪	۶	۰

۶۰	!	۷۱	۰.۰۱
۶۱	[	۶۱	۰.۰۱

۳-۳-۲-۲- مدل زبانی بایگرام

در این مرحله با بررسی متون فارسی، احتمال قرار گرفتن هر یک از کاراکترهای زبان، بعد از یک کاراکتر خاص، محاسبه می‌شود. به عنوان مثال با بررسی‌های صورت گرفته، نتایج جدول (۳-۶) ، برای کاراکتر "ا" بدست آمده است.

در جدول (۳-۶) ، منظور از ستون "کاراکتر"، کاراکتری است که مورد بررسی قرار گرفته است. همچنین ستون " تعداد تکرار" ، تعداد دفعات تکرار کاراکتر مورد بررسی، بعد از کاراکتر "ا" و ستون "احتمال" ، احتمال رخداد این کاراکتر ، بعد از کاراکتر "ا"، در متون آموزشی را نشان می‌دهد.

از جدول (۳-۶) اینگونه برداشت می‌شود که در داده‌های آموزشی ما احتمال آمدن کاراکتر "ر" بعد از کاراکتر "ا" ۱۳.۵۷٪ می باشد (این احتمال با تقسیم تعداد زوج کاراکتر "ار" بر تعداد کل دوتایی‌هایی که با کاراکتر "ا" شروع شده، بدست آمده است). همچنین می‌توان گفت بر اساس نتایج بدست آمده احتمال آمدن کاراکتر "ب" بعد از کاراکتر "ا" ۹.۴۲٪ می‌باشد، که این امر بیانگر آن است که احتمال رخداد کاراکتر" ر" بعد از کاراکتر " ا "، بیشتر است.

جدول (۳-۶)، نتایج مرحله بایگرام در ارزیابی کاراکتر "ا"

ردیف	کاراکتر	تعداد تکرار	احتمال (٪)
۱	آ	۰	۰
۲	ا	۱۸	۰.۰۲
۳	ب	۹۸۵۲	۹.۴۲
۴	پ	۱۲۰۸	۱.۱۶
۵	ت	۴۰۶۹	۳.۸۹

۶	ث	۲۲۷	۰.۲۲
۷	ج	۱۹۹۸	۱.۹۱
۸	چ	۲۶۱	۰.۲۵
۹	ح	۱۳۵۸	۱.۳۰
۱۰	خ	۱۶۰۳	۱.۵۳
۱۱	د	۸۵۸۸	۸.۲۱
۱۲	ذ	۸۶۱	۰.۸۲

۱۳	ر	۱۴۱۸۹	۱۳.۵۷
۱۴	ز	۳۰۵۷	۲.۹۲
۱۵	ژ	۱۲۹	۰.۱۲
۱۶	س	۵۲۰۱	۴.۹۸
۱۷	ش	۲۲۰۵	۲.۱۱
ردیف	**کاراکتر**	**تعداد تکرار**	**احتمال (٪)**
۱۸	ص	۹۹۳	۰.۹۵
۱۹	ض	۵۷۴	۰.۵۵
۲۰	ط	۳۲۶	۰.۳۱
۲۱	ظ	۳۹۵	۰.۳۸
۲۲	ع	۱۷۰۱	۱.۶۳

۲۳	غ	۳۳۲	۰.۳۲
۲۴	ف	۱۵۹۳	۱.۵۲
۲۵	ق	۲۰۸۵	۱.۹۹
۲۶	ک	۲۹۵۶	۲.۸۳
۲۷	گ	۲۲۶۹	۲.۱۷
۲۸	ل	۴۳۷۳	۴.۱۸
۲۹	م	۷۸۷۴	۷.۵۴
۳۰	ن	۴۱۷۶	۳.۹۹
۳۱	و	۵۷۶۹	۵.۵۲
۳۲	ه	۸۴۱۷	۸.۰۵
۳۳	ی	۵۸۷۵	۵.۶۲

با نگاهی به جدول (۳-۶) مشخص می‌شود که احتمال قرار گرفتن برخی از کاراکترها بعد از حرف "الف" بسیار بیشتر از سایر کاراکترهاست. از این ویژگی برای فشرده‌سازی بهتر متون بهره می‌بریم.

۳-۳-۲-۳- مدل زبانی ترایگرام[1]

مرحله سوم تحلیل بر اساس مدل زبانی[2]، ترایگرام می‌باشد. در این مرحله احتمال یک کاراکتر خاص بعد از هر زوج ترکیب حروف زبان بررسی می‌شود. بعنوان مثال احتمال قرار گرفتن کاراکتر "ب" بعد از زوج ترکیب "ار" و یا احتمال قرار گرفتن کاراکتر"ب" بعد از زوج ترکیب "می" و ... مورد بررسی قرار می‌گیرد.

در این مدل لازم است تمام زوج ترکیب‌های ممکن زبان؛ و احتمال قرار گرفتن هر یک از کاراکترها بعد از تمامی زوج کاراکترهای موجود محاسبه شود. نتایج بدست آمده برای زوج ترکیب "ار" در جدول شماره (۳-۷) نشان داده شده است.

