به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

محمد ناصری

۸۱۰۱۰۰۴۸۶

تمرین اول

آبان ۱۴۰۰

مقدمه

امروزه با توجه به افزایش حجم داده های متنی، موتورهای جستجو اصلی ترین ابزار جستجو در اینترنت محسوب میشوند. آنها با توجه به سوال ورودی کاربر و با استفاده از توابع بازیابی و میزان ارتباط یک سند با پرس وجو، امتیازی به سند تخصیص میدهند تا در نهایت اسناد بر اساس امتیازشان، رتبه بندی و نمایش داده شوند. در این تمرین، هدف آشنایی با معیارهای ارزیابی و توابع امتیازدهی به اسناد است. یک تابع امتیازدهی با توجه به میزان ارتباط یک سند با پرس وجو، امتیازی به سند تخصیص میدهد تا در نهایت اسناد براساس امتیازشان، رتبه بندی و نمایش داده شوند. در نهایت رتبه بندی حاصل عموما با رتبه بندی طلایی مقایسه شده و کارایی تابع بازیابی گزارش میگردد.

ابزار جستجوی متنی مورد استفاده در این تمرین گالاگو میباشد

اهداف تمرین

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

شرح دادگان

برای انجام این تمرین فایلهای زیر بر روی صفحه مربوط به درس قرار داده شده اند:

پیکره متنی(فایل اسناد):

این فایل مجموعهای از مقاله های خبری در قالب **TREC**می باشد. هر سند شامل چندین فیلد است:

OCNO:شناسه هر سند

Head:عنوان سند

Text:متن سند

فایل پرس و جوها :

این فایل شامل پرس و جوها میباشد.

فایل قضاوتهای مرتبط :

این فایل شامل قضاوتهای مرتبط میباشد. در مرحله نهایی جهت ارزیابی توابع بازیابی، نتایج بدست آمده با این قضاوتها مقایسه میشوند

فایل کدهای مرتبط:

این فایل شامل کدهای مرتبط با قسمت‌های مختلف سوالات تمرین میباشد

فایل‌های پاسخ بازیابی:

این فایل‌ها شامل پاسخ بازیابی‌های انجام شده مرتبط با قسمت‌های مختلف سوالات تمرین میباشد

فایل‌های لاگ:

این فایل‌ها شامل لاگ‌های مقایسه نتیجه‌های بازیابی اطلاعات بر اساس معیارهای ارزیابی میباشد.

سوال ۱- تابع بازیابی BM25

هدف از این سوال آشنایی با مولفه های روش BM25و تاثیر هر یک بر روی کیفیت رتبه بندی میباشد

- **معیارهای ارزیابی nDCG، MAP، Recallو P@5 میباشند**

سوالات:

1 -روش بازیابی BM25:

الف) دراین قسمت شما بایستی بازیابی را به روش BM25 انجام دهید و تاثیر پارامترهای kو b را بررسی کنید. مقادیر مختلف bو k را آزمایش کنید تا به مقداردهی بهینه برای پرس و جو های ۱۰۱-۱۵۰برسید. هنگام تفسیر مقادیر بهینه BM25 به تاثیر هر یک مولفه های تابع امتیازدهنده دقت کنید

$$f(q,d) = \sum_{w \in q \cap d} IDF(w) \frac{c(w,d)}{c(w,d)+k\left(1-b+b\frac{|d|}{avdl}\right)}$$

پاسخ

برای انجام اعمال جستجو و بازیابی اطلاعات روی داده ارائه شده لازم است ابتدا فایل مورد نظر شاخص‌گذاری شود تا بتوان بر روی آن پردازش انجام داد.

پس از عمل شاخص گذاری و استخراج پرس و جوهای ۱۰۱-۱۵۰ برای بدست آوردن مقادیر بهینه پارامترهای bو k بازیابی‌های مختلف با مقادیر مختلف بر روی داده‌ها صورت داده و نتایج آنها را توسط توابع ارزیابی مقایسه و امتیازدهی میکنیم تا به مقادیر بهینه دست پیدا کنیم.

