به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین اول

مهرماه ۱۴۰۲

*فهرست

بخش ۱- سؤالات عملی

با توجه به افزایش روزافزون حجم اطلاعات متنی، موتورهای جستجو از مهم‌ترین ابزارهایی هستند که جهت بازیابی اطلاعات مورد استفاده قرار می‌گیرند. موتورهای جستجو با در نظر گرفتن پرس‌وجو ورودی کاربر، اسناد موجود را به‌کمک روش‌ها و توابع بازیابی، از لحاظ شباهت به پرس‌وجو امتیازدهی و سپس رتبه‌بندی کرده و به کاربر نمایش می‌دهند. با توجه به اهداف درس بازیابی هوشمند اطلاعات، هدف از این تمرین آشنایی با ابزارهای جستجوی متنی و همچنین آشنایی با معیارهای ارزیابی و توابع امتیازدهی به اسناد است. یک تابع امتیازدهی با توجه به میزان ارتباط یک سند با پرس‌وجو، امتیازی به سند تخصیص می‌دهد تا در نهایت اسناد براساس امتیازشان، رتبه‌بندی و نمایش داده‌شوند. رتبه‌بندی حاصل عموماً با رتبه‌بندی طلایی[۱] مقایسه شده و کارایی تابع بازیابی گزارش می‌گردد.

برای جستجوی متنی در این تمرین از ابزار گالاگو[۲] استفاده می‌شود.

اهداف تمرین:

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

نکات قابل توجه در هنگام پاسخ به سؤالات:

- 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، تأثیر مثبت در نمره شما خواهد داشت.
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، به تفسیرهایی که بدون آزمایش و صرفاً به‌صورت فرضی بیان‌گردند نمره‌ای تعلق نمی‌گیرد.

---

[۱] Golden Rankings

[۲] Galago

شرح دادگان

برای این تمرین داده‌های زیر بر روی سایت درس قرار داده‌شده‌اند.

پیکره متنی[۱] (فایل اسناد):

مجموعه‌ای از ۱۴۰۰ سند به‌دست آمده از چکیده‌های علمی که هر سند شامل فیلدهای زیر می‌باشد:

۱. DOCNO: شناسه هر سند
۲. FILEID: شناسه فایل
۳. HEAD: عنوان سند
۴. TEXT: متن سند

فایل پرس‌وجوها[۲]:

این فایل شامل ۱۶۰ پرس‌وجو می‌باشد و فیلدهای زیر را شامل می‌شود:

۱. Number: شناسه پرس‌وجو
۲. Text: متن پرس‌وجو

فایل دادگان طلایی[۳]:

این فایل شامل قضاوت‌های مرتبط[۴] می‌باشد در مرحله نهایی جهت ارزیابی کارایی توابع بازیابی مورد استفاده قرار می‌گیرد.

---

۱ Corpus
۲ Queries
۳ Golden Dataset
۴ Relevance Judgments

پیش‌نیازها – ایجاد شاخص[1]

همان طور 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:

- نوع فایل را trectext قرار دهید.
- از Porter Stemmer جهت ریشه‌یابی کلمات استفاده کنید.
- از Tokenizer جهت جداسازی کلمات موجود در فیلد text استفاده کنید.

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

[1] Indexing

سؤال ۱- تابع بازیابی BM25

هدف از این سؤال آشنایی با مولفه‌های روش BM25 و تأثیر هر یک بر روی کیفیت رتبه‌بندی می‌باشد. فرض کنید جستجوهای موجود در این سامانه انجام شده و برای هر جستجو می‌خواهیم ۳۰ صفحه حاوی ۱۰ مقاله به کاربر نمایش بدهیم.