جدول (۳- ۷)، نتایج مرحله ترایگرام در ارزیابی زوج کاراکتر "ار"

[1] Trigram
[2] Language model

۱۷	ش	۸۶۵	۸.۴۶
۱۸	س	۳۶۵	۳.۵۷
۱۹	ص	۳	۰.۰۳
۲۰	ض	۴۴	۰.۴۳
۲۱	ط	۰	۰
۲۲	ظ	۰	۰
۲۳	ع	۲۳	۰.۲۲
۲۴	غ	۱۲	۰.۱۲
۲۵	ف	۴۸	۰.۴۷
۲۶	ق	۲۴	۰.۲۳
۲۷	ک	۲۰۹	۲.۰۴
۲۸	گ	۲۰۸	۲.۰۳
۲۹	ل	۴۱	۰.۴۰
۳۰	م	۲۷۳	۲.۶۷
۳۱	ن	۷۶۳	۷.۴۶
۳۲	و	۲۸۲	۲.۷۶
۳۳	ه	۱۱۲۴	۱۰.۹۹
۳۴	ی	۲۳۱۷	۲۲.۶۶

ردیف	کاراکتر	تعداد تکرار	احتمال (٪)
۱	آ	۱۵	۰.۱۵
۲	ا	۷۴۳	۷.۲۷
۳	ب	۵۶	۰.۵۵
۴	پ	۰	۰
۵	ت	۷۲۷	۷.۱۱
۶	ث	۹	۰.۰۹
۷	ج	۳۶۵	۳.۵۷
۸	چ	۵۶	۰.۵۵
۹	ح	۵	۰.۰۵
۱۰	خ	۲۷	۰.۲۶
۱۱	د	۱۳۷۸	۱۳.۴۸
ردیف	**کاراکتر**	**تعداد تکرار**	**احتمال (٪)**
۱۲	ذ	۰	۰
۱۳	ر	۳	۰.۰۳
۱۴	ز	۲۳۷	۲.۳۲
۱۵	ژ	۳	۰.۰۳
۱۶	س	۳۶۵	۳.۵۷

۳-۳-۲-۴- مدل زبانی فورگرام[1]

فورگرام چهارمین مرحله‌ی تحلیل بر اساس مدل زبانی است. در این مرحله تمام ترکیب-هایی سه تایی ممکن، در زبان و احتمال قرار گرفتن هر یک از کاراکترها، بعد از این ترکیب‌ها، محاسبه می‌شود. بعنوان نمونه، احتمال قرار گرفتن کاراکتری مانند "د" بعد از ترکیب "آبا" مورد ارزیابی قرار می‌گیرد. نمونه‌ای از ترکیب‌های چهارتایی بدست آمده از این مرحله در جدول (۳-۸) نشان داده شده است. مشاهده می‌گردد در این مرحله خروجی بصورت یک ماتریس اسپارس می‌باشد.

جدول (۳- ۸)، نتایج مرحله فورگرام در ارزیابی ترکیب‌های چهارتایی کاراکترها

ردیف	ترکیب سه‌تایی	آ	ا	ب	پ	ت	ث	ج	چ	ح	خ	د	ذ	ر	ز	ژ	س	ش	ص	ض

[1] Fourgram

۱	اتآ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲	اتا	۰	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۳	اتب	۰	۴	۰	۰	۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۴	اتپ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۵	اتت	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۶	اتث	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۷	اتج	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۰	۰	۰	۰	۰	۰	۰
۸	اتچ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۹	اتح	۰	۴۲	۰	۰	۰	۰	۰	۰	۰	۰	۴۲	۰	۰	۰	۰	۰	۰	۰	۰
۱۰	اتخ	۰	۷	۱	۰	۰	۰	۰	۰	۰	۰	۰	۷	۰	۰	۰	۰	۰	۰	۰
۱۱	اتد	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۲	اتذ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۳	اتر	۰	۱	۰	۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۴	اتز	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۵	اتژ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۶	اتس	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۷	اتش	۰	۴	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۸	اتص	۰	۸	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۹	اتض	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۰	اتط	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۱	اتظ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۲	اتع	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۳	اتغ	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۴	اتف	۰	۱۴۸	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۵	اتق	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۶	اتک	۰	۸	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۷	اتگ	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۰	۰	۰	۰	۰	۰	۰
۲۸	اتل	۰	۶	۰	۰	۲۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲۹	اتم	۰	۱۹	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۰	۰	۰	۰
۳۰	اتن	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۳۱	اتو	۰	۶	۷	۰	۰	۰	۵	۰	۰	۰	۰	۱۹	۰	۱	۰	۰	۰	۰	۰
۳۲	اته	۰	۱۷	۰	۰	۰	۰	۰	۰	۰	۰	۰	۵	۰	۰	۰	۰	۰	۰	۰
۳۳	اتی	۰	۱	۰	۰	۳	۰	۰	۰	۴	۰	۷	۰	۰	۰	۴	۰	۰	۰	۰

۳-۴- مکانیزم بازیابی اطلاعات (خارج کردن از حالت فشرده)

در زمان حذف فشردگی، عملیات رمزگشایی صورت گرفته و داده‌های کد شده رمزگشایی شده و با استفاده از اطلاعات موجود تفسیر می‌شوند. سمبول‌های متناظر تولید شده و در خروجی برنامه قرار می‌گیرد. الگوریتم ایجاد شده در این مرحله بر حسب زمان کم مورد نیاز جهت حذف فشردگی یک سیستم مناسب می‌باشد.

فصل چهارم
تجزیه و تحلیل داده‌ها

۴-۱- داده‌های آموزشی

همانند هر مدل آماری دیگر در این سیستم نیز به داده‌هایی برای آموزش یا در واقع تنظیم پارامترهای آزاد مدل احتیاج داریم. برای تهیه یک مدل آماری، داده‌های آموزشی باید دارای توزیع احتمالی باشند که بعداً مدل با آنها سروکار دارد. یعنی به طور کلی داده‌های آموزشی باید فضای احتمال را پوشش دهند. در نتیجه اگر بخواهیم مدل آماری به دست آمده در همه جا قابل استفاده باشد، واضح است که باید انواع مختلف متون - از جمله ادبی، سیاسی، اقتصادی، طنز، ورزشی و ... را برای آموزش سیستم به کار گیریم.

برای تهیه داده‌های آموزشی با چنین حجم زیادی تنها راه عملی استفاده از متون موجود در شبکه اینترنت است. لذا بمنظور تهیه داده‌های آموزشی و ارزیابی روش پیشنهادی و حصول نتیجه، از متون فارسی استخراج شده از وب‌سایت‌های خبری فارسی زبان استفاده نموده‌ایم. در جدول (۴-۱) لیست منابع مورد استفاده، بمنظور جمع‌آوری متون آزمایشی، ذکر شده است.