(میدانیم مقادیر b در بازه $[0,1]$ و مقادیر k در بازه $[0, +\infty)$ قرار دارند. همچنین طبق مشاهدات مقادیر k برای اکثریت موارد در بازه $[0.5, 3]$ پاسخ بهتری میدهند هرچند که میتواند مقادیر بیشتری بگیرد)

{فایل‌های Q01P01A}

جدول مقادیر تست:

شماره بازیابی	مقدار b	مقدار K
bm25Test01	0.1	0.1
bm25Test02	1	0.1
bm25Test03	0.5	0.1
bm25Test04	0.5	1
bm25Test05	0.5	2
bm25Test06	0.5	3
bm25Test07	0.5	4
bm25Test08	0.5	10
bm25Test09	0.8	2
bm25Test10	0.3	2
bm25Tes11	0.4	2
bm25Test12	0.4	2.5
bm25Test13	0.4	2.8
bm25Test14	0.4	2.9
bm25Test15	0.4	2.7
bm25Test16	0.5	2.8
bm25Test17	0.3	2.8

روند انتخاب این مقادیر بر اساس پاسخ ارزیابی‌های انجام شده در هر مرحله بازیابی و مقایسه نتیجه با مراحل قبلی بوده است تا در هر قدم به مقادیر بهینه نزدیکتر شده و پس از یافتن مقادیر بهینه از صحیح بودن این مقادیر اطمینان حاصل شود.

در ادامه به بررسی و ارزیابی نتایج بازیابی بر اساس مقادیر ذکر شده در جدول بالا میپردازیم.

جدول نتیجه ارزیابی:

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5	Recall
01	5000	4805	971	0.143	0.289	0.352	0.202
02	5000	4805	828	0.122	0.283	0.280	0.172
03	5000	4805	962	0.139	0.289	0.328	0.200
04	5000	4805	937	0.142	0.295	0.336	0.195
05	5000	4805	936	0.144	0.299	0.332	0.194
06	5000	4805	897	0.138	0.293	0.344	0.186
07	5000	4805	864	0.133	0.285	0.332	0.179
08	5000	4805	730	0.111	0.254	0.300	0.151
09	5000	4805	1101	0.169	0.340	0.380	0.229
10	5000	4805	1174	0.180	0.345	0.392	0.244
11	5000	4805	1167	0.180	0.346	0.404	0.242
12	5000	4805	1168	0.182	0.348	0.420	0.243
13	5000	4805	1171	0.183	0.350	0.432	0.243
14	5000	4805	1164	0.182	0.348	0.428	0.242
15	5000	4805	1167	0.182	0.349	0.420	0.242
16	5000	4805	1158	0.180	0.349	0.420	0.240
17	5000	4805	1173	0.182	0.348	0.416	0.244

با توجه به جدول گذشته و همانطور که در جدول بالا مشاهده میشود، از آزمایش‌های ۱ تا ۳ نتیجه میشود که مقدار بهینه b یک مقدار میانه‌ای بوده و در ابتدا و انتهای بازه [0,1] نیست. پس از آن در مراحل ۴ تا ۸ آزمایش نتیجه میشود که مقدار بهینه k نیز در بازه [2,3] قرار دارد. در مراحل بعدی با آزمایش مقادیر مختلف در بازه‌های احتمالی بدست آمده، در آزمایش ۱۳ به مقادیر بهینه احتمالی دست پیدا میکنیم که (b=0.4) و (k=2.8) که پس از یافتن این مقادیر با آزمایش‌های با گام کوچک از صحیح بودن نتایج بدست آمده اطمینان حاصل میکنیم.

نکته قابل ذکر در این سری از آزمایش‌ها دو مورد آزمایش ۱۰ و ۱۷ هستند که از نظر معیار ارزیابی Recall و تعداد سند مرتبط بازگردانده شده آمار بهتری دارند ولی به علت پایین بودن باقی معیارهای ارزیابی (برای مثال P@5 که نشان‌دهنده دقت بازگردانی ۵ سند اول است) ما همان نتایج آزمایش ۱۳ را بعنوان بهینه در نظر میگیریم.

ب) بازیابی برای پرس و جوهای ۱۰۰ –۵۱را یک بار با مقادیر پیشفرض گالاگو برای پارامترهای kو bو بار دیگر با پارامترهای بهینه به دست آمده در قسمت الف انجام دهید. آیا MAPبرای این پرس و جوها با مقادیر بهینه به دست آمده در قسمت الف افزایش پیدا میکند؟ نتایج را تحلیل کنید

{فایل‌های Q01P01B}

پاسخ

برای این منظور ۲ بار بازیابی انجام میشود که نتایج آن در جدول زیر قابل مشاهده است

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5	Recall
default	4801	6100	1242	0.172	0.305	0.392	0.203
optimal	4801	6100	1248	0.172	0.304	0.396	0.204

همانطور که مشاهده میشود بر خلاف پرس و جوهای ۱۰۱-۱۵۰ ، مقادیر بهینه گذشته تفاوت بسیار اندکی با مقادیر پیشفرض یعنی (b=0.75) و (k=1.2) دارند و از لحاظ معیار MAP هیچ تفاوتی حاصل نمیشود. فلذا نتیجه میشود مقدارهای بهینه b , k برای پرس و جوهای متفاوت مقادیر مختلفی دارند. با انجام آزمایش مانند قسمت الف مقادیر بهینه برای ۵۱-۱۰۰ را بدست می آوریم.

