# حانیه ناصری - 810101286

- ببخشید..من در لحظات اخر آپلود متوجه شدم که انگار حتی قسمت تپوری هم باید تایپ می شد.قسمت سوالات تئوری را پی دی اف دست نویس آپلود کردم. لطفا این یکبار را نمره کسر نکنید. واقعا ممنون میشم.

: Porter stemmer

یک روش برای به دست آوردن ریشه کلمات میباشد

:Trectest

یک فرمت که برای داکیومنت ها استفاده می شود و با تگ قسمت ها را از هم جدا می کند.

:Tokenizer

بخش هایی از فایل که میخواهیم توکن بندی شود را مشخص میکند.

در تمامی گزارش های زیر سرچ را با requested=50 انجام می دهیم.

در سوال 1 قسمت ب در مقایسه روش های پیشنهادی،
روش پیشنهادی i ام = bm25_i

در سوال 2، روش مدل اصلی (bm25_plan_1) و روش بدون مولفه تو در تو (bm25_plan_2) می باشد.

## سوال 1

الف)

## Which b is best? = 0.8

## bm25 b=0.2 k=0.7

| | | |
|---|---|---|
| num_rel_ret | all | 9666.00000 |
| map | all | 0.40799 |
| ndcg | all | 0.59642 |
| P5 | all | 0.67067 |

# bm25 b=0.5 k=0.7

| | | |
|---|---|---|
| num_rel_ret | | all 9672.00000 |
| map | all | 0.40819 |
| ndcg | all | 0.59646 |
| P5 | all | 0.67107 |

# bm25 b=0.8 k=0.7

| | | |
|---|---|---|
| num_rel_ret | | all 9678.00000 |
| map | all | 0.40823 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67127 |

K را ثابت می گیریم و b را در 3 مقدار 0.2 و 0.5 و 0.8 تست میکنیم. در b بزرگتر (0.8) نتیجه بهتر شد. (تاثیر طول اسناد در جست و جو کم نیست!)

## Which k is best? 0.65

# bm25 b=0.8 k=1

| | | |
|---|---|---|
| num_rel_ret | | all 9662.00000 |
| map | all | 0.40546 |
| ndcg | all | 0.59531 |
| P5 | all | 0.66727 |

# bm25 b=0.8 k=1.2

| | | |
|---|---|---|
| num_rel_ret | | all 9631.00000 |
| map | all | 0.40251 |
| ndcg | all | 0.59420 |
| P5 | all | 0.66306 |

## bm25 b=0.8 k=0.7

| | | |
|---|---|---|
| num_rel_ret | all | 9678.00000 |
| map | all | 0.40823 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67127 |

## bm25 b=0.8 k=0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## bm25 b=0.8 k=0.6

| | | |
|---|---|---|
| num_rel_ret | all | 9674.00000 |
| map | all | 0.40877 |
| ndcg | all | 0.59599 |
| P5 | all | 0.67207 |

## bm25 b=0.8 k=0.55

| | | |
|---|---|---|
| num_rel_ret | all | 9674.00000 |
| map | all | 0.40877 |
| ndcg | all | 0.59582 |
| P5 | all | 0.67167 |

با ثابت گیری b در مقدار 0.8 که در قسمت قبل آن را به دست آوردیم k را با مقادیر 1.2 و 1 و 0.7 و 0.65 و 0.6 و 0.55 تست کردیم. معیار های ارزیابی با تغییر آن از 1.2 تا 0.65 بهتر شد. 0.6 ریکال کمتری نسبت به 0.65 داشت و در k=0.55 هم ریکال نسبت به 0.65 کمتر شد و معیارهای ارزیابی دیگر تغییر خاصی نکردند و بهتر نشدند. پس 0.65 را برگزیدیم.

به نظر میرسد ارزش پارامتر تعداد کلمات مشترک یک سند و کوئری خیلی زیاد نیست و همان تعداد واقعی کلمه (بدون ضرب شدن در k بزرگ) نسبت به judgement های ارائه شده منطقی تر است. (احتمالا تعداد خوبی از داکیومنت هایی که به عنوان مرتبط تشخیص داده شده اند لزوما به علت تعداد کلمه بیشتر مرتبط نبوده اند و عوامل دیگری نیز دخیل بوده است!)