جدول (۴-۱)، منابع مورد استفاده بمنظور جمع‌آوری متون آزمایشی

ردیف	نوع خبر	نام خبرگزاری	آدرس سایت
۱	سیاسی	ایسنا	http://www.isna.ir/
۲	ورزشی	خبر آنلاین	http://www. khabaronline.ir/
۳	علمی	ایرنا	http://www.irna.ir
۴	فرهنگی	فارس نیوز	http://www. farsnews.com/
۵	هنری	مهر	http://www.mehrnews.com/
۶	اقتصادی	خبرنگاران جوان	http://www.yjc.ir/
۷	اجتماعی	تابناک	http://www.tabnak.ir/

در تهیه این مجموعه، خبرهای استخراج شده‌ی سه روز ۱، ۲ و ۳ فروردین ماه ۱۳۹۴ به‌عنوان داده‌ی آموزشی، خبرهای ۴ فروردین ماه ۱۳۹۴ به عنوان مجموعه‌ی اعتبارسنجی یا توسعه و خبرهای ۵ فروردین ماه ۱۳۹۴ به‌عنوان داده‌ی تستی مورد استفاده قرار گرفته است.

لازم به ذکر است تمامی کلمات یکسان در این متون همسان‌سازی شده‌اند. به‌عنوان مثال از آنجا که کلمه "آئین نامه "می‌تواند به چهار شکل متفاوت "آئین نامه"، "آئین نامه"، "آیین نامه" و "آیین-نامه" نوشته شود، اشکال مختلف آن توسط افزونه‌ی ویراستیار که در ورد ۲۰۱۳ استفاده شده، به شکل "آئین نامه" تبدیل شده است. حروف "ی" به "ی" و "ک" به "ک" تبدیل شده‌اند که این امر به جهت سهولت در امر پردازش متون فارسی صورت گرفته است.

داده‌های آموزشی در قالب ۵ فایل train0.txt، train1.txt، train2.txt، train3.txt و train4.txt به فرمت UTF-8 آماده شده است. این فایل‌ها مجموعاً شامل ۲۲۸۳۰۵ کلمه می‌باشند. همچنین داده-های مجموعه توسعه در یک فایل بنام Dev.txt و داده‌های تستی در قالب یک فایل بنام test.txt در ذخیره شده است. این فایل‌ها به ترتیب شامل ۴۷۲۳۶ و ۵۵۳۳۵ کلمه می‌باشند.

۴-۲- ترتیب بیت‌ها بمنظور نمایش مدل‌های زبانی

به منظور بهینه‌سازی روش فشرده‌سازی ذکر شده با کمک تحلیل مدل زبانی، می‌بایست به این سئوال پاسخ داد که، در نظر گرفتن چند بیت برای نمایش پرکاربردترین کلمات و کاراکترها، بهینه‌ترین نتیجه ممکن را در بر خواهد داشت؟

برای پاسخ به سئوال فوق ما نتایج بدست آمده برای تعداد تکرار تمامی کلمات و کاراکترها را در نظر گرفته و از ترتیب بیتی مطابق با جدول (۴-۲)، برای نمایش مدل‌های زبانی، استفاده می‌نماییم. در این مدل برای نمایش هر مدل زبانی از یک بایت (۸ بیت) استفاده می‌شود که در آن اولین بیت نمایانگر نوع مدل زبانی مورد استفاده (کلمه‌ای=۱ و کاراکتری=۰) می‌باشد. در صورتی‌که بیت دوم صفر باشد نشان‌دهنده استفاده از مدل زبانی یونیگرام و در غیر اینصورت بیت‌های دوم و سوم نشان-دهنده‌ی نوع مدل زبانی بایگرام خواهند بود. (بایگرام=۱۰) سه بیت شماره ۲، ۳ و ۴ نشان دهنده‌ی استفاده از مدل زبانی ترایگرام و فورگرام خواهند بود (ترایگرام = ۱۱۰ و فورگرام =۱۱۱)

در ترتیب بیتی پیش‌رو اعداد ۶۴، ۳۲، ۱۶ و ۱۶ برای مدل‌های زبانی یونیگرام، بایگرام، ترایگرام و فورگرام (به منظور نمایش عبارات یک، دو، سه و چهار کلمه‌ای/کاراکتری) با توجه به تعداد تکرار کلمه یا کاراکتر در متون آموزشی در نظر گرفته شده است.

فرآیند اختصاص دادن هر یک از این اعداد به مدل‌های زبانی موجود، با توجه به تعداد تکرار بیشتر عبارات در نظر گرفته شده است. بطوریکه به تعداد تکرار بیشتر عبارات، عدد بزرگتری اختصاص یافته است. به عنوان نمونه با توجه به اینکه تعداد دو کلمه‌ای‌های پرتکرار بسیار بیشتر از چهار کلمه‌ای‌های پر تکرار است عدد بزرگتری برای دو کلمه‌ای‌ها (۳۲) نسبت به چهار کلمه ای‌ها (۱۶) در نظر گرفته شده است.

جدول (۴-۲)، ترتیب بیتی نمایش مدل‌های زبانی

شماره بیت		۸	۷	۶	۵	۴	۳	۲	۱
کلمه-ی زبانی	یونیگرام	۲۶ یک کلمه‌ای						۰	۱

	بایگرام	2^5 دو کلمه‌ای		۰	۱	۱
	ترایگرام	2^4 سه کلمه‌ای	۰	۱	۱	۱
	فورگرام	2^4 چهار کلمه‌ای	۱	۱	۱	۱
مدل زبانی کاراکتری	یونیگرام	2^6 یک کاراکتری			۰	۰
	بایگرام	2^5 دو کاراکتری		۰	۱	۰
	ترایگرام	2^4 سه کاراکتری	۰	۱	۱	۰
	فورگرام	2^4 چهار کاراکتری	۱	۱	۱	۰

با توجه به نمایش بیتی در نظر گرفته شده مطابق با جدول (۴-۲)، نمونه‌ای از کدهای اختصاص داده شده به ۱۶ سه کاراکتری پر تکرار در مدل‌زبانی ترایگرام به شرح ذیل می‌باشد.

