بازیابی هوشمند اطلاعات

تمرین اول

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

آذر ۱۴۰۰

فهرست

۱- هدف تمرین

- شاخص‌گذاری تمامی اسناد
- به کارگیری و آشنایی با توابع بازیابی موجود
- استفاده از معیارهای ارزیابی و گزارش کارایی توابع ارزیابی

۲- شاخص‌گذاری

شاخص‌گذاری فرآیندی مهم در سیستم های بازیابی اطلاعات است و کارکرد اصلی آن را تشکیل می دهد زیرا اولین گام در بازیابی اطلاعات است و به بازیابی کارآمد اطلاعات کمک می کند. شاخص‌گذاری اسناد را به ترم‌های حاوی اطلاعات مفید موجود در آنها کاهش می دهد.

```
galago build index-parameters.json
```

پس از تبدیل دستی پسوند فایل Documents.txt به trectext ، با استفاده از دستور بالا و تنظیمات زیر برای index-parameters.json (که می‌توانیم آن‌ها را بدون استفاده از json در ترمینال نیز مقداردهی کنیم) ، آن را شاخص گذاری می‌کنیم.

```
    "fileType" : "trectext" ,
    "inputPath" : "/home/sadoldman/Desktop/CA1-
Resources/Corpus/corpus/Documents.trectext" ,
    "indexPath" : "/home/sadoldman/Desktop/CA1-Resources/Indexed" ,
    "stemmer" : ["porter"] ,
    "tokenizer" : {
        "fields" : ["TEXT" , "HEAD" , "DOCNO"] ,
        "format" : {

            "TEXT" : "string" ,
            "HEAD" : "string" ,
            "DOCNO" : "string"
        }
    }
```

در ادامه بخش‌های مهم این فایل کانفیگ تشریح می‌شود.

۲٫۱- fileType

گالاگو معمولا به صورت خودکار نوع فایل ورودی را به کمک بررسی ۱۰۰ خط ابتدایی فایل تشخیص می‌دهد اما گاهی مشخص کردن نوع فایل ورودی به صورت صریح ساده‌تر است. در این تمرین نوع فایل ورودی را trectext انتخاب می‌کنیم.

۲٫۲- Porter Stemmer

ریشه‌یابی فرآیند حذف بخشی از یک کلمه، یا کاهش یک کلمه به ریشه آن است اما لزوما به معنی نیست که ما کلمات را به ریشه لغوی آنها کاهش میدهیم. یکی از این الگوریتم‌ها الگوریتم ریشه‌یابی پورتر یا Porter stemmer نام دارد. بر اساس ایده‌ی این

الگوریتم ، پسوندها در زبان انگلیسی از ترکیب پسوندهای کوچکتر و ساده‌ترساخته می‌شوند. یکی از مزایای این روش سرعت بالای آن است.

۲٫۳- Tokenizer

توکن سازی روشی برای جدا کردن یک متن به واحدهای کوچکتری به نام توکن است. در تمرین ما توکن‌ها Head ، Text و DOCNO در نظر گرفته شده‌اند و در بخش فرمت هر سه توکن از جنس رشته(string) تعریف شده‌اند.

۳- سوال ۱

در این سوال روش بازیابی BM25 و نسخه‌های تغییر یافته آن بررسی و ارزیابی می‌شوند.

۳.۱- بخش ۱ : روش بازیابی BM25

در بخش ۱ سوال ۱ روش بازیابی BM25 بررسی می‌شود.

۳.۱.۱- الف

از ما خواسته شده که مقدار توابع ارزیابی Precision at 5 ، MAP ، ndcg و recall را برای مقادیر مختف پارامترهای b و k را بیابیم و b و k با بهترین نتایج را انتخاب کنیم.

با توجه به بازه مجاز b (بین ۰ و ۱) و k (بین ۰ و بی‌نهایت) ابتدا گام‌هایی نسبتا بلند برای یافتن بازه حدودی این پارامترها برمیداریم. این گام‌های بلند برای b برابر 0.25 و برای k برابر 2.5 می‌باشد.