راهنمایی:

- با توجه به اینکه روش BM25 در ابزار گالاگو پیاده‌سازی شده، به‌راحتی 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 قسمت روش‌های پیشنهادی (مولفه‌های تابع بازیابی BM25) را بایستی در گالاگو پیاده‌سازی کنید. برای این کار می‌توانید 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 نام دلخواهی ذخیره نمایید. (هر نام‌گذاری می‌توانید انجام دهید.) در ادامه برای آنکه بتوانید کلاس (فایل ایجاد شده) خود را از طریق command_line صدا بزنید، کلاس خود را با نام مربوطه در قسمت Score iteratorsدر فایل FeatureFactory.java اضافه کنید. در پایان، build را مجدداً انجام دهید تا تغییرات انجام‌شده، اعمال شوند
- **معیارهای ارزیابی Recall، MAP، nDCG، P@10 می‌باشند.**

سؤالات:

الف) در مرحله اول شما باید بازیابی را به روش BM25 برای پرس‌وجوها انجام دهید و تأثیر پارامترهای این روش (b, k) را بررسی کنید. شما باید مقادیر مختلف را برای پارامترها آزمایش کنید تا به مقدار بهینه برای این دو مقدار برسید. هنگام تفسیر مقادیر بهینه، به تأثیر هر یک از مولفه‌های تابع امتیازدهنده دقت کنید.

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

ب) در این قسمت بازیابی را با سایر روش‌های پیشنهادی انجام دهید و مرتبط بودن یا نبودن اسناد با پرس‌وجوهای به‌دست آمده از توابع بازیابی را با دادگان طلایی مقایسه کنید تا معیارهای ارزیابی را به‌دست آورید. در نهایت با توجه به معیارهای ارزیابی، تمامی توابع را با یکدیگر مقایسه کنید و نتایج به‌دست‌آمده را تفسیر کنید.

۱) روش پیشنهادی اول

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

۲) روش پیشنهادی دوم

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

۳) روش پیشنهادی سوم (PL2)

$$f(q,d) = \sum_{w \in q \cap d} c(w,d) * log(1 + \frac{avdl}{|d|}) * \log\ (\frac{N}{IDF(w)})$$

۴) روش پیشنهادی چهارم (وزن بیشتر عناصر کمیاب)

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)^2 \frac{c(w,d)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})}$$

۵) روش پیشنهادی پنجم (وزن بیشتر ترم‌های تکرارشونده)

$$f(q,d) = \sum_{w \in q \cap d} IDF(w) \frac{c(w,d)^2}{c(w,d)^2+k(1-b+b\frac{|d|}{avdl})}$$

۶) روش پیشنهادی ششم (وزن بیشتر ترم‌های تکرارشونده)

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

۷) روش پیشنهادی هفتم (BM25+)

مقادیر مختلف برای δ بررسی شود و بهترین مقدار گزارش شود

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

سؤال ۲ – تابع بازیابی Pivoted Length Normalization

هدف از این سؤال آشنایی با تأثیر تابع تبدیل استفاده‌شده برای مولفه TF در کیفیت رتبه‌بندی می‌باشد.

این روش برای اولین بار در مقاله‌ای[۱] با عنوان Pivoted Document Length Normalization معرفی گردید.

- در این قسمت باید روش‌های موجود را در گالاگو پیاده‌سازی کنید. برای این کار می‌توانید 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 نام دلخواه ذخیره نمایید. (هر نام گذاری می‌توانید انجام‌دهید.) در ادامه برای آنکه 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 مجدداً انجام دهید تا تغییرات انجام‌شده، اعمال شود.
- **معیارهای ارزیابی Recall، MAP، nDCG، P@10 می‌باشند.**

الف) در این قسمت بازیابی پرس‌وجوهای موجود را با استفاده از مقادیر پیش‌فرض ابزار گالاگو و روش‌های زیر انجام دهید و سپس رتبه‌بندی به‌دست آمده را با فایل دادگان طلایی مقایسه کرده تا معیارهای ارزیابی را به‌دست آورید. در نهایت با توجه به نتایج معیارهای ارزیابی، توابع را با یکدیگر مقایسه کنید و تفسیر خود را بیان نمایید.