جدول (۴-۳)، ترتیب بیتی اختصاص یافته به ۱۶ سه کاراکتری پرتکرار

دنباله سه کاراکتری پرتکرار	بیت شماره					
	۸	۷	۶	۵	۲، ۳ و ۴	۱
ایی	۰	۰	۰	۰	۱۱۰	۰
انی	۱	۰	۰	۰		
اند	۰	۱	۰	۰		
هان	۱	۱	۰	۰		
ران	۰	۰	۱	۰		
دار	۱	۰	۱	۰		
مان	۰	۱	۱	۰		
اری	۱	۱	۱	۰		
وان	۰	۰	۰	۱		
روز	۱	۰	۰	۱		
دان	۰	۱	۰	۱		
یان	۱	۱	۰	۱		
ارد	۰	۰	۱	۱		
کار	۱	۰	۱	۱		
اشت	۰	۱	۱	۱		
شود	۱	۱	۱	۱		

۴-۳- ارزیابی مدل‌های زبانی

یکی از چالش‌های اساسی موجود در مدل‌سازی نحوه‌ی ارزیابی دنباله‌های استخراج شده از متون آزمایشی است. [28] هدف از ارزیابی سیستم‌ها نزدیک کردن ارزیابی خودکار به ارزیابی توسط انسان می‌باشد. از آنجا که ارزیابی سیستم‌های اطلاعات زبانی توسط انسان بسیار وقت گیر و هزینه‌بر است، سیستم های ارزیابی خودکار در دنیای امروز بسیار پر اهمیت و پرکاربرد می‌باشند. این سیستم‌ها از طریق امتیازدهی به متن، بر اساس معیارهای خاص و تعریف شده‌ای عمل می‌کنند. این معیارها می-تواند با توجه به هدفی مشخص از سیستمی به سیستم دیگر (به منظور دستیابی به، سیستمی بهینه که تا حد مطلوبی معیارهای ارزیابی را پوشش دهد)، متفاوت باشند.

از جمله مشکلات ارزیابی روش‌های مبتنی بر N-gram ها یکسان در نظر گرفتن اهمیت کلمات و N-gram هاست. در حالی که از دیدگاه انسان هر کلمه ارزش و اهمیت متفاوتی نسبت به بقیه کلمات دارد. از جمله روش‌هایی که با توجه به تطابق با در نظر گرفتن اهمیت کلمات و معیارهای انسانی برای ارزیابی N-gram ها در نظر گرفته می‌شود، وزن‌دهی است. [29]

در این پژوهش وزن‌دهی با توجه به دو فاکتور تعداد تکرار و طول دنباله‌ها در متن صورت می‌پذیرد. به صورتی‌که طول و تعداد تکرار بیشتر ارزش بیشتری را برای دنباله‌ی مورد بررسی به ارمغان خواهد آورد. محاسبه میزان ارزش یک دنباله بر روی مجموعه‌ی آموزشی، با توجه به دو فاکتور فوق از طریق فرمول زیر انجام می‌شود؛

$$value = (lambda * x_1) + (1 - lambda * x_2) \qquad (۱-۴)$$

x_1 : ارزش عبارت با توجه به تعداد تکرار

x_{21} : ارزش عبارت با توجه به طول آن

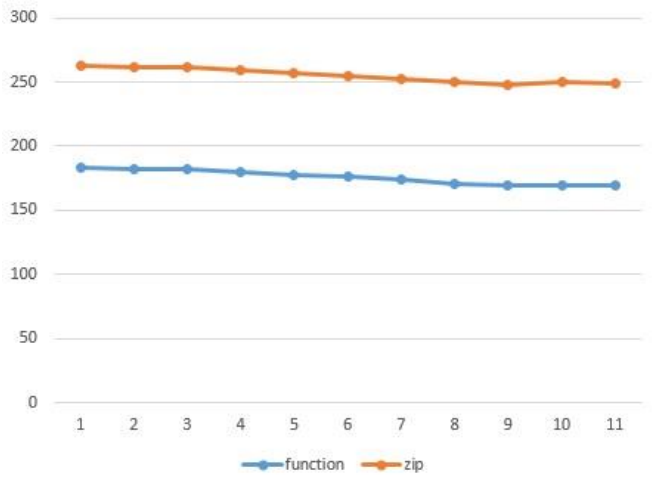
$lambda$: ضریبی با مقادیری بین ۰ تا ۱ (۰، ۰.۱، ۰.۲، ۰.۳، ۰.۴، ۰.۵، ۰.۶، ۰.۷، ۰.۸، ۰.۹، ۱)

به منظور مشخص نمودن مقادیر پارامترهای x_1 , x_2 از فرآیند نرمال‌سازی استفاده می‌نماییم. با در نظر گرفتن نرمال‌سازی به عنوان فرآیندی جهت نگاشت تمامی مقادیر پارامترها (ارزش عبارت با توجه به تعداد تکرار و طول آن) به مقادیری بین صفر تا یک، جهت شرح فرآیند نرمال‌سازی به ذکر مثالی در این مورد، در مدل زبانی فورگرام کلمات می‌پردازیم.