فایلی از نوع json به منظور قرار دادن تنظیمات لازم برای انجام جستجو به شکل زیر ایجاد می‌شود. مقادیر b و k در هر فایل و جستجو متفاوت است و در بخش Queries کوئری‌های ۱۰۱ تا ۱۵۰ قرار می‌گیرند.

```
"index" : "/home/sadoldman/Desktop/CA1-Resources/Indexed" ,
"scorer" : "bm25" ,
"showNoResults" : true ,
"b" : "different-values",
"k": "different-values" ,
"verbose" : true ,
"queries" : [

]
```

مقدار اختصاص یافته به scorer روش بازیابی مورد نظر را مشخص می‌کند و مقدار اختصاص یافته به shownoresults مشخص می‌کند که برای کوئری‌های بدون سند مرتبط dummy در نظر بگیرد.

این روش‌ها به وسیله خط زیر در ترمینال به اجرا در می‌آیند و نتیجه آنها در فایلی که نام آن پس از علامت > ذکر شده است ، ذخیر می‌شود.

```
galago batch-search configure.json > results.txt
```

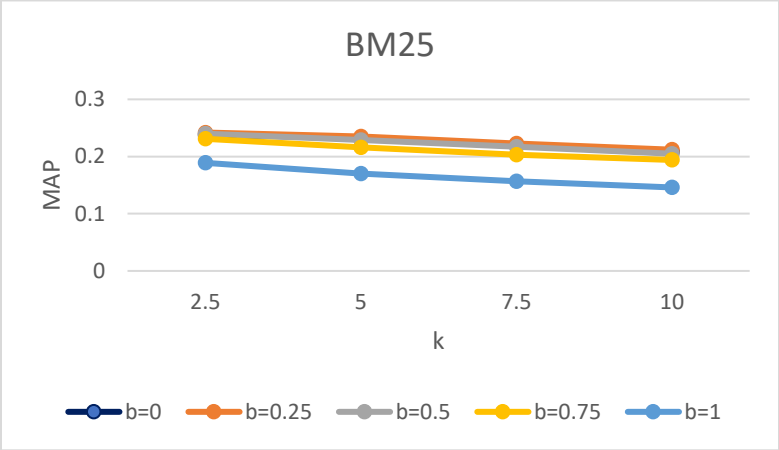
پس از به دست آوردن نتایج مرتبط با هر کدام از مقادیر پارامتر در فایل های جدا ، با استفاده از کانفیگ زیر و دستور پس از آن و با توجه به دادگان طلایی ، معیارهای ارزیابی را برای هر کدام محاسبه می‌کنیم.(در قسمت runs آدرس فایل‌های نتایج قرار می‌گیرد)

```
    "judgments" : "/home/sadoldman/Desktop/CA1-
Resources/Relevance/relevance_judgment.txt" ,
    "runs" : [] ,
    "metrics" : [ "num_rel_ret" , "num_rel" , "p5" , "ndcg" , "map"]
```

```
galago eval configure.json > eval.txt
```

مقدار k از ۰ تا ۱۰ با گام‌های ۲.۵ واحدی و مقدار b از ۰ تا ۱ با گام‌های ۰.۲۵ واحدی حرکت می‌کند.

هنگامی که مقدار k برابر صفر باشد بدترین عملکرد روش بازیابی را شاهد خواهیم بود.در این مقدار برای k هیچ اهمیتی ندارد که چه مقداری برای b انتخاب کنیم و در هر حالت مقادیر تابع ارزیابی نسبت به دیگر مقادیر k ارزش بسیار کمتری را نشان می‌دهد.

