به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده‌ی مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین اول

آبان ۱۴۰۰

مقدمه

امروزه با توجه به افزایش حجم داده‌های متنی، موتورهای جستجو اصلی‌ترین 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، رتبه‌بندی 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، رتبه‌بندی و نمایش داده شوند. در نهایت رتبه‌بندی حاصل عموما با رتبه‌بندی طلایی[1] مقایسه شده و کارایی تابع بازیابی گزارش می‌گردد.

ابزار جستجوی متنی مورد استفاده در این تمرین گالاگو[2] می‌باشد.

اهداف تمرین:

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

نکات قابل توجه درهنگام پاسخ به سوالات:

- در **تمامی تمرین‌ها**، نمره اصلی به تفسیر دانشجویان تعلق می‌گیرد (تفسیر اجباری است).
- استفاده از نمودارها و کشف نمونه‌های مرتبط از اسناد و پرس‌وجوها در صورتی که موجب افزایش کیفیت تفسیرها گردد، تاثیر مثبت در نمره شما خواهد داشت.
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، به تفسیرهایی که بدون آزمایش و صرفا به صورت فرضی بیان گردند نمره‌ای تعلق نمی‌گیرد.

[1] Golden
[2] Galago

شرح دادگان

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

پیکره متنی[3] (فایل اسناد):

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

OCNO: شناسه هر سند

Head: عنوان سند

Text: متن سند

(دقت کنید که این فایل xml صحیح نمی‌باشد و با کتابخانه xml قابل تجزیه نیست.)

فایل پرس و جوها [4] :

این فایل شامل پرس‌وجوها می‌باشد.

فایل قضاوتهای مرتبط [5] :

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

---

[3] corpus
[4] Queries
[5] Relevance Judgment

پیش نیاز- ایجاد شاخص[6]

همانطور که در مطالب درسی بیان گردید، جهت استفاده از اسناد در توابع بازیابی، بایستی اسناد ابتدا شاخص‌گذاری گردند تا دسترسی به آماره‌های مورد نیاز برای محاسبه‌ی مقادیر امتیازها ساده شود. جهت شاخص‌گذاری می‌توانید از دستورات موجود در ابزار گالاگو استفاده کنید.

هنگام شاخص‌گذری به نکات زیر توجه کنید:

1- از Porter Stemmer جهت ریشه‌یابی کلمات استفاده کنید.

2- از Tokenizerجهت جداسازی کلمات موجود در فیلدهای text و head استفاده کنید.

3- نوع فایل را trectext قرار دهید.

در گزارش خود نقش هر کدام از این پارامترهای ذکرشده را بیان کنید.

[6] index

سوال 1- تابع بازیابی BM25

هدف از این سوال آشنایی با مولفه‌های روش BM25 و تاثیر هر یک بر روی کیفیت رتبه‌بندی می‌باشد.

راهنمایی:

- با توجه به اینکه روش BM25 در ابزار گالاگو پیاده‌سازی گردیده، به راحتی می‌توانید از مولفه‌های آن در پیاده‌سازی روش‌های پیشنهادی استفاده کنید.
- در مقداردهی برای پارامترها بهتر است ابتدا گام‌های بلند و سپس گام‌های کوچک آزمایش گردند تا منابع محاسباتی تلف نشود.
- در این قسمت روش‌های پیشنهادی (مولفه‌های تابع بازیابی BM25) را بایستی در گالاگو پیاده‌سازی کنید. برای این کارمی‌توانید فایل BM25ScoringIterator.java در پوشه گالاگو جستجو کنید، نمونه‌هایی از آن ایجاد کنید، سپس توابع score آن را تغییر دهید و فایل خود را با نامی ذخیره نمایید. (هرنام‌گذاری می‌توانید انجام دهید.) در ادامه برای آن که بتوانید کلاس (فایل ایجاد شده) خود را از command line صدا بزنید، کلاس خود را با نام مربوطه در قسمت Score iterators در فایل FeatureFactory.java اضافه کنید. در پایان، build را مجددا انجام دهید تا تغییرات انجام شده، صورت پذیرد.
- **معیارهای ارزیابی Recall، MAP، nDCG و P@5 می‌باشند.**

سوالات:

1- روش بازیابی BM25 :

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

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

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

در ادامه بازیابی برای پرس و جوهای ۵۱ –۱۰۰ را با سایر توابع پیشنهادی (۲،۳،۴،۵،۶) انجام دهید، و مرتبط بودن/نبودن اسناد با پرس وجوهای به دست آمده از توابع بازیابی را با فایل دادگان طلایی مقایسه کرده تا معیارهای ارزیابی را بدست آورید. در نهایت با توجه به نتایج معیارهای ارزیابی، تمامی توابع را با یکدیگر مقایسه کنید و تفسیر خود را بیان نمایید.