بالاترین تعداد تکرار و طول رشته را در میان عبارات موجود را یافته و به نرمال‌سازی آن می‌پردازیم. در این عبارات بالاترین تکرار ۹۷ و بالاترین طول رشته ۳۷ حرف می‌باشد. به بالاترین تکرار (۹۷) ارزش ۱، به حد وسط آن (۴۹) ارزش ۰.۵ و به همین نسبت به سایر تعداد تکرارها ارزش‌هایی بین صفر تا یک را نسبت می‌دهیم. همین عمل را برای پارامتر طول رشته انجام می‌دهیم. به ماکزیمم طول رشته موجود (۳۷) ارزش ۱، به حد وسط آن (۱۹) ارزش ۰.۵ و به سایر طول رشته‌های موجود به همین تناسب مقادیری بین صفر تا یک را نسبت می‌دهیم.

این فرآیند علاوه بر مدل زبانی فورگرام برای سایر مدل‌های زبانی نیز محاسبه می‌گردد. در ادامه با در نظر گرفتن مقادیر مختلف برای متغیر lambda (مقادیری بین ۰ تا ۱- ۱۱ مورد)، بر اساس فرمول (۴-۱) به محاسبه ارزش یک عبارت می‌پردازیم. سپس با توجه به فرمول (۴-۱)، به ازای هر lambad به انتخاب ۶۴ تک کلمه‌ای، ۳۲ دو کلمه‌ای، ۱۶ سه کلمه‌ای و ۱۶ چهار کلمه‌ای پر ارزش در هر گروه می‌پردازیم. با توجه به نتایج بدست آمده از محاسبات فوق و در نظر گرفتن مجموعه عبارت‌های منتخب در هر یک از ۱۱ گروه (بر اساس امتیازهای کسب شده) به فشرده‌سازی مجموعه‌ی توسعه (از پایگاه دادگان ایجاد شده) می‌پردازیم. بر اساس نتایج حاصل از میزان فشرده‌سازی هر گروه، بهینه‌ترین حالت را انتخاب و به ارزیابی آن بر مجموعه تست می‌پردازیم.

نمودار زیر نتایج را با در نظر گرفتن مقدار لامبدای یکسان برای مدل‌های زبانی مختلف نشان می‌دهد.



شکل (۴-۱) ، نتایج جستجوی مقادیر بهینه

علاوه بر ارزیابی به ازای لامبداهای یکسان برای هر مدل زبانی، این ارزیابی به ازای در نظر گرفتن مقدار لامبدای متفاوت برای هر مدل زبانی نیز انجام شده است.

۴-۳-۱- استفاده از معیار perplexity در ارزیابی

معمول‌ترین راه برای ارزیابی یک مدل احتمالی اندازه‌گیری لگاریتم احتمال وقوع برای یک مجموعه داده‌ی آزمون که قبلاً مشاهده نشده، می‌باشد. احتمال وقوع متون مشاهده نشده در زمان آموزش می‌تواند برای مقایسه‌ی مدل‌های زبانی مختلف استفاده شود. مدل‌های با احتمال وقوع بالاتر مدل‌های بهتری هستند.

در تئوری اطلاعات یکی از معیارهای کمی مورد استفاده برای ارزیابی یک مدل زبانی و بررسی میزان مناسب بودن یک توزیع احتمالی یا پیش‌بینی مدل احتمالی معیار Perplexity[1] است، معیاری که به‌طور سنتی برای ارزیابی مدل‌های زبانی استفاده می‌شود. Perplexity تابعی نزولی از لگاریتم احتمال وقوع متون مشاهده نشده است. هر چه میزان Perplexity کمتر باشد مدل بهتر است. در واقع سرگشتگی یک معیار مستقل از سیستم و متناسب با احتمال‌های نسبت داده شده به دنباله‌هاست و

[1] سرگشتگی

می‌توان از آن برای مقایسه مدل‌های احتمالاتی استفاده کرد. در این مدل‌ها، دنباله‌ای بهتر است که در آن پیش‌بینی‌ها احتمال بیشتری (معکوس سرگشتگی) داشته باشند. [30]

$$PP(W) = \frac{1}{p(w_1 w_2 \dots . w_N)^{\frac{1}{N}}} = p(w_1 w_2 \dots . w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots . w_N)}}$$

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

از معیار سرگشتگی تحت عنوان میانگین فاکتور انشعاب (میانگین تعداد کلمات ممکن بعد از هر رشته یا کلمه) نیز یاد می‌شود. مدل زبانی بهتر (قوی‌تر)، سرگشتگی کمتری را نتیجه می‌دهد. در مجموعه داده موجود معیار سرگشتگی بر روی مجموعه آزمون مورد ارزیابی قرار می‌گیرد. جهت ارزیابی این ویژگی از برنامه آماده شده در محیط Python استفاده شده است. در این پژوهش مجموعه داده‌ها با اعمال هموارسازی جمع با یک (لاپلاس)[1] جهت رفع مشکل احتمال‌های صفر در N-gram مورد ارزیابی قرار گرفته است. نمونه‌ای از خروجی‌های بدست آمده در این مرحله به شرح زیر می‌باشد.

Smoothing function: LaPlace

training $2-$ gram language model
training finished, calculating perplexity
train set perplexity $=$ 114.694439
test set perplexity $=$ 114.693661

training $3-$ gram language model
training finished, calculating perplexity
train set perplexity $=$ 280.724717
test set perplexity $=$ 280.720527

training $4-$ gram language model

[1] Laplace Smoothing

training finished, calculating perplexity

train set perplexity = 466.851562

test set perplexity = 466.840437

در مقایسه کارایی مدل‌های زبانی مختلف با استفاده از اعداد حاصل از ارزیابی فوق می‌توان گفت مدل زبانی 2-gram قوی‌تر از 3-gram و مدل 4-gram قوی‌تر از 3-gram می‌باشد. این ارزیابی بیانگر صحت مدل کردن جملات تست توسط مدل آموزش داده شده با مجموعه آموزش می‌باشد.