جدول مقادیر تست:

شماره بازیابی	مقدار b	مقدار K
bm25Test00	0.75	1.2
bm25Test01	0.4	2.8
bm25Test02	0.75	2
bm25Test03	0.4	1.2
bm25Test04	0.4	1.5
bm25Test05	0.4	1
bm25Test06	0.4	1.1

bm25Test07	0.4	1.3
bm25Test08	0.5	1.2
bm25Test09	0.3	1.2
bm25Test10	0.2	1.2
bm25Test11	0.1	1.2

جدول نتیجه ارزیابی:

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5
00	4801	6100	1242	0.172	0.305	0.392
01	4801	6100	1248	0.172	0.304	0.396
02	4801	6100	1223	0.169	0.198	0.384
03	4801	6100	1284	0.178	0.316	0.412
04	4801	6100	1273	0. 178	0.315	0.412
05	4801	6100	1281	0. 178	0.313	0.408
06	4801	6100	1283	0. 178	0.313	0.412
07	4801	6100	1283	0. 178	0.316	0.416
08	4801	6100	1280	0. 178	0.313	0.388
09	4801	6100	1289	0. 178	0.317	0.420
10	4801	6100	1285	0. 178	0.317	0.429
11	4801	6100	1282	0.178	0.317	0.449

از نتایج برمی‌آید که اولا تاثیر مقادیر b, k در پرس و جوهای ۵۱-۱۰۰ از لحاظ معیار ارزیابی MAP به نسبت بسیار کمتر از پرس و جوهای ۱۰۱-۱۵۰ میباشد و این بدان معناست که میانگین دقت بازیابی اطلاعات این پرس و جوها نسبت به مقادیر b, k تغییرات کمتری دارد از طرفی نمونه آزمایش های ۱۰ و ۱۱ به علت P@5 بالاتر نسبت به باقی، از دید کاربر بهتر و کاراتر به نظر خواهند رسید.

{فایل‌های Q01P02-06}

2 -روش پیشنهادی اول

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)$$

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5	Recall
default	4801	6100	169	0.015	0.039	0.037	0.027

$$f(q,d) = \sum_{w \in q \cap d} \frac{c(w,d)(k+1)}{c(w,d)+k}$$

3 -روش پیشنهادی دوم

جدول مقادیر تست:

شماره بازیابی	مقدار b	مقدار K
bm25Test2	-	-
bm25Test21	-	2
bm25Test22	-	0.5
bm25Test23	-	1
bm25Test24	-	1.3
bm25Test25	-	1.5

جدول نتیجه ارزیابی:

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5
2	4801	6100	1189	0.159	0.290	0.424

0.412	0.289	0.156	1178	6100	4801	**2_1**
0.424	0.288	0.156	1183	6100	4801	**2_2**
0.433	0.290	0.160	1186	6100	4801	**2_3**
0.424	0.292	0. 159	1189	6100	4801	**2_4**
0.424	0.290	0. 158	1190	6100	4801	**2_5**

4 - روش پیشنهادی سوم

$$f(q,d) = \sum_{w \in q \cap d} I(w,d)$$

$$I(w,d)=1 \text{ (if count(w)}\neq 0),\ 0(\text{o.w.})$$

Recall	P@5	NDCG	MAP	Num_Rel_Ret	Num_Rel	Num_Ret	شناسه اجرا
0.122	0.249	0.184	0.078	746	6100	4801	**default**

5 - روش پیشنهادی چهارم

$$f\left(q,d\right) = \sum_{w \in q \cap d} IDF(w) \frac{(k+1)\left(\frac{c\left(w,d\right)}{\left(1-b+b\frac{|d|}{avdl}\right)}+\delta\right)}{k+\left(\frac{c\left(w,d\right)}{\left(1-b+b\frac{|d|}{avdl}\right)}+\delta\right)}$$

جدول مقادیر تست:

شماره بازیابی	مقدار b	مقدار K
bm25Test4	-	-
bm25Test41	0.75	2

1.5	0.75	bm25Test42
1.2	0.5	bm25Test43
1.2	0.4	bm25Test44
1	0.4	bm25Test45
0.8	0.2	bm25Test46
0.7	0.3	bm25Test47
0.6	0.2	bm25Test48
0.1	0.5	bm25Test49

جدول نتیجه ارزیابی:

P@5	NDCG	MAP	Num_Rel_Ret	Num_Rel	Num_Ret	شناسه اجرا
0.392	0.300	0.170	1228	6100	4801	4
0.376	0.297	0.168	1219	6100	4801	4_1
0.384	0.298	0.169	1223	6100	4801	4_2
0.392	0.310	0.175	1264	6100	4801	4_3
0.400	0.313	0. 177	1269	6100	4801	4_4
0.412	0.315	0. 178	1273	6100	4801	4_5
0.420	0.318	0. 179	1287	6100	4801	4_6
0.420	0.317	0. 178	1289	6100	4801	4_7
0.433	0.318	0. 178	1290	6100	4801	4_8
0.453	0.318	0. 179	1282	6100	4801	4_9

6 -روش پیشنهادی پنجم

$$f\left(q,d\right)=\sum_{w\in q\cap d} IDF(w)\left(\frac{c\left(w,d\right)(k+1)}{c\left(w,d\right)+k\left(1-b+b\frac{|d|}{avdl}\right)}+\delta\right)$$

جدول مقادیر تست:

شماره بازیابی	مقدار b	مقدار K
bm25Test5	-	-
bm25Test51	0.75	2
bm25Test52	0.75	0.5
bm25Test53	0.5	0.5
bm25Test54	0.3	0.5

جدول نتیجه ارزیابی:

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5
5	4801	6100	1242	0.172	0.305	0.392
5_1	4801	6100	1223	0.166	0.298	0.384
5_2	4801	6100	1244	0.171	0.305	0.400
5_3	4801	6100	1268	0.174	0.310	0.412
5_4	4801	6100	1271	0. 174	0.310	0.429

با توجه به آزمایشهای بالا، در مقوله مقایسه روش‌های پیشنهادی نتایج زیر حاصل میشود:

- ✓ پایین ترین دقت مربوط به روش اول است زیرا تنها پارامتر دخیل در این روش نمره‌دهی مقدار IDF توکن‌ها میباشد که همانطور که میدانیم، این پارامتر به معنی اعتبار ترم در میان سندهای کالکشن است یعنی میزان و اهمیت اطلاعاتی که ترم یا توکن به ما میدهد را می‌سنجد فلذا برای بازیابی اطلاعات همانطور که در مطالب درس نیز داشتیم، از دقت خوبی برخوردار نیست.
- ✓ در روش پیشنهادی دوم که از حالت‌های پایه‌ای TF Transformation(Bm25 Transformation)، در قالب $\frac{(k+1).x}{k+x}$ است، به وضوح میتوان افزایش چشمگیر دقت بازیابی را در اطلاعات بدست آمده مشاهده کرد. (در روش‌هایی که از پارامترهای b, k استفاده کرده‌اند علاوه بر حالت پیشفرض سعی شده تا با تکرار آزمایش مقادیر بهینه بدست بیاید تا نتیجه بهینه بازیابی اطلاعات توسط روش مورد نظر نیز در دسترس باشد که مقادیر در جداول آورده شده‌اند)

- ✓ در روش پیشنهادی سوم تنها حضور یا عدم حضور عبارت در سند مورد بررسی قرار داده میشود که همانطور که از نتایج پیداست با اینکه از روش اول نتایج بهتری دارد ولی در پاسخ پرس و جوها سندهای نامرتبط بیشتری را برمیگرداند که میتوان نتیجه گرفت که صرف حضور یک عبارت در سند به معنی مرتبط بودم سند با پرس و جو نیست ولی در مقایسه با در نظر گرفتن اعتبار ترم یا توکن نتایج مناسب‌تری بازمیگرداند.
- ✓ در روش‌های چهارم و پنجم تاثیر Pivot Length Normalization را بر نتایج بازیابی مشاهده میکنیم. طبق تحقیقات، فرمول BM25 سندهای بسیار طولانی را بعضا بیش از حد جریمه میکند و این باعث میشود برخی از اسناد بازیابی نشوند.
- ✓ هردو روش‌های ۴و۵ تلاش بر رفع مشکل امتیازدهی ناعادلانه به اسناد مرتبط طولانی را دارند که در روش پنجم تنها یک پارامتر دلتا با معادله افزوده شده که در مثال ما مشاهده شد که این پارامتر (با هر مقدار دلخواهی با توجه به معادلات ذکر شده در داکیومنت مرجع که در ذیل آورده شده) هیچ گونه تاثیری در نتایج بازیابی ندارد. پس در رفع این مشکل راه حل چهارم بهتر عمل میکند و اسناد مرتبط بیشتری بازیابی میشوند.

