به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش زبان طبیعی

استاد ممتازی

تمرین اول

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش دوم: مدل‌های زبانی آماری

پیش از هر چیز لازم است توضیحاتی در مورد نرمال‌سازی استفاده شده بدهم. فرمول نرمال‌سازی Absolute Discounting در حالت عادی و برای Bigram عبارت است از:

چنانچه برای یک جفت کلمه در زمان آموزش وجود داشته باشد، روابط فوق قابل استفاده است. در پیاده‌سازی من برای سرعت بیشتر مقادیر B به ازای کلمات موجود در لغت‌نامه یک بار محاسبه و نگهداری می‌شود. همچنین اگر یک کلمه یک بار و در انتهای یک مصرع آمده باشد دارای مقدار B‌ برابر با صفر خواهد بود. در نتیجه مقدار α هم برابر با صفر می‌شود. بدین ترتیب رابطه فوق امکان تولید احتمال صفر را خواهد داشت که این امر باعث مشکلاتی در قسمت‌های بعدی می‌شود. برای حل این مشکل من تمامی B‌ هایی که برابر با صفر هستند را یک واحد افزایش دادم.

مسئله دیگری که باید به آن پرداخت استفاده از رابطه نرمال‌سازی مذکور برای جفت‌هایی است که دیده نشده است. برای این حالت مقدار B، و همگی برابر با صفر می‌شود. طبیعتا اولین جمله در رابطه نرمال‌سازی و α دیگر قابل تعریف نیستند. برای این حالت بنا به سادگی نرمال‌سازی را به شکل زیر تعریف کرده‌ام:

احتمال هم مطابق معمول برابر با Unigram کلمه تعریف می‌شود.

حال نوبت به تعریف Unigram مطابق با نرمال‌سازی Absolute Discounting می‌رسد! برای Unigram احتمال برابر با Zerogram می‌شود. یعنی:

در این رابطه V برابر با لغت‌نامه داده‌های آموزشی است. از آنجایی که در Unigram‌ کلمه تعریف نشده است، می‌توان فرض کرد که اندازه کل کلمات پیکره یعنی N را برای درنظر گرفت و را معادل دانست. با منطقی مشابه می‌توان اندازه لغت‌نامه را برای مقدار B تعیین کرد. پس روابطی که برای Bigram داشتیم به این شکل تغییر می‌کند:

نهایتا برای یک احتمال یک کلمه خواهیم داشت:

در روابط نرمال‌سازی باید ابرپارامتر وجود دارد که باید آن را تنظیم دقیق کرد. برای تنظیم این پارامتر مطابق درخواست سوال از مقدار Perplexity‌ روی مجموعه اعتبارسنجی کمک گرفته‌ام. مقدار بهینه برای Unigram برابر با 0.96 و این مقدار برای Bigram برابر با 0.88 بدست آمد.

مقدار Perplexity برای دو مدل Unigram‌ و Bigram با مقادیر بهینه و برای سه مجموعه‌داده در جدول زیر آورده شده است:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train | Validation | Test |
| مدل Unigram | *1768* | *1828* | *1810* |
| مدل Bigram | *322* | *1326* | *1318* |

بخش سوم: تکمیل جملات ناقص با استفاده از مدل‌های زبانی آماری