

تمرین اول پردازش زبان های طبیعی

فرشاد برجعلی زاده

۶۱۰۳۹۹۰۱۵

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

داده های متنی موجود در این تمرین به صورت دستی از خبرهای سایت ورزش ۳ 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 فارسی از کتابخانه مفید hazm استفاده میکنیم و همچنین برای پیش پردازش و نرمال سازی 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 بدست 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 رسیدند و برای داده های تست هم ۲۲۴۰ به ۱۰۳۲ توکن کاهش یافتند، لغات منحصر به فرد را در یک فایل vocab.txt ذخیره می کنیم و همچنین داده های نرمال شده را دو فایل train.txt و test.txt ذخیره می کنیم. تصویر زیر تعدادی از کلمات به همراه تکرارشان را در پیکره متنی را مشاهده می کنیم.

{ 'اشراف' : '1' ,	'بخران' : '1' ,	'انگیزه' : '3' ,
'هفته' : '14' ,	'لازم' : '1' ,	'شادابی' : '1' ,
'ایلاغیه' : '1' ,	'توضیح' : '5' ,	'گرفته‌ایم' : '1' ,
'کمیت' : '9' ,	'بدهم' : '3' ,	'داوری' : '12' ,
'داوران' : '16' ,	'رخ' : '17' ,	'اینطوری' : '1' ,
'دست' : '25' ,	'داده_است' : '6' ,	'میشود' : '20' ,
'باتوجه' : '2' ,	'بازی' : '156' ,	'صحبت' : '18' ,
'لیگ' : '75' ,	'برتر' : '24' ,	'یکطوری' : '1' ,
'جاهای' : '1' ,	'داور' : '21' ,	'خسته' : '2' ,
'حساسی' : '2' ,	'آمدند' : '2' ,	'هرکسی' : '2' ,
'اتفاقاتی' : '7' ,	'قضاوت' : '5' ,	'زحمتی' : '1' ,
'افتاده' : '4' ,	'گرفته‌اند' : '3' ,	'کشیده' : '1' ,
'خرده' : '2' ,	'نقر' : '8' ,	'چشم' : '1' ,
'اهالی' : '4' ,	'کمک' : '17' ,	'فضا' : '1' ,
'فوتبال' : '84' ,	'واقع' : '2' ,	'سمی' : '1' ,
'دل' : '5' ,	'باتجربه' : '3' ,	'خطرناک' : '1' ,
'نگران' : '1' ,	'نکرده‌اند' : '2' ,	'کار' : '23' ,
'لازم' : '1' ,	'اکبریان' : '5' ,	'کاری' : '7' ,
	'انگیزه' : '3' ,	

حال با استفاده از SRILM toolkit می‌خواهیم که N-gram ها را محاسبه کنیم و از روش هموار سازی “good-turing” هم برای این کار بهره می‌گیریم که در جدول زیر خلاصه ای از آنچه که حاصل شده است را می‌بینیم.

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 همین‌طور در تصویر زیر میتوان ۱۰ بایگرم که بیشترین تعداد تکرار را دارن مشاهده کرد.

```
[ ('ورزش' , 'سه' ) ,
  ('جام' , 'جهانی' ) ,
  ('قدراسیون' , 'فوتبال' ) ,
  ('لیگ' , 'برتر' ) ,
  ('لیگ' , 'قهرمانان' ) ,
  ('کره' , 'شمالی' ) ,
  ('تیم' , 'ملی' ) ,
  ('خوشه' , 'طلایی' ) ,
  ('لیگ' , 'بیستم' ) ,
  ('قهرمانان' , 'آسیا' ) ]
```