6.1 Lower-Bounded BM25 (BM25+)

It is trivial to verify that BM25+ still satisfies LB1 unconditionally. To examine LB2, we apply an analysis method that is consistent with our analysis for BM25 in Section 5.1. The LB2 constraint on BM25+ is equivalent to

$$\frac{k_1}{k_1+2} < \frac{(k_1+1)\cdot 1}{k_1\left(1-b+b\frac{|D_2|}{avdl}\right)+1}+\delta \qquad (17)$$

which can be shown to be satisfied unconditionally if

$$\delta \geq \frac{k_1}{k_1+2} \qquad (18)$$

Clearly, if we set δ to a sufficiently large value, BM25+ is able to satisfy LB2 unconditionally, which is also confirmed in our experiments that BM25+ works very well when we set $\delta = 1$.

- ✓ در میان روش‌های ذکر شده بهترین روش از دید کاربر (P@5) و همچنین از نظر تعداد سند مرتبط بازیابی شده و دگیر معیارها، روش چهارم (با مقادیر b, k بهینه بدست آمده) میباشد

سوال دوم: تابع بازیابی Pivoted Length Normalization

هدف از این سوال، آشنایی با تاثیر تابع تبدیل استفاده شده برای مولفه ی TFدر کیفیت رتبه بندی میباشد. این روش برای اولین بار در مقاله ای با عنوان، Pivoted Document Length Normalization معرفی گردید

{فایل‌های Q02P01-02}

پاسخ

برای این منظور ۲ بار بازیابی انجام میشود که نتایج آن در جدول زیر قابل مشاهده است

شناسه اجرا	Num_Ret	Num_Rel	Num_Rel_Ret	MAP	NDCG	P@5
Default(0)	4801	6100	745	0.076	0.173	0.233
Not_nested(1)	4801	6100	474	0.048	0.113	0.176

در مسائل سوال اول برای محاسبه تاثیر TF از بهترین روش ممکن موجود یعنی روش‌های BM25 استفاده میشد ولی در این سوال هدف مقایسه روش‌های کلاسیک تر و قدیمی‌تر محاسبه و تاثیر TF میباشد. میدانیم در ابتدا برای محاسبه TF از یک تابع خطی (y = x) استفاده میشد ولی بعدتر برای کنترل بهتر ترم‌هایی که به تعداد زیاد تکرار میشوند از توابع لگاریتمی استفاده شد که ۲ روشی که در مطالب درس بررسی شد در اینجا آورده شده و به صورت عملی مورد بررسی قرار میگیرد.

در روش BM25 میتوان با انتخاب مناسب مقدار پارامتر k علاوه بر کنترل وزن‌دهی ترم‌ها، حد بالای وزن نیز برای ترم‌ها مشخص کرد که این قابلیت در توابع لگارتیمی وجود ندارد و این توابع حد بالا ندارند.

با توجه به موارد بالا و نتایج بدست آمده از آزمایش‌ها، به وضوح میتوان کارایی این ۳ روش را مقایسه کرد. به وضوح میتوان کارایی بهترِ روش BM25 را نسبت به روش‌های لگاریتمی مشاهده کرد. علاوه بر این از بین روش‌های لگاریتمی، روش لگاریتم تودرتو به مراتب نتیجه بهتری نسبت به روش بدون لگاریتم تودرتو را بازیابی میکند. علت این امر میتواند ناشی از نزدیک‌تر بودن این تابع به تابع خطی باشد که با وجود اینکه نتایج بهتری نسبت به روش پیشنهادی اول سوال قبل(که تنها IDF را در نظر میگیرد) ارائه میدهد اما همچنان از دقت پایینی برخوردار است. این درحالیست که روش تودرتو نتایج نزدیکی با حالت سوم سوال قبل دارد که شاید بتوان نتیجه گرفت که با لگاریتم تودرتوی استفاده شده معیار ما نزدیک به حالت سوم (تنها مشاهده ترم در سند کافی باشد) شده است.