فصل پنجم
نتیجه گیری و پیشنهادات

۵-۱- نتیجه‌گیری

فشرده‌سازی داده‌ها یک نوع عمل کدینگ است که در آن داده‌های ورودی به طریقی کد می‌شوند که فضای کمتری را اشغال نموده و نیز بتواند دوباره در هر زمان دلخواه بازیابی شده و داده اصلی را بازگرداند. [2] هدف اصلی در این پژوهش ارائه تکنیکی به منظور فشرده‌سازی متون فارسی می‌باشد. تا در زمان انتقال فضای کمتری را اشغال نموده و سرعت انتقال آن بیشتر شود.

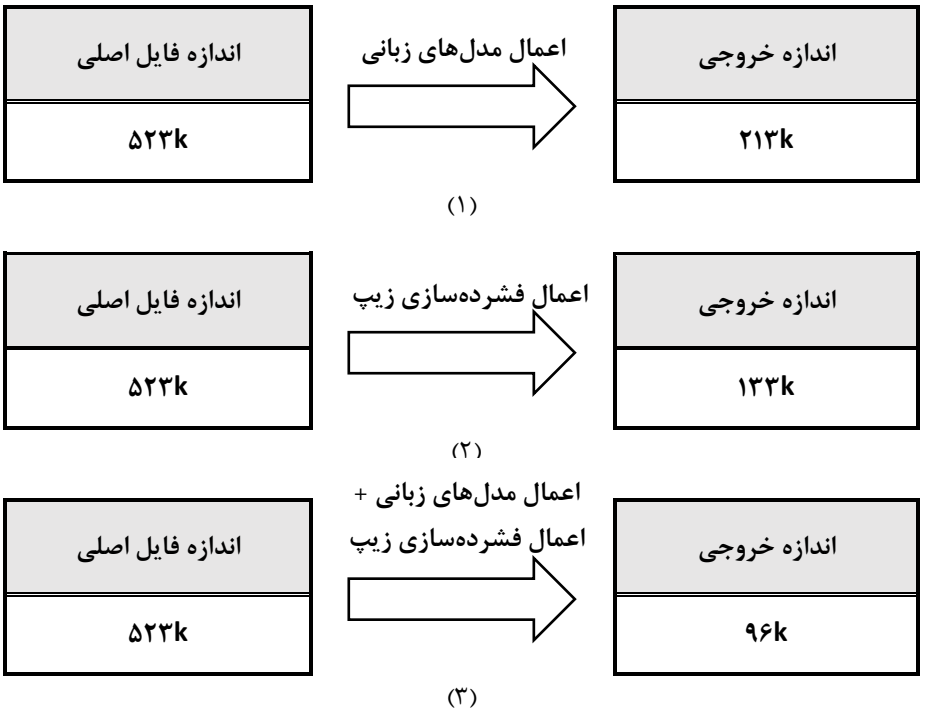
در روش ارائه شده با در نظر گرفتن شیوه‌ی جدیدی جهت استفاده همزمان از مدل‌های زبانی کلمات و کاراکترها و همچنین تعیین مدلی به منظور وزن‌دهی به عبارات و تعیین مقادیر بهینه سعی بر بهبود میزان فشرده‌سازی در متون فارسی با استفاده از اطلاعات زبانی داشته‌ایم. در نهایت با استفاده از معماری جدید، بر مبنای معیارهای مشخص موجود و با استفاده از سناریوی موجود به یافتن مقادیر بهینه پرداخته و در نهایت به ارزیابی آن بر روی متون مجموعه آزمون می‌پردازیم.

هر یک از کاراکترهای استفاده شده در متون فارسی که از استاندارد یونیکد پیروی می‌نماید، ۱۶ بیت فضا را اشغال می‌کند. در این پایان‌نامه فشرده‌سازی در دو مرحله صورت پذیرفته است. مرحله اول استفاده از مدل زبانی کلمات است که در نتیجه آن نگاشت ۱۶ دنباله چهار کلمه‌ای، ۱۶ دنباله سه کلمه‌ای، ۳۲ دنباله دو کلمه‌ای و ۶۴ دنباله یک کلمه‌ای با ارزش با یک کد ۸ بیتی صورت می‌پذیرد. و در مرحله دوم از خروجی‌های مدل زبانی کاراکترها برای فشرده‌سازی بیشتر استفاده می‌گردد. در این راستا هر یک از مدل‌های زبانی یونیگرام، بایگرام، ترایگرام و فورگرام بر روی متون آموزشی برای بدست آوردن توالی‌های ۴، ۳، ۲ و ۱ کاراکتری پرکابرد مورد ارزیابی قرار گرفته‌اند. جهت رسیدن به جواب بهینه‌تر آزمون‌های لازم برای رسیدن به راه‌حل صحیح‌تر و مقادیر با ارزش‌تر بر روی مجموعه توسعه انجام و بهترین راه حل بر روی مجموعه تست مورد ارزیابی قرار گرفته است.

در آزمون دیگری با توجه به عدم استفاده الگوریتم فشرده‌سازی زیپ از اطلاعات زبانی، با اعمال فشرده‌سازی زیپ بر روی خروجی حاصل از فشرده‌سازی بدست آمده با استفاده از

مدل‌های زبانی توصیف شده، به بهبود در اندازه فایل خروجی فشرده شده رسیدیم. نتایج حاصله به شرح ذیل می‌باشد؛

جدول (۵-۱)، نتایج ارزیابی بر روی مجموعه تست



در جدول فوق در بخش (۱) نتیجه اعمال مدل‌های زبانی و در بخش (۲) نتیجه اعمال فشرده‌سازی زیپ و در بخش (۳) نتیجه اعمال متوالی مدل‌های زبانی و فشرده‌سازی زیپ بر مجموعه داده‌ی تست نشان داده شده است.