2- روش پیشنهادی اول

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

3- روش پیشنهادی دوم

$$f(q_و d) = \sum_{W \in q \cap d} \frac{c(w_و d)(k+1)}{c(w_و d) + k}$$

4- روش پیشنهادی سوم

$$f(q_و d) = \sum_{W \in q \cap d} I(w_و d)$$

I(w,d)=1 (if count(w)≠0), 0(o.w.)

5- روش پیشنهادی چهارم (BM25L[7])

(مقدار ۰.۵ برای δ در نظر گرفته شود)

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

6- روش پیشنهادی پنجم (BM25+[8])

(مقادیر مختلف برای δ بررسی شود و بهترین مقدار گزارش شود)

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

[7] Lv, Y., C. Zhai, When documents are very long, BM25 fails! SIGIR 2011, p. 1103-1104

[8] Lv, Y., C. Zhai, Lower-bounding term frequency normalization, CIKM 2011, p. 7-16

سوال 2- تابع بازیابی Pivoted Length Normalization

هدف از این سوال، آشنایی با تاثیر تابع تبدیل استفاده شده برای مولفه‌ی TF در کیفیت رتبه‌بندی می‌باشد. این روش برای اولین بار در مقاله‌ای با عنوان، [9] Pivoted Document Length Normalization معرفی گردید.

- در این قسمت روش‌های موجود را بایستی در گالاگو پیاده‌سازی کنید. برای این کارمی‌توانید فایل BM25ScoringIterator.java در پوشه گالاگو جستجو کنید، نمونه‌هایی از آن ایجاد کنید، سپس توابع score آن را تغییر دهید و فایل خود را با نامی ذخیره نمایید. (هرنام‌گذاری می‌توانید انجام دهید.) در ادامه برای آن که بتوانید کلاس (فایل ایجاد شده) خود را از command line صدا بزنید، کلاس خود را با نام مربوطه در قسمت Score iterators در فایل FeatureFactory.java اضافه کنید. در پایان، build را مجددا انجام دهید تا تغییرات انجام شده، صورت پذیرد.
- **معیارهای ارزیابی Recall، MAP، nDCG و P@5 می‌باشند.**

در این قسمت بازیابی پرس و جوهای ۵۱-۱۰۰ را با استفاده از مقادیر پیشفرض ابزار گالاگو، با روش‌های زیر را انجام دهید و سپس رتبه‌بندی به دست آمده را با فایل دادگان طلایی مقایسه کرده تا معیارهای ارزیابی را بدست آورید. در نهایت با توجه به نتایج معیارهای ارزیابی، توابع را با یکدیگر مقایسه کنید و تفسیر خود را بیان نمایید.

1- مدل اصلی :

$$f\left(q_و d\right) = \sum_{W \in q \cap d} c(w_و q) \frac{\ln\left[1 + \ln\left[1 + c(w_و d)\right]\right]}{1 - b + b\frac{|d|}{avdl}} \log\frac{M+1}{df(w)}$$

2- مدل بدون مولفه لگاریتمی تودرتو

$$f\left(q_و d\right) = \sum_{W \in q \cap d} c(w_و q) \frac{\ln[1 + c(w_و d)]}{1 - b + b\frac{|d|}{avdl}} \log\frac{M+1}{df(w)}$$

3- نتایج مدل اصلی نسبت به روش‌های BM25 و BM25+ با توجه به معیارهای ارزیابی مقایسه شود. علت تغییر در نتایج را در صورت مشاهده بیان کنید.

[9] http://singhal.info/pivoted-dln.pdf

ملاحظات (حتما مطالعه شود)

- تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA1_StudendID تحویل داده شود. این پوشه باید شامل موارد زیر باشد:

1- گزارش به صورت تایپ شده در قالب PDF، شامل شرح آزمایش های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد. (از توضیح دادن کد در گزارش جدا خودداری نمایید. در صورت نیاز به توضیح بر روی کد کامنت بگذارید.) در صورت نیاز، تنها نام فایل، که شامل کدهای مربوطه می‌باشد آورده شود.

2- یک پوشه به نام code، که در آن فایل کدهای خواسته شده با نام گذاری و ساختار مناسب قرار گیرند.

- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود، که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری وبه اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد گزارش می‌شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند ترتیب اثری داده نخواهد شد. گزارش نهایی خود را حتما به صورت PDF در سایت درس بارگذاری نمایید.
- مهلت تحویل بدون جریمه: 05 آذر 1400
- مهلت تحویل با تاخیر و جریمه 30 درصد: 12 آذر 1400
- در صورت بروز هر گونه ابهام و سوال با دستیاران آموزشی طراح سوال تماس بگیرید.

beheshti.7676@gmail.com

Hoomanshirvani@ut.ac.ir