به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس جستجو و بازیابی اطلاعات در وب

استاد ممتازی

تمرین اول

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

پیش‌پردازش

برای پیش‌پردازش علائم نگارشی تماما حذف شدند و حروف کوچک و بزرگ یکسان‌سازی شدند. با توجه به محدودیتی که در استفاده از کتابخانه‌ها وجود داشت پیش‌پردازش‌هایی نظیر Lemmatization را به ناچار کنار گذاشتم. در عین حال ایده حذف ایست‌واژه‌ها (Stop Words) را به صورت ضمنی برای مدل TF-IDF استفاده کردیم که در قسمت مربوط به آن بیان می‌شود.

نهایتا توجه کنید که نحوه توکن‌کردن یک متن به کلمات آن با استفاده از کاراکتر فاصله (Space) انجام می‌گیرد.

بخش دوم – بازیابی با استفاده از مدل فضای برداری

برای انتخاب کلمات مناسب از DF استفاده کردیم؛ یعنی بررسی کردیم که هر کلمه در چه بخشی از اسناد آمده است. کلماتی که DF زیادی دارند جز ایست‌واژه‌ها هستند و کلمات با DF کم جز کلمات نادر (Rare Words) به حساب می‌آیند. با بررسی کلی به نظر می‌آید که حذف صد کلمه با بیشترین DF به عنوان ایست‌واژه منطقی باشد و سپس دو هزار کلمه بعدی مناسب‌ترین کلمات خواهند بود.

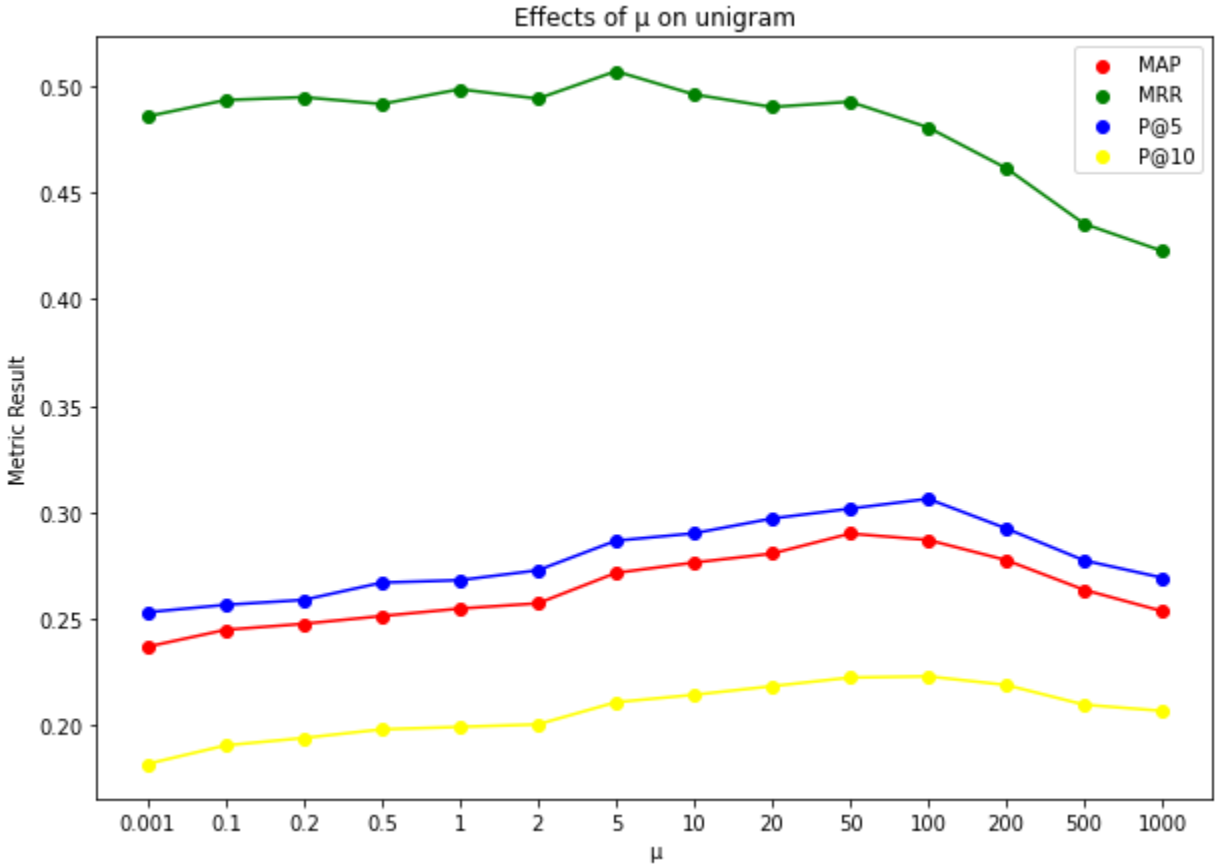
بخش سوم – بازیابی سندها برپایه مدل زبانی

پیش از هر چیز باید نکات مهم در مورد پیاده‌سازی این قسمت را بیان کنم:

- برای نگهداری مناسب‌تر احتمالات و ضربشان از لگاریتم احتمالات استفاده کردم و سعی کردم لگاریتم ضرب احتمالات که برابر با جمع لگاریتم هر احتمال می‌شود را بیشینه کنم. طبیعتاً سندی که بیشترین لگاریتم احتمال را داشته باشد، بیشترین احتمال را هم خواهد داشت.
- برای پیدا کردن مقدار بهینه سعی شده است که چهار معیار دقت معرفی شده در بخش چهارم تمرین را بر روی مجموعه اعتبارسنجی (validation) بیشینه کنم.

