به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین دوم

آبان‌ماه ۱۴۰۲

*فهرست

بخش ۱- سوالات عملی

مقدمه

در تمرین اول با معیارهای ارزیابی و توابع امتیازدهی به اسناد آشنا شدید. دیدید که 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 داده شوند. در این تمرین قصد داریم روش‌های مختلف هموارسازی توابع بازیابی و پارامترهای آنها را مورد مطالعه قرار بدهیم.

نکات قابل توجه درهنگام پاسخ به سؤالات:

- معیارهای ارزیابی در این تمرین MAP و P@10 می‌باشند..
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، به تفسیرهایی که بدون آزمایش و صرفا به صورت فرضی بیان‌گردند نمره‌ای تعلق نمی‌گیرد.

شرح دادگان

برای این تمرین داده‌های زیر بر روی سایت درس قرار داده شده‌اند.

پیکره متنی[1] (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[2]:

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

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

فایل دادگان طلایی[3]:

این فایل شامل قضاوت‌های مرتبط[4] می‌باشد در مرحله نهایی جهت ارزیابی کارایی توابع بازیابی مورد استفاده قرار می‌گیرد.

پیش‌نیازها

مشابه تمرین قبلی جهت استفاده از اسناد در توابع بازیابی، بایستی اسناد ابتدا شاخص‌گذاری گردند تا دسترسی به آماره‌های مورد نیاز برای محاسبه‌ی مقادیر امتیازها ساده شود.

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.

[1] Corpus
[2] Queries
[3] Golden Dataset
[4] Relevance Judgments

سؤال ۱- هموارسازی

یکی از مشکلات حوزه بازیابی اطلاعات و روش‌های Likelihood، وجود احتمال‌های صفر برای کلماتی است که در اسناد مشاهده نمی‌شوند. روش‌های هموارسازی برای حل این مشکل مطرح‌شده‌اند تا احتمال رخداد کلمات دیده نشده پرس‌وجو در اسناد را تخمین بزنند.

ابزار گالاگو، به صورت پیش فرض بازیابی را به‌روش Query-Likelihood انجام می‌دهد. هدف از این سوال آشنایی با روش‌های هموارسازی می‌باشد. با استفاده از نمودار مناسب برای هر یک از روش‌های خواسته شده مقادیر مختلف λ، μ و δ را بررسی کنید و مقدار بهینه را گزارش نمایید. (تعداد اسناد بازیابی شده را ۱۰۰ قرار دهید و برای ریشه‌یابی از porter stemmer استفاده‌کنید.)

روش‌هایی که قصد داریم در این تمرین مورد بررسی قرار دهیم عبارتند از:

- روش Additive Smoothing با پارامتر δ
- روش JM با پارامتر λ
- روش Dirichlet Prior با پارامتر μ
- روش هموارسازی دو مرحله‌ای با پارامترهای λ و μ[1]

$$p(w|d) = (1-\lambda)\frac{c(w,d)+\mu p(w|C)}{|d|+\mu} + \lambda p\,(w|C)$$

راهنمایی:

- با توجه به اینکه روش JM و Dirichlet prior در ابزار گالاگو پیاده‌سازی شده‌است، می‌توانید از توابع پیش‌فرض گالاگو استفاده‌کنید.
- می‌توانید با تغییر تابع DirichletScoringIterator روش‌ها را پیاده‌سازی کنید.
- در مقداردهی برای پارامترها بهتر است ابتدا گام‌های بلند و سپس گام‌های کوچک آزمایش گردند تا منابع محاسباتی تلف نشود.

[1] هموارسازی دو مرحله‌ای از ترکیب دو روش هموارسازی Dirichlet و JM بدست می‌آید.

سؤال ۲– پیاده‌سازی تابع وزن‌دهی با استفاده از Pseudo Relevance Feedback

در این سؤال قصد داریم به پیاده‌سازی تابع وزن‌دهی با استفاده از Pseudo Relevance Feedback بپردازیم. برای این منظور فایل‌های زیر در صفحه درس قرار گرفته است:

- wResult.java
- wordWeight.java
- bSearch.java
- fbData.java
- fbMixtureModel.java

فایل‌های فوق شامل پیاده‌سازی روش Mixture Model است، برای پیاده‌سازی این سؤال ابتدا فرمت پرس‌وجوهای موجود در فایل q3.json را به فرمت زیر با پسوند tsv ذخیره کنید:

- #queryNumber [/t(tab)] queryTittle [\n(enter)]

برای مثال:

- 1 what similarity law….

برای انجام این سؤال در ابتدا نیاز است که به API گالاگو متصل شوید. برای این منظور در ابتدا نیاز است که در محیط IDE یک پروژه MAVEN ایجاد گردد. پس از ساخت این پروژه می‌توانید با ویرایش فایل POM.xml و با اضافه کردن API پروژه را به گالاگو متصل کنید. سپس فایل‌های فوق را به پروژه ایجاد شده اضافه نمایید. همچنین نیاز است تا فایل‌های کتابخانه گالاگو موجود در پوشه appassembler را نیز به پروژه خود اضافه کنید.

پس از اتصال به API گالاگو حال می‌توانید با استفاده از تغییراتی بر روی تابع محاسبه وزن مطابق خواسته مسئله، مدل EM را پیاده سازی کنید. در نهایت مدل خود را برای پرس‌وجوهای موجود در فایل q3.json اجرا کنید.

الف) با استفاده از نمودار مناسب رابطه بین تعداد سندهای منتخب برای بازخورد به ازای مقادیر بزرگتر از ۱ و معیارهای ارزیابی را نمایش دهید و سپس به تحلیل نتایج بپردازید. در نهایت مقدار بهینه را شناسایی و گزارش کنید.