۱) مدل اصلی:

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

۲) مدل بدون مولفه تودرتو

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

ب) نتایج مدل اصلی نسبت به‌روش‌های BM25 و BM25+ با توجه به معیارهای ارزیابی مقایسه کنید. علت تغییر در نتایج را در صورت مشاهده بیان کنید.

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

بخش ۲- سؤالات تئوری

سؤال ۱ – Rocchio relevance feedback

فرض کنید پرس‌وجوی کاربر "بازیابی هوشمند اطلاعات" می‌باشد. کاربر تعداد ۴ سند را مورد بررسی قرار می‌دهد. پس از بررسی نتایج، کاربر اسناد را به دو دسته مرتبط و نامرتبط دسته بندی می‌کند. گزارش کاربر به‌صورت زیر است:

شماره سند	محتوای سند	گزارش کاربر
۱	بازیابی اطلاعات بازیابی منابع	مرتبط
۲	کلاس بازیابی هوشمند اطلاعات	مرتبط
۳	سیستم مدیریت هوشمند کلاس	نامرتبط
۴	مدیریت هوشمند منابع سیستم	نامرتبط
Words = (بازیابی، اطلاعات، هوشمند، منابع، کلاس، سیستم، مدیریت)		

۱- پرس‌وجوی جدید را به‌دست آورده و تغییرات آن را تحلیل کنید (اگر بعد از آپدیت وزن‌های به‌دست آمده منفی بود صفر در نظر بگیرید)

۲- تأثیر پارامترها را در نظر گرفته و تغییرات آن‌ها را بررسی کنید.

۳- در نهایت مقادیر پیشنهادی خود را برای پارامترها با ذکر دلیل اعلام کنید.

فرضیات:

۱. بدون اعمال مقیاس‌بندی (TF-Scaling) مستقیماً فقط از فرکانس مربوط به کلمات استفاده می‌گردد.

۲. مقادیر α و β و γ را در ابتدا برابر ۱ در نظر بگیرید.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j\in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j\in D_{nr}}\vec{d}_j$$

سؤال ۲- معیارهای ارزیابی

۳ سیستم بازیابی اطلاعات پیش روی شما قرار دارد و رتبه‌بندی بازگردانده شده آنها برای یک پرس‌وجو نشان داده شده است. پایگاه داده اطلاعاتی سیستم شامل ۲۰ سند می‌باشد که:

- اسناد فرد مرتبط به پرس‌وجو هستند
- اسناد زوج مرتبط به پرس‌وجو نیستند.

Rank	R1	R2	R3
1	d1	d1	d1
2	d2	d2	d2
3	d5	d4	d4
4	d6	d5	d5
5	d13	d6	d9
6		d7	d10
7		d8	d12
8		d9	d13
9		d10	d14
10		d11	d15
11		d12	d20
12		d13	
13		d19	
14		d14	
15		d17	
16		d3	
17		d15	
18		d16	
19		d18	
20		d20	

برای این سامانه بر اساس پاسخ‌های داده شده معیارهای زیر را محاسبه کنید.

الف) Precision

ب) recall

ج) p@4 و p@7 و p@12

د) R-precision

ه) MAP

ز) gMAP

ح) رسم نمودار 11-point precision-recall

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA1_StudentID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ‌شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را به تفکیک هر سؤال شامل شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به‌صورت کاغذی تحویل داده شوند یا به‌صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- در پاسخ به سؤالات عملی، بایستی آزمایش‌های انجام‌شده، پارامترهای آزمایش، نتایج و تحلیل‌ها را به‌طور کامل شرح دهید.
- 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 تأخیر تحویل تمرین تا یک هفته به ازای هر روز ۱۵ درصد است.
- توجه کنید این تمرین باید به‌صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت‌کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل‌های زیر در ارتباط باشید:

mailto: mohammad.na3ri@gmail.com

mailto: mj.kamyab@ut.ac.ir

مهلت تحویل بدون جریمه: **۲۹ مهرماه ۱۴۰۲**

مهلت تحویل با تأخیر، با جریمه ۱۵ درصد: **۶ آبان‌ماه ۱۴۰۲**