گزارش کار تمرین اول درس پردازش زبان‌های طبیعی
طراحی مدل‌های زبانی تولیدکننده متن

محمد لشکری ۴۰۰۱۱۲۰۸۷

۲۰ اسفند ۱۴۰۰

۱ جمع‌آوری دادگان

۲۷۳۴ عنوان خبری از بخش آرشیو سایت‌های ایرنا و ایسنا جمع‌آوری شد. دادگان با نسبت ۰/۲ به آموزش و آزمایش تقسیم‌بندی شده‌اند. عناوین خبری شامل نیم‌فاصله، خط فاصله و عدد بودند که مدیریت آن‌ها در بخش بعد تشریح می‌گردد. کدهای مربوط به جمع‌آوری دادگان در فایل title-scraper/scraper.py و دادگان جمع‌آوری شده در فایل title-scraper/titles.txt قرار دارد.

۲ پیش پردازش دادگان

کاراکترهای اضافی(به جز ؟،حروف فارسی، عدد، نقطه) حذف و اعداد نیز با کاراکتر N جایگزین شدند. مجموعه آموزشی و آزمایشی به ترتیب شامل ۷۱۲ و ۵۰۸ کلمه یکتا هستند. فایل frequent.txt شامل ۲۰۰ کلمه پرتکرار دادگان آموزشی است. چون تعداد کلمات یکتای مجموعه آموزشی از ۱۰۰۰۰ کمتر است unk در مجموعه لغات وجود ندارد. فایل‌های dataset/titles_train.txt و dataset/titles_test.txt به ترتیب مجموعه‌های آموزشی و آزمایشی هستند. دادگان این دو فایل جداسازی شده دادگان فایل title-scraper/titles.txt با نسبت ۰/۲ هستند.

۳ طراحی مدل زبانی

مدل‌های زبانی bigram و trigram برای تولید عناوین خبری پیاده‌سازی شدند. توابع cal_2gram_prob و cal_3gram_prob برای محاسبه احتمال وقوع کلمه بعدی به کلاس اضافه شدند.

تابع Average_log_likelihood برای محاسبه معیار ارزیابی مدل‌ها به کلاس اضافه شد. پارامتر prepared_test_data در تابع evaluate_model از جنس prepared_data است که عضو داده‌ای کلاس DataProcessor است.

۴ ارزیابی

مقدار Average log likelihood روی ۳۰۰ نمونه از دادگان آزمایشی محاسبه شد که نتایج آن در جدول زیر قابل مشاهده است:

3gram	2gram
−۲/۴۶	−۲/۵۷

جدول ۱: نتایج

همان‌طور که انتظار می‌رفت مقدار فوق برای 2gram کمتر از 3gram به‌دست آمد. دلیل این امر آن است که 3gram برای محاسبه احتمال وقوع کلمه در جایگاه فعلی، دو کلمه قبل از آن را در نظر می‌گیرد. در حالی که 2gram تنها کلمه قبل را در نظر می‌گیرد. لازم به ذکر است به دلیل دامنه مقادیر احتمال(بین صفر و یک) منفی شدن مقدار فوق طبیعی است و هر چه به صفر نزدیک‌تر باشد مدل عملکرد بهتری داشته است.

۱.۴ ارزیابی بر مبنای log perplexity

اگر معیار Average log likelihood را به اختصار با ALLL نشان دهیم:

$$ALLL(W) = \frac{۱}{N}\sum_{i=۱}^{N}\log P(w_i)$$

که $W = w_۱w_۲\cdots w_N$ سند حاصل از الصاق [۱] جملات آزمایشی است و w_iها نشان‌دهنده کلمات هستند و N تعداد کلمات است.

$$\begin{aligned}\log perplexity &= \log \sqrt[N]{\frac{۱}{P(w_۱w_۲\cdots w_N)}} \\ &= -\frac{۱}{N}\log\prod_{i=۱}^{N}P(w_i) \\ &= -\frac{۱}{N}\sum_{i=۱}^{N}\log P(w_i) = -ALLL(W)\end{aligned}$$

بنابراین قرینه مقادیر جدول ۱ مقدار log perplexity را به ما می‌دهد.

[1] Concatenate