- برای محاسبه bigram یک توکن شروع به ابتدای توکن‌ها اضافه کردم تا محاسبه احتمال اولین توکن واقعی جمله مانند سایر توکن‌های آن باشد.
- در bigram اگر اولین توکن یک جفت توکن دیده نشده باشد، طبیعتاً قسمت bigram آن تعریف نشده خواهد بود. برای حل این مشکل و به سادگی احتمال صفر را برای آن در نظر گرفتم. طبیعتاً به واسطه قسمت unigram یک احتمال غیر صفر برای احتمال نهایی بدست خواهد آمد.

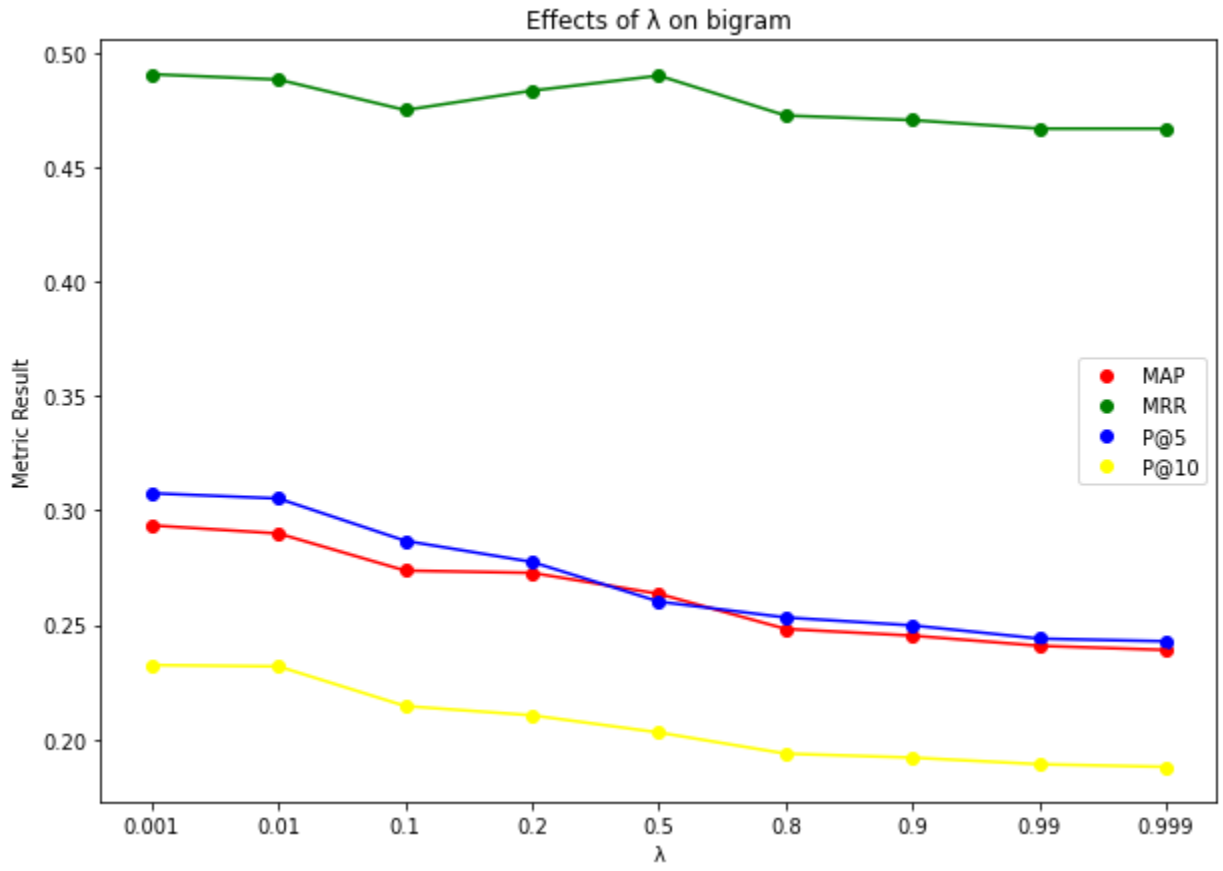
برای یافتن مقدار مناسب ابرپارامتر μ لیستی از مقادیر ۰.۰۰۱، ۰.۱، ۰.۲، ۰.۵، ۱، ۲، ۵، ۱۰، ۲۰، ۵۰، ۱۰۰، ۲۰۰، ۵۰۰ و ۱۰۰۰ مورد بررسی قرار گرفت. نتایج این آزمایش در تصویر ۱ آورده شده است.