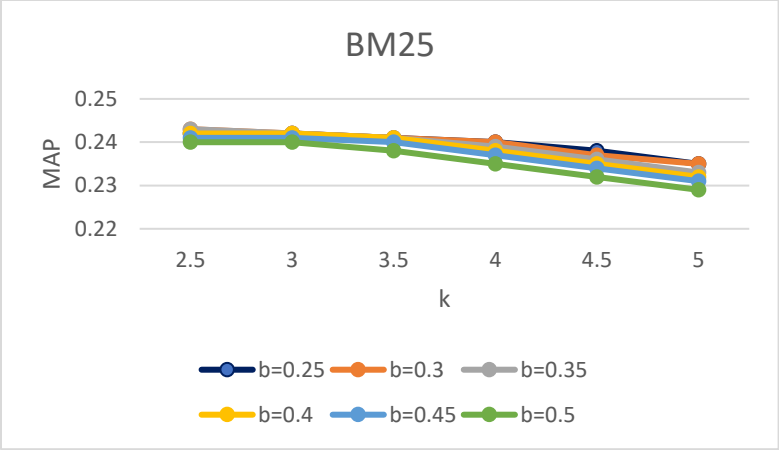


نمودار بالا معیار ارزیابی Map را در مقادیر b و k متفاوت نشان می‌دهد. به راحتی قابل مشاهده است هنگامی که مقدار k بین ۲.۵ و ۵ باشد ، BM25 در بهترین حالت ممکن قرار دارد. همچنین قابل تشخیص است که نمودارهای b=2.5 و b=5 در بهترین حالات map قرار دارند.(برای سایر معیارهای ارزیابی نیز نتیجه مشابهی به دست می‌آید)

با توجه به نمودار بالا تصمیم می‌گیریم که برای پارامتر k بازه [۵ و ۲.۵] با گام‌های ۰.۵ و برای پارامتر b بازه [۰.۵ و ۰.۲۵] با گام‌های ۰.۰۵ را مورد بررسی قرار دهیم و بهترین مقادیر آن‌ها را بیابیم.

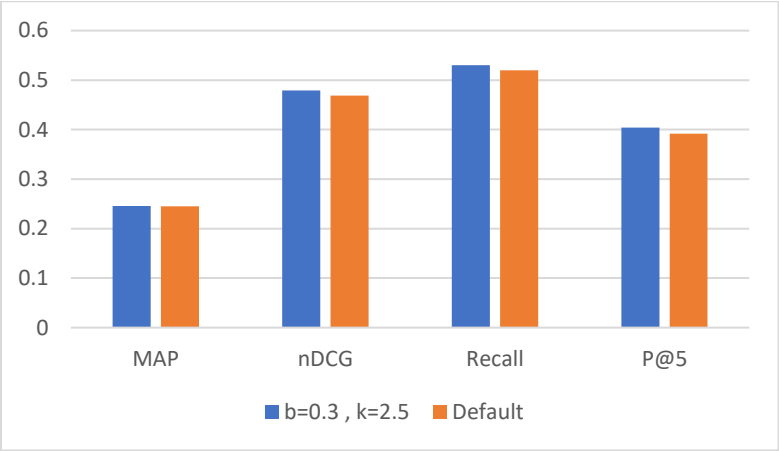
با استفاده از تنظیماتی که قبلا گفته شد (و البته مقادیر b و k متفاوت) و با اجرای دستور batch-search و سپس eval نتایج ارزیابی را برای روش BM25 و مقادیر پارامترهای متفاوت به دست می‌آوریم.

با توجه به نمودار زیر بهترین مقدار MAP به ازای k=2.5 و b=0.35 به دست می‌آید اما با توجه به دیگر معیارهای ارزیابی از جمله ndcg و Recall ، در مجموع به ازای k=2.5 و b=0.3 روش بازیابی BM25عملکرد بهتری از خود نشان می‌دهد. در این نقطه بهترین مقدار Recall و همچنین مشترکا با نقطه k=2.5 و b=0.35 بهترین MAP را داراست. همچنین مقادیر ndcg و precision at 5 نیز برای این نقطه جزو بالاترین مقادیر به ازای کل b و k است.