با توجه به جدول (۵-۱)، می‌توان به این نتیجه رسید که با استفاده از اطلاعات زبانی در فشرده‌سازی فایل‌های متنی می‌توان به نتایج مطلوبی در فشرده‌سازی، کاهش حجم فایل مبدا (۶۰ ٪) و میزان صرفه‌جویی بیشتری در فضای حافظه مصرفی رسید. علاوه بر این با اعمال فشرده‌سازی زیپ بر فایل خروجی بدست آمده، به میزان فشرده‌سازی بیشتری نسبت به اعمال مستقیم آن بر روی فایل ورودی خواهیم رسید. (۸۲ ٪) و این امر بیانگر تاثیر مثبت کاربرد اطلاعات زبانی در الگوریتم فشرده‌سازی زیپ می‌باشد.

با توجه به فرمول (۲-۵) که جهت محاسبه نرخ فشرده‌سازی ارائه شده است، میزان نرخ فشرده‌سازی ۸۱.۶۵٪ با استفاده از اطلاعات زبانی در فشرده‌سازی زیپ بدست آمده است. همچنین مقایسه سرعت فشرده‌سازی الگوریتم فشرده‌سازی زیپ و روش ارائه شده در این پژوهش بیانگر سرعت تقریباً دو برابری (۱.۹ برابر) الگوریتم فشرده‌سازی زیپ می‌باشد.

در ارزیابی دیگری بر روی فایلی با اندازه ۱.۹۲ MB نتایجی به شرح زیر به دست آمده است؛

جدول (۵-۲)، نتایج ارزیابی بر روی مجموعه تست دوم

اندازه فایل زیپ شده		اندازه خروجی با استفاده از مدل‌های زبانی	اندازه فایل اصلی
مدل زبانی	فایل اصلی		
۳۶۴k	۴۶۹k	۸۸۳k	۲,۰۲۳,۲۰۹k

نتایج فوق بیانگر صحت نتایج بدست آمده در مرحله آزمون می‌باشد.

۵-۲- نتیجه ارزیابی روش پیشنهادی بر روی متون انگلیسی

در آزمون دیگری روش پیشنهادی بر روی متون انگلیسی مورد ارزیابی قرار گرفته است. در این مرحله پس از آماده‌سازی پایگاه دادگان انگلیسی، فرآیندهای لازم جهت آموزش سیستم و انتخاب مدل‌های زبانی با کارایی بهتر (مطابق با مراحل طی شده برای متون فارسی) انجام و پس از انجام محاسبات مورد نیاز، نتایجی به شرح ذیل بدست آمده است.

جدول (۵-۳)، نتایج ارزیابی بر روی مجموعه متون انگلیسی

اندازه فایل زیپ شده		اندازه خروجی با استفاده از مدل‌های زبانی	اندازه فایل اصلی
مدل زبانی	فایل اصلی		
۷۱۶k	۸۰۴k	۲,۰۹۵k	۳,۳۰۹k

جدول (۵-۳) بیانگر تاثیر مثبت استفاده همزمان از مدل‌های زبانی کلمات و کاراکترها در فشرده سازی متون انگلیسی و تاثیر مثبت کاربرد اطلاعات زبانی در الگوریتم فشرده‌سازی زیپ می‌باشد.

۵-۳- پیشنهادات

به عنوان ایده‌های پژوهشی جهت ادامه کار در این حوزه می‌توان به موارد زیر اشاره نمود:

۱) در پژوهش حاضر مدل‌های زبانی بیشتر از 4-gram مورد بررسی قرار نگرفته است. قطعاً مدل‌های بزرگتر ماتریس‌های اسپارس بزرگی را تولید خواهند نمود اما ممکن است نرخ فشرده‌سازی بیشتری را ایجاد نموده و به یک راه حل مناسب در فشره‌سازی تبدیل شود. لذا بررسی تاثیر N-gram های بزرگتر یکی از جنبه‌های تحقیقاتی کار ما در آینده خواهد بود.

۲) روش ارائه شده در این پژوهش با حداقل فرضیات در مورد ساختار متون در نظر گرفته شده است. بنابراین می‌توان آن را برای سایر زبان‌ها با حداقل تغییرات بکار برد. هدف ما از این‌کار تست روش ارائه شده برای زبان‌هایی است که اختلاف ساختاری با زبان فارسی دارند. (اختلافاتی از قبیل تعداد حروف و تعداد تکرار اطلاعات زبانی در ساختار آن). لذا بررسی روش ارائه شده فوق در سایر زبان‌ها، می‌تواند یکی از اهداف پژوهشی آینده محسوب شود.

فهرست منابع

[1] نان‌گیر م، (۱۳۹۱)، پایان نامه ارشد :" تحلیل الگوریتم‌های کدگـذاری جـامع بـرای کـلاس‌هـای خاصـی از منـابع اطلاعاتی"، دانشکده مهندسی برق ،دانشگاه صنعتی شریف

[2] ربانی ع، (۱۳۹۲)، پایان نامه ارشد :" شبیه‌سازی کدینگ تصـویری بـا تبـدیل DCT"، دانشـکده مهندسـی بـرق، دانشگاه زنجان

[3] Ling Sun T. , Sei Ping L. and Chong Eng T. , "Optimizing LZW Text Compression Algorithm via Multithreading Programming", Malaysia International Conference on Communications , CSI-02974 , 2009

[4] جهانگیر ا،(۱۳۷۵)، " فشرده‌سازی متون فارسی "، دومـین کنفـرانس سـالانه انجمـن کـامپیوتر ایـران ، دانشـکده مهندسی کامپوتر،دانشگاه صنعتی شریف

[5] Batista L. and Alexandre L. A. , "Text Pre-processing for Lossless Compression", pp. 506, 2008

[6] Kalajdzic K. , Samaher H. A. and Petal A. , "Rapid lossless comparison of short text message", Computer Standard & Interfaces , CSI-02974 , 2014.