تصویر ۱ - تاثیر تغییر μ بر دقت‌های unigram

مطابق تصویر ۱ در دو معیار P@n مقدار ۱۰۰، برای معیار MAP مقدار ۵۰ و برای MRR مقدار ۵ بهترین بوده است. در مجموع به نظر می‌رسد انتخاب ۵۰ به عنوان عدد نهایی گزینه‌ای باشد که هر سه معیار را در حد معقول و مناسبی نگه دارد چراکه دو انتخاب دیگر هر کدام باعث کاهش قابل توجه یک معیار می‌شود.

حال با تثبیت ۵۰ برای مقدار μ تلاش کردم تا مقدار بهینه ابرپارامتر λ را بدست بیاورم. برای تعیین مقدار λ اعداد ۰.۰۰۱، ۰.۰۱، ۰.۱، ۰.۲، ۰.۵، ۰.۸، ۰.۹، ۰.۹۹ و ۰.۹۹۹ را بررسی کردم که نتایج آن برای معیارهای دقت و مجموعه اعتبارسنجی در تصویر ۲ آورده شده است.



تصویر ۲ - تاثیر تغییر λ بر دقت‌های bigram

مطابق نتایج تصویر ۲ به نظر می‌رسد به طور کلی هر چقدر λ کم‌تر باشد معیارهای ارزیابی نتایج بهتری را نشان می‌دهند. ما می‌دانیم که مقدار ابرپارامتر مذکور میزان اهمیت bigram را نشان می‌دهد. به طور عادی باید عدد کمی باشد تا احتمالات بیشتر بر مبنای bigram و کمتر بر مبنای unigram باشد به گونه‌ای که unigram برای کلمات کمتر دیده شده قادر به تولید احتمال باشد. اما چنین چیزی رخ نداده است و نشان می‌دهد که مدل bigram ما چندان مناسب نیست و هر چه کمتر تاثیر بگذارد بهتر است.

بخش چهارم – کار با روش‌های ارزیابی

پس از پیاده‌سازی معیارهای ارزیابی دقت مدل‌های مختلف بدست آمد و در جدول ۱ آورده شده است.

جدول ۱ – دقت‌های معیار ارزیابی مدل‌ها

	P@5	P@10	MAP	MRR
TF-IDF	۴۴.۵۲٪	۳۱.۷۱٪	۳۹.۴۸٪	۶۲.۷۷٪
Unigram	۳۹.۸۶٪	۲۸.۴۲٪	۳۷.۵۵٪	۶۱.۰۵٪
Bigram	۴۰.۲۷	۲۸.۱۵٪	۳۶.۷۲٪	۶۰.۷۵٪

مطابق اعداد حاصل‌شده می‌توان دید که مدل TF-IDF عملکرد بهتری را نسبت به مدل‌های زبانی داشته است. از طرفی و با تنظیم $\lambda = 0.001$ عملا تفاوت دو مدل Unigram و Bigram چندان زیاد نخواهد بود.

همچنین می‌توان مقایسه‌ای را نیز بین خود معیارهای ارزیابی انجام دهیم. معیار MRR اعداد بالای نیم را نشان می‌دهد که بیانگر آن است که به طور میانگین در جایگاه اول یا دوم یک جواب مرتبط می‌توانیم بیابیم. معیار P@5 از P@10 بهتر است که یک دلیل آن این است که برای هر داده تست تعداد نمونه‌های مثبت محدود است. با بررسی انجام شده به ازای هر کوئری تنها ۶.۷ نمونه مثبت وجود دارد. این امر باعث

می‌شود که بهترین مدل‌ها هم نتواند دقت بالا در P@10 بدست آورد ولی باز در P@5 دقت‌های بالاتر را بهتر می‌توان بدست آورد. معیار MAP هم از P@10 بالاتر است که علت آن هوشمندی این معیار در در نظر گرفتن تعداد نمونه مثبت است.

در جدول ۲ زمان‌های اجرای مربوط به مدل‌های مختلف به ثانیه آورده شده است. توجه کنید که زمان آزمون شامل اجرای مستقل هر چهار معیار بر روی تمام دادگان آزمون است با در نظر گرفتن این مورد می‌توان گفت که زمان آزمون تمام مدل بر روی کل دادگان آزمون بین نیم تا یک دقیقه است. مدل TF-IDF زمان خیلی بیشتری را نسبت به مدل‌های زبانی دارد. این مسئله با توجه به محاسبات مختلف و ایجاد بردارها منطقی به نظر می‌رسد. از طرفی مدل Bigram هموارسازی‌شده نزدیک به دو برابر Unigram زمان نیاز دارد این مورد نیز همچنان منطقی است.

جدول ۲ – زمان اجرای مدل‌ها

	زمان آموزش (s)	زمان آزمون (s)
TF-IDF	۳.۴۵	۲۰۰.۶۹
Unigram	۰.۵۷	۹۱.۱
Bigram	۰.۹۱	۱۶۵.۴۵