ب)

مقایسه روش های زیر را با یکدیگر و با bm25 با همان b و k برگزیده از مرحله قبل (b = 0.8 و k = 0.65) انجام می دهیم.

**روش پیشنهادی اول (idf)**

## BM25_1 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9678.00000 |
| map | all | 0.40854 |
| ndcg | all | 0.59580 |
| P5 | all | 0.67247 |

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25 b = 0.8 k = 1

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40546 |
| ndcg | all | 0.59531 |
| P5 | all | 0.66727 |

## BM25_1 b = 0.8 k = 1

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40538 |
| ndcg | all | 0.59481 |
| P5 | all | 0.66727 |

میبینیم که idf کمی بدتر (نه خیلی!) عمل کرده است. اینکه بدتر عمل کرده است دلیلش واضح است. تاثیر فرکانس کلمات و طول داکیومنت در نظر گرفته نشده است. اما احتمالا داکیومنت هایی که در judgement مرتبط تشخیص داده شده اند در کلمات مشترک با کوئری idf بالایی داشتند و بالعکس داکیومنت های غیر مرتبط idf شان در آن کلمات بالا نبوده یا به عبارتی اکثر کلمات کوئری ها در داکیومنت مرتبط کم تکرار شده و در داکیومنت غیر مرتبط به دفعات بیشتر بوده و باعث شده idf علارغم بدتر عمل کردن خیلی هم غیر منطقی جواب ندهد و برتری نسبی داکیومنت مرتبط به غیر مرتبط با همین پارامتر حفظ شود.

**روش پیشنهادی دوم**

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_2 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 3779.00000 |
| map | all | 0.11927 |
| ndcg | all | 0.20563 |
| P5 | all | 0.21301 |

در روش دوم میبینیم که اگر تاثیر طول داکیومنت ها را حذف کنیم نتیجه خیلی بدتر میشود که شاهدی بر تاثیر طول سند در این مجموعه اسناد میباشد.

**روش پیشنهادی سوم**

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_3  b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9014.00000 |
| map | all | 0.37306 |
| ndcg | all | 0.56789 |
| P5 | all | 0.64484 |

در این روش هم تاثیر پارامتر b در تاثیر طول داکیومنت ها را میبینیم که در صورت حذف آن از فرمول نتیجه بدتر شده است.

**روش پیشنهادی چهارم**

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_4 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 7928.00000 |
| map | all | 0.29598 |
| ndcg | all | 0.49404 |
| P5 | all | 0.51211 |

در اینجا هم تاثیر تعداد کلمات هم تاثیر طول اسناد و هم idf در نظر گرفته نشده است و صرفا با حضور/عدم حضور کلمه پرسش درسند قضاوت شده است که خوب نیست. (به طور مثال سندی که تعداد بیشتری از آن کلمه را دارد احتمالا مرتبط تر باشد که در نظر گرفته نمیشود)

**روش پیشنهادی پنجم**

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |

| | | |
|---|---|---|
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_5 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9639.00000 |
| map | all | 0.40354 |
| ndcg | all | 0.59453 |
| P5 | all | 0.66507 |

در این روش هم نتیجه اندکی نسبت به فرمول اصلی بدتر شده است. در این فرمول تاثیر تعداد کلمات و طول سند حذف نشده است اما هردو در صورت و مخرج تاثیر داده شده. از طرفی ضریب k برخلاف فرمول اصلی پشت عبارت شامل ضریب b ضرب نشده. در کل نتیجه تقریبا نزدیک با فرمول اصلی شده.