۳.۱.۲- ب

با تنظیمات گفته شده در بخش‌های قبل ، خروجی پرس‌وجوهای ۵۱ تا ۱۰۰ را یک بار با مقادیر b و k به دست آمده و یک بار با مقادیر پیشفرض به دست می‌آوریم و سپس مقدار معیارهای ارزیابی را به دست می‌آوریم. نتیجه به شکل زیر است:



تمام معیارهای ارزیابی برتری مقادیر به دست آمده برای b و k را نسبت به مقادیر پیشفرض نشان می‌دهند.

به طور خلاصه مقادیر به دست برای b و k :

- اسناد مرتبط بیشتری را بازیابی می‌کند.(Recall)
- در ۵ رتبه اول اسناد بازیابی شده ، اسناد مرتبط بیشتری را نشان می‌دهد.(p@5) این اختلاف بسیار کم است اما برتری محسوسی به چشم می‌خورد.
- میانگین دقت بهتری برای پرس‌وجوهای متفاوت ارائه می‌دهد.(MAP)
- مقدار ndcg بهتری نسبت به حالت پیشفرض دارد.

۳٫۲- روش پیشنهادی اول

با استفاده از این روش ۴۵۴۰۰ سند بازیابی شدند که ۲۷۹۴ عدد از آنها مرتبط بودند.(مجموع اسناد مرتبط برابر ۶۱۰۰ است)
در این روش recall=0.458 ، ndcg=0.384 ، p@5=0.261 و MAP=0.153 می‌باشد.

۳٫۳- روش پیشنهادی دوم

با استفاده از این روش ۴۵۴۰۰ سند بازیابی شدند که ۲۹۲۵ عدد از آنها مرتبط بودند.(مجموع اسناد مرتبط برابر ۶۱۰۰ است)
در این روش recall=0.479 ، ndcg=0.438 ، p@5=0.408 و MAP=0.212 می‌باشد.

۳٫۴- روش پیشنهادی سوم

با استفاده از این روش ۴۵۴۰۰ سند بازیابی شدند که ۲۵۱۰ عدد از آنها مرتبط بودند.(مجموع اسناد مرتبط برابر ۶۱۰۰ است)
در این روش recall=0.411 ، ndcg=0.346 ، p@5=0.249 و MAP=0.130 می‌باشد.

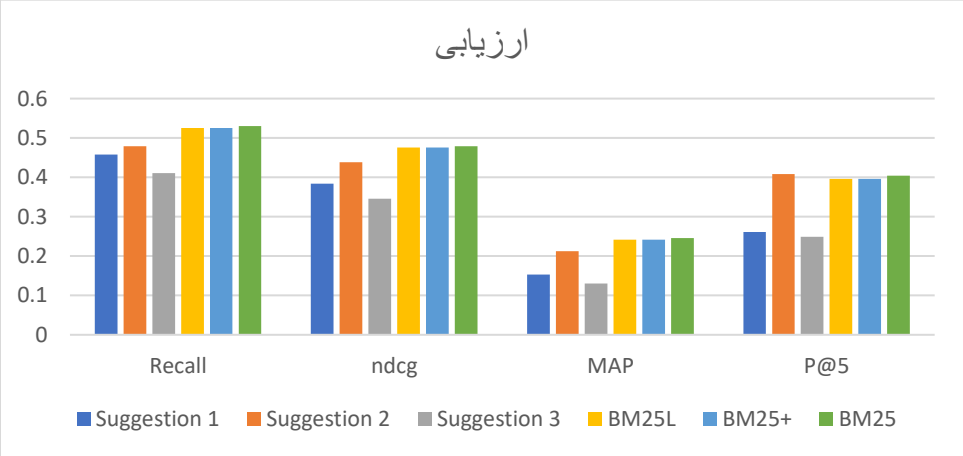
۳٫۵- روش پیشنهادی چهارم(BM25L)

با استفاده از این روش ۴۵۴۰۰ سند بازیابی شدند که ۳۲۰۴ عدد از آنها مرتبط بودند.(مجموع اسناد مرتبط برابر ۶۱۰۰ است)
در این روش recall=0.525 ، ndcg=0.476 ، p@5=0.396 و MAP=0.242 می‌باشد.

