تمرین اول پردازش زبان های طبیعی

فرشاد برجعلی زاده

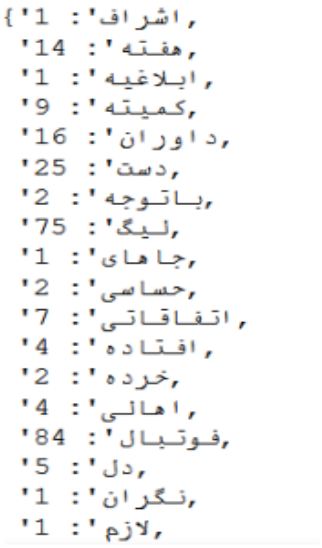
۶۱۰۳۹۹۰۱۵

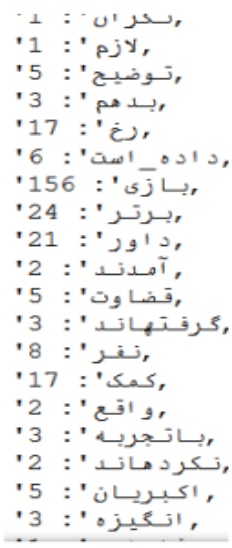
در این تمرین قصـــد داریم که مدل های زبانی unigram, bigram, trigram را پیاده سازی کنیم.

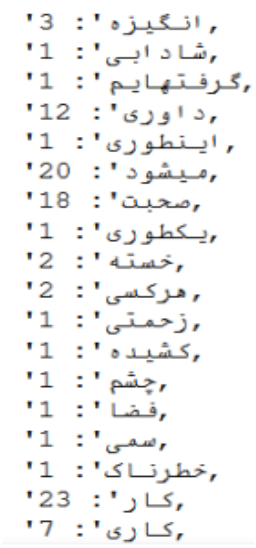
داده های متنی موجود در این تمرین به صـورت دسـتی از خبرهای سـایت ورزش۳ جمع آوری شده است . فایل های دیتاست به صورت جدا گانه در دو پوشه train و test وجود دارند. از آنجایی که بسیاری از کلمات از جمله اعداد و نماد های نگارشی و ایست واژه ها تاثیری در پیکره متنی ندارند آن ها را می توانیم حذف کنیم و با تعداد کمتری توکن مدل را آموزش دهیم.

برای پیش پردازش متن فارسـی از کتابخانه مفید hazm اسـتفاده میکنیم و همچنین برای پیش پردازش و نرمال سـازی بهتر تعدادی از ویژگی ها را به عنوان مثال حذف اعداد ریاضی و نیم فاصله ها را خود به صورت دستی و کد نویسی بر روی پیکره متنی اعمال می کنیم.

برای اینکه توکن ها را بدسـت آوریم از تابع "word_tokenize" موجود در کتابخانه hazm استفاده میکنیم داده های آموزش قبل از نرمال کردن ۱۳۹۳۳ توکن بودند که بعد از آن به ۶۲۵۶ عدد رسـیدند و برای داده های تسـت هم ۲۲۴۰ به ۱۰۳۲ توکن کاهش یافتند، لغات منحصـر به فرد را در یک فایل vocab.txt ذخیره می کنیم و همچنین داده های نرمال شـده را دو فایل train.txt و test.txt ذخیره می کنیم. تصـویر زیر تعدادی از کلمات به همراه تکرارشان را در پیکره متنی را مشاهده می کنیم.







حال با استفاده از SRILM toolkit می خواهیم که N-gram ها را محاسبه کنیم و از روش هموار سازی "good-turing" هم برای این کار بهره می گیریم که در جدول زیر خلاصه ای از آنچه که حاصل شده است را می بینیم.

N	Smoothing method	Perplexity
bigram	Kneser-Ney	۳۸۰.۵۲۶
bigram	good-turing	۴۰۱.۰۲۳
trigram	Kneser-Ney	۲۹۰.۲۶۸
trigram	good-turing	۳۰۹.۱۲۵

و همین طور در تصویر زیر میتوان ۱۰ بایگرم که بیشترین تعداد تکرار را دارن مشاهده کرد.