**روش پیشنهادی ششم**

## BM25_6 b = 0.8 k = 0.65 e=0.1

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40854 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.2

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40854 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.3

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.4

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40854 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.5

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.6

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.7

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40857 |
| ndcg | all | 0.59632 |
| P5 | all | 0.67187 |

## BM25_6 b = 0.8 k = 0.65 e=0.8

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 0.65 e=0.9

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40857 |
| ndcg | all | 0.59632 |
| P5 | all | 0.67187 |

## BM25_6 b = 0.8 k = 0.65 e=1

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40854 |
| ndcg | all | 0.59631 |
| P5 | all | 0.67167 |

## BM25_6 b = 0.8 k = 1 e=0.1

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40545 |
| ndcg | all | 0.59530 |
| P5 | all | 0.66727 |

## BM25_6 b = 0.8 k = 1 e=0.2

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40546 |
| ndcg | all | 0.59531 |
| P5 | all | 0.66727 |

## BM25_6 b = 0.8 k = 1 e=0.3

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40545 |
| ndcg | all | 0.59530 |
| ndcg5 | all | 0.67137 |
| P5 | all | 0.66727 |

## BM25_6 b = 0.8 k = 1 e=0.4

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40545 |
| ndcg | all | 0.59530 |
| P5 | all | 0.66727 |

## BM25_6 b = 0.8 k = 1 e=0.5

| | | |
|---|---|---|
| num_rel_ret | all | 9662.00000 |
| map | all | 0.40546 |
| ndcg | all | 0.59531 |

P5 all 0.66727

## BM25_6 b = 0.8 k = 1 e=0.6

num_rel_ret all 9662.00000
map all 0.40545
ndcg all 0.59530
P5 all 0.66727

## BM25_6 b = 0.8 k = 1 e=0.7

num_rel_ret all 9662.00000
map all 0.40546
ndcg all 0.59531
P5 all 0.66727

## BM25_6 b = 0.8 k = 1 e=0.8

num_rel_ret all 9662.00000
map all 0.40546
ndcg all 0.59531
P5 all 0.66727

## BM25_6 b = 0.8 k = 1 e=0.9

num_rel_ret all 9662.00000
map all 0.40545
ndcg all 0.59530
P5 all 0.66727

## BM25_6 b = 0.8 k = 1 e=1

num_rel_ret all 9662.00000
map all 0.40546
ndcg all 0.59531
P5 all 0.66727

فرمول را برای ابلیسون های 0.1 تا 1 و به ازای ۲ k برابر با 0.65 و 1 بررسی کردیم. در مقایسه نتایج با k ثابت خیلی در معیار های ارزیابی تغییر نداشتیم و در مقایسه نتایج با k = 0.65 و k = 1 معیار های ارزیابی با k = 0.65 بهتر هستند.

سوال ۲)

الف )

**مدل اصلی**

## BM25 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 9679.00000 |
| map | all | 0.40853 |
| ndcg | all | 0.59630 |
| P5 | all | 0.67167 |

## BM25_plan_1 b = 0.8 k = 0.65

| | | |
|---|---|---|
| num_rel_ret | all | 7109.00000 |
| map | all | 0.23416 |
| ndcg | all | 0.41991 |
| P5 | all | 0.42102 |

## BM25 b = 0.75 k = 1.2 (default)

| | | |
|---|---|---|
| num_rel_ret | all | 9632.00000 |
| map | all | 0.40258 |
| ndcg | all | 0.59424 |
| P5 | all | 0.66306 |

## BM25_plan_1 b = 0.75 k = 1.2 (default)

| | | |
|---|---|---|
| num_rel_ret | all | 7111.00000 |
| map | all | 0.23416 |
| ndcg | all | 0.41993 |
| P5 | all | 0.42122 |

**در مقایسه Bm25 و مدل اصلی:**

## Bm25 >>>> bm25_plan_1 (مدل اصلی)

## BM25_plan_1  b = 0.8 k = 0.65

num_rel_ret all 7109.00000
map all 0.23416
ndcg all 0.41991
P5 all 0.42102

## BM25_6 (BM25+)  b = 0.8 k = 0.65 e=0.5

num_rel_ret all 9679.00000
map all 0.40853
ndcg all 0.59630
P5 all 0.67167

## BM25_plan_1 b = 0.75 k = 1.2

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

## BM25_6 (BM25+) b = 0.75 k = 1.2 e=0.5

num_rel_ret all 9632.00000
map all 0.40257
ndcg all 0.59423
P5 all 0.66306

**در مقایسه Bm25+ و مدل اصلی:**

## BM25+ >>>> BM25_PLN_1 (مدل اصلی)

## BM25_plan_1 b = 0.75 k = 1.2 (default)

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

## BM25_plan_1 b = 0.75 k = 1

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

## BM25_plan_1 b = 0.75 k = 0.8

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

این فرمول وابسته به پارامتر k نمی باشد و با افزایش/کاهش آن معیار های ارزیابی تغییر نمیکنند.