[7] Koditowakku S. R. and Amarasinghe U. S. , "Comparison of lossless data compression algorithms for text data", Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 416-425, 2010

[8] Shannon C.E. ," A mathematicatlh eory of communications,B" ell Syst. Tech. J. , vol. 27, pp. 379-423, July. 1948

[9] فرسی ح. و اعتضادی فر پ ،(۱۳۹۱) " فشرده‌سازی اطلاعات متغیر با زمان با اسـتفاده از کـد هـافمن" **فصـلنامه پردازش علائم و داده‌ها** ، شماره ۲، پیاپی۱۸: ص۶۱

[10] Islam M. R. and Mahamud A. , "AN ENHANCED STATIC DATA COMPRESSION SCHEME OF BENGALI SHORT MESSAGE", (IJCSIS) International Journal of Computer Science and Information Security Vol. 4, No. 1 & 2, 2009

[11] خسروی فرد م، (۱۳۷۷)، پایان نامه ارشد :" فشرده‌سـازی عمـومی داده‌هـا"، دانشـکده مهندسـی بـرق، دانشـگاه صنعتی شریف.

[12] Ziv J. and Lempel A. "A universal algorithm for sequential data compression" IEEE Transactions on Information Theory, 23, pp. 337–343, 1977

[13] Ziv J. , Lempel A. "Compression of individual sequences via variable-rate coding". IEEE Transactions on Information Theory, 24, pp. 530–536, 1978

[14] مبینی م.(۱۳۹۲) " روشی جدید برای افزایش ظرفیت پنهان‌نگاری در متون فارسی مبتنی بر فشرده‌سازی LZW "، هشتمین کنفرانس ماشین بینایی و پردازش تصویر ایران ، دانشکده مهندسی کامپوتر،دانشگاه صنعتی امیرکبیر

Hoang D. T. , Long P. M. , Vitter J. S. "Dictionary Selection using Partial Matching". Information sciences, 119, pp. 57–72, 1999

[16] R. Steinmettz and K. Nahrstedt, "Multimedia: Computing, Communications and Applications", Prentice Hall, 1995.

[17] جعفری نجفی م، (۱۳۷۵)، پایان نامه ارشد :" فشرده‌سازی داده‌ها و بکارگیری آن در سیستم‌های تبادل اطلاعات"، دانشکده مهندسی برق ،دانشگاه صنعتی شریف.

[18] A.Wesley. "The Unicode Consortium, The Unicode Standard, Version 3.0", Addison- Press, 2000

[19] سلطانی نژاد ف.(۱۳۸۸) " بررسی فونت های یونی کد رایانه ای فارسی و عربی" ادبیات تطبیقی، سال سوم، شمارهٔ 11 ؛ ص1 143

[20] Rosenfeld R. , Chen S. F. and Zhu, X. . "Whole-sentence exponential language models: a vehicle for linguistic-statistical integration", Computer Speech & Language, vol. 15(1), pp. 55-73, January. 2001

[21] Corazza A. , De Mori R. , Gretter R. and Satta G. "Language modeling using stochastic context-free grammars", Speech Communication, vol. 13(1-2), pp. 163-170, October. 1993

[22] Rosenfeld R. "Two decades of statistical language modeling: Where do we go from here?", Proceedings of the IEEE, vol. 88, pp. 1270-1278, 2000

[23] Gunter Bolch ,Stefan Greiner, Hermann de Meer Kishor S. Trivedi.Queueing Networks and Markov Chains Modeling and Performance Evaluation with Computer Science Applications., 2006

[24] TAGHIYAREH F. , DARRUDI E. , OROUMCHIAN F. and ANGOSHTARI N. "Compression of Persian Text for Web-Based Applications, Without Explicit Decompression", Computer Standard & Interfaces , CSI-02974 , 2014.

[25] Nagao M. and Mori S., "A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese", In Proceedings of International Conference on Computational Linguistics, pp. 611-615, 1994

[26] Stolke A. , "SRILM — AN EXTENSIBLE LANGUAGE MODELING TOOLKIT", in Conf. On Spoken Language Processing (ICSLP), vol. 2, pp. 901-904, 2002

[27] Katz M. , "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer ", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, pp. 400-401, 1997

[28] بعثت ز، عبدالله زاده بـارفروش ا.(۱۳۸۸)، " Aut-BLEU ، سیسـتم ارزیابی ترجمه‌ی ماشینی با رویکردی نو در وزن‌دهی به N-gram ها بر اساس نقش کلمه در جمله"، پانزدهمین کنفـرانس بـین‌المللـی انجمـن کـامپیوتر ایـران، دانشگاه صنعتی امیرکبیر

[29] Arun R. , Suresh V. , Madhavan C. V. , Murthy M. N. , "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations", Advances in Knowledge Discovery and Data Mining, vol. 6118, pp. 391-402, 2010

[30] Bharadwaj S. and Hasegawa-Johnson M. , "A PAC-Bayesian Approach to Minimum Perplexity Language Modeling", in Proc. COLING, pp.130-140, 2014

Abstract

The growth of digital data in recent years, cause of increasing attention has been text compressing. Information of text type that every day, witness to send and receive it. The need to reduce the amount of data and saving storage space, compression has become a critical phenomenon. With increasing non-English and non-Latin texts, the need for the development of compression algorithms also be felt in other languages. This thesis is an attempt to provide a technique to compress Persian texts.

In this study, the purpose is using the rules and techniques of modeling languages. The rules that in the known and common compression algorithms such as zip is not considered. In this technique, with use a statistical model N-gram, we investigate the probability of the sequence of words and language characters after another, considering the number of repetitions and the length parameters. To evaluate and select the model with the more performance is using from perplexity criteria that is independent of the system and fit with probability attributed to expressions. Compare the results obtained, show the compression rate 82% with regard to language information the output the zip compression. In the future stages of analysis based on language modeling, results will be described the procedures and result obtained.

Keywords: Compression, Coding data, Persian literature, language model

Shahrood University of Technology

Faculty Computer Engineering

e-Learning Center

Text Compression Using Artificial Intelligence Techniques

Mahboubeh Soleymanian

Supervisor: Dr. Ali Akbar Pouyan

Date: September 2015