۳٫۶- روش پیشنهادی پنجم(BM25+)

با استفاده از این روش ۴۵۴۰۰ سند بازیابی شدند که ۳۲۰۴ عدد از آنها مرتبط بودند.(مجموع اسناد مرتبط برابر ۶۱۰۰ است)
در این روش مقادیر متفاوت دلتا در فرمول قرار گرفتند و نتایج معیارهای ارزیابی برای آنان کاملا یکسان بود. مقادیر معیارهای ارزیابی بدین شکل است. recall=0.525 ، ndcg=0.476 ، p@5=0.396 و MAP=0.242 می‌باشد.

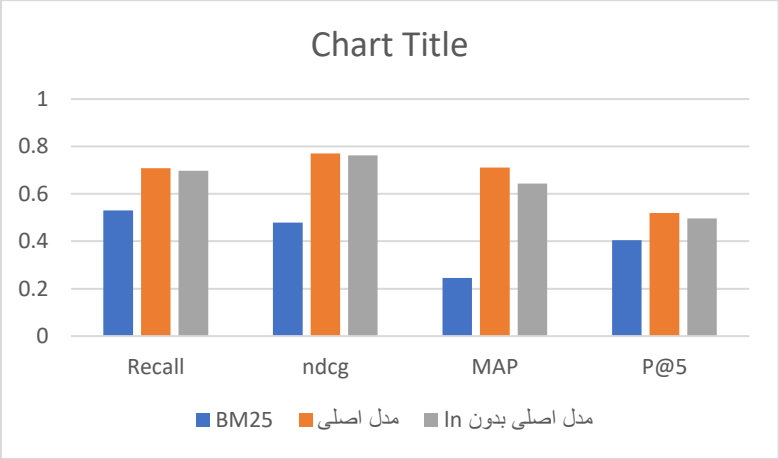
۳٫۷- مقایسه تمام روش‌های پیشنهادی و BM25



نتیجه ارزیابی بازیابی اسناد مرتبط با پرس‌وجوهای ۵۱ تا ۱۰۰ با استفاده از روش‌های ۱ تا ۶ به شکل فوق است. مشاهده می‌شود که سه مورد از معیارهای ارزیابی(MAP ، ndcg و Recall) به BM25 به عنوان بهترین روش بازیابی بین روش‌های فوق اشاره

می‌کنند و پس از آن با اختلاف کمی BM25L و BM25+ بهترین عملکرد را دارند.
در روش P@5 رتبه بندی کمی متفاوت است و بهترین عملکرد را روش پیشنهادی ۲ دارد.با توجه به عدم وجود مولفه نرمال سازی طول اسناد در این روش دلیل برتری روش پیشنهادی ۲ می‌تواند برابری تقریبی طول اسناد مورد جستجو باشد که باعث شده اسناد مرتبط بیشتری در بین ۵ سند اول بازیابی شده توسط آن باشد.
جالب توجه است که روش‌های BM25L و BM25+ در هیچکدام از روش های ارزیابی تفاوتی ندارند که دلیل آن می‌تواند این باشد که پارامتر دلتا اضافه شده در روش BM25+ ، تحت تاثیر هیچ پارامتر دیگری نیست و مستقیما و بدون انجام هیچ محاسباتی روی آن ، به مقدار نهایی هر ترم اضافه می‌شود.

۴- سوال ۲



با توجه به شکل بالا مدل اصلی و مدل اصلی بدون ln تو در تو برتری واضحی نسبت به BM25 دارند و در بین این دو روش مدل اصلی اندکی بهتر عمل می‌کند.