ب) با توجه به تعداد اسناد منتخب، رابطه بین تعداد کلمات استخراج شده برای بازخورد و معیار ارزیابی را با نمودار مناسب نمایش دهید، مقدار بهینه را گزارش کرده و در نهایت به تحلیل نتایج به‌دست آمده بپردازید.

راهنمایی:

۱. در کلاس fbMixtureModel.java تابعی با نام computeWeights() ایجاد شده‌است، می‌بایست این تابع را به‌نحوی ویرایش کنید که با استفاده از روشEM، وزن‌دهی به کلمات استخراج شده از سندهای منتخب فراهم شود.

۲. با ویرایش کلاس bSearch.java می‌توانید فایل‌های ورودی و خروجی را برای پروژه خود تنظیم کنید.

بخش ۲- سؤالات تئوری

سؤال ۱ (امتیازی) – الگوریتم Expectation Maximization

یک مدل مخلوط دو بخشی برنولی[1] برای تعین اسناد مرتبط و نامرتبط مانند زیر دارید:

$$P(D|Q) = \mathrm{P}(\mathrm{R}=1|\mathrm{Q})\,\mathrm{P}(\mathrm{D}|\mathrm{R}=1,\mathrm{Q}) + \mathrm{P}(\mathrm{R}=0|\mathrm{Q})\,\mathrm{P}(\mathrm{D}|\mathrm{R}=0,\mathrm{Q})$$

در این عبارت R نشان دهنده‌ی کلاس ارتباط است. (R=0 نامرتبط، R=1 مرتبط)

الگوریتم EM در دو گام Expectaion و Maximization به تخمین احتمالات $P(R|Q)$ و $\mathrm{P}(\mathrm{w}|\mathrm{R},\mathrm{Q})$ می‌پردازد و به صورت زیر است:

گام Expectation: مقدار $\gamma(\mathrm{R}|\mathrm{D},\mathrm{Q})$ را محاسبه کنید.

گام Maximization: با استفاده از γ پارامترها را دوباره تخمین بزنید.

بعد از ۵ مرحله اجرا[2] احتمالات کلمه‌ی «retrieval» برابر مقدار زیر است:

$$\mathrm{P}(\mathrm{retrieval}|\mathrm{R}=1,\mathrm{Q}) = 0.8$$
$$\mathrm{P}(\mathrm{retrieval}|\mathrm{R}=0,\mathrm{Q}) = 0.3$$

۱. یک سند که شامل ۳ کلمه‌ی retrieval است چگونه دسته بندی می‌شود؟ پاسخ خود را شرح دهید. (مقادیر $\mathrm{P}(\mathrm{R}=1|\mathrm{Q}) = \mathrm{P}(\mathrm{R}=0|\mathrm{Q}) = 0.5$)

راهنمایی: باید مقدار زیر را حساب کنید

$$\gamma(\mathrm{R}=1|\mathrm{D},\mathrm{Q}) = \mathrm{P}(\mathrm{R}=1|\mathrm{D},\mathrm{Q})$$

سند به‌صورت:

D: “retrieval retrieval retrieval”

۲. نقطه‌ی ضعف و قدرت این الگوریتم در بازیابی اطلاعات چیست؟

[1] two-component Bernoulli mixture model
[2] iterations

سؤال ۲- مرتب سازی اسناد

اسناد زیر را در نظر بگیرید

D1: The cat sat on the mat. The cat was black and white.

D2: The dog played with the cat. The animals played in the park.

D3: The girl played in the park with her dog. They played for hours.

شما باید برای پرس‌وجوهای زیر اسناد را مرتب کنید:

Q1: cat

Q2: cat dog

Q3: cat dog park

۱. Okapi TF[1] را برای هر کدام از پرس‌وجوها در هر یک از اسناد محاسبه کنید. مقدار k1=1.5

۲. مقدار BM25 را بین هر کوئری و سند با استفاده از پارامترهای زیر محاسبه کنید:

k1 = 1.5, b = 0.75, avgdl = 10

۳. برای هر کدام از کوئری‌ها و روش‌ها اسناد را بر حسب امتیاز مرتب کنید.

۴. بررسی کنید چه عواملی باعث تفاوت در ترتیب اسناد در این دو روش شده‌است، همچنین بررسی کنید چگونه اهمیت کلمات در هر سند مشخص می‌شود.

[1] Okapi TF/BM25 TF

سؤال ۳- مرتب سازی اسناد با استفاده از هموارساز

اسناد و پرس‌وجو زیر را در نظر بگیرید:

D1: Machine learning models like regression are used for prediction tasks. Neural networks are a common model.

D2: Models used in machine learning include SVM, decision trees, k-NN. These models have applications in many areas.

Q: machine learning models

۱. مقدار P (neural|D1) با استفاده از هموار ساز JM و مقدار $\lambda=0.3$ به‌دست آورید.
(P (neural|C) = 0.001)

۲. مقدار P (SVM|D2) را با استفاده از هموارساز Dirichlet به‌دست آورید. ($\mu = 1500$)

۳. مقدار query likelihood برای P (Q|D1) و P (Q|D2) با استفاده از هموارساز JM و Dirichlet محاسبه کنید. (مقادیر پارامترها را $\lambda=0.3$ و $\mu = 1500$ در نظر بگیرید)

۴. با استفاده از هموارسازهای متفاوتی که استفاده کردید، برای هر بخش کدام سند بازگردانده می‌شود؟

۵. برتری‌های JM را نسبت به Dirichlet برای مرتب سازی اسناد تشریح کنید.

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA2_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 به‌ازای 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: **۱۲ آبان‌ماه ۱۴۰۲**

مهلت تحویل با تأخیر، با جریمه ۱۵ درصد: **۱۹ آبان‌ماه ۱۴۰۲**