## BM25_plan_1 b = 0.25 k = 0.65

num_rel_ret all 7102.00000
map all 0.23434
ndcg all 0.42037
P5 all 0.42222

## BM25_plan_1 b = 0.5 k = 0.65

num_rel_ret all 7114.00000
map all 0.23451
ndcg all 0.42060
P5 all 0.42182

## BM25_plan_1 b = 0.75 k = 0.65

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

این فرمول در b میانی (حدود ۰.۵) بهتر از b های کم یا زیاد عمل میکند.

ب )

**مدل بدون مولفه تو در تو**

## BM25 b = 0.8 k = 0.65

num_rel_ret all 9679.00000
map all 0.40853
ndcg all 0.59630
P5 all 0.67167

## BM25_plan_2 b = 0.8 k = 0.65

num_rel_ret all 6889.00000
map all 0.22345
ndcg all 0.40195
P5 all 0.40440

## BM25 b = 0.75 k = 1.2

num_rel_ret all 9632.00000
map all 0.40258
ndcg all 0.59424
P5 all 0.66306

## BM25_plan_2 b = 0.75 k = 1.2

num_rel_ret all 6885.00000
map all 0.22348
ndcg all 0.40197
P5 all 0.40420

## **BM25 >>> BM25_PLN_2**

## BM25_plan_1 b = 0.75 k = 1.2

num_rel_ret all 7111.00000
map all 0.23416
ndcg all 0.41993
P5 all 0.42122

## BM25_plan_2 b = 0.75 k = 1.2

num_rel_ret all 6885.00000
map all 0.22348
ndcg all 0.40197

P5 all 0.40420

## BM25_plan_1 b = 0.8 k = 0.65

num_rel_ret all 7109.00000
map all 0.23416
ndcg all 0.41991
P5 all 0.42102

## BM25_plan_2 b = 0.8 k = 0.65

num_rel_ret all 6889.00000
map all 0.22345
ndcg all 0.40195
P5 all 0.40440

**(مدل بدون تو در تو) BM25_PLAN_2 <<< (مدل اصلی) BM25_PLAN_1**
**مدل اصلی تاثیر تعداد کلمات را به دلیل ۲ بار لگاریتم گرفتن کمتر می کند که انگار بهتر است (در قسمت قبل هم و برای BM25 دیدیم که برای K کوچکتر ۱ جواب بهتری می گیریم که در تایید علت ذکر شده است)**

## BM25_plan_2 b = 0.75 k = 1.2

num_rel_ret all 6885.00000
map all 0.22348
ndcg all 0.40197
P5 all 0.40420

## BM25_plan_2 b = 0.75 k = 1

num_rel_ret all 6885.00000
map all 0.22348
ndcg all 0.40197
P5 all 0.40420

## BM25_plan_2 b = 0.75 k = 0.8

| | | |
|---|---|---|
| num_rel_ret | all | 6885.00000 |
| map | all | 0.22348 |
| ndcg | all | 0.40197 |
| P5 | all | 0.40420 |

این فرمول وابسته به پارامتر k نمی باشد و با افزایش/کاهش آن معیار های ارزیابی تغییر نمیکنند.

## BM25_plan_2 b = 0.25 k = 1.2

| | | |
|---|---|---|
| num_rel_ret | all | 6875.00000 |
| map | all | 0.22360 |
| ndcg | all | 0.40209 |
| P5 | all | 0.40420 |

## BM25_plan_2 b = 0.5 k = 1.2

| | | |
|---|---|---|
| num_rel_ret | all | 6882.00000 |
| map | all | 0.22361 |
| ndcg | all | 0.40211 |
| P5 | all | 0.40400 |
| P10 | all | 0.35138 |

## BM25_plan_2 b = 0.75 k = 1.2

| | | |
|---|---|---|
| num_rel_ret | all | 6885.00000 |
| map | all | 0.22348 |
| ndcg | all | 0.40197 |
| P5 | all | 0.40420 |

الگو خاصی برای پارامتر b مشاهده نشد!