به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده‌ی مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین دوم

آذر ۱۴۰۰

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 پرس وجو، 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 بدیم و همچین به بسط پرس‌وجوی کاربر با استفاده از روش Pseudo Relevance Feedback خواهیم پرداخت.

نکات قابل توجه درهنگام پاسخ به سوالات:

- در تمامی تمرین‌ها، نمره اصلی به تفسیر دانشجویان تعلق می‌گیرد (تفسیر اجباری است).
- معیارهای ارزیابی در این تمرین MAPو p@20 است.
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، به تفسیرهایی که بدون آزمایش و صرفا به صورت فرضی بیان گردند نمره‌ای تعلق نمی‌گیرد.

شرح دادگان

برای انجام این تمرین فایل‌های مشابه تمرین ۱ مورد استفاده قرار می‌گیرد، که بر روی صفحه مربوط به درس قرار دارند:

پیکره متنی[1] (فایل اسناد):

این فایل شامل اسنادی می‌باشد که 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 نمی باشد و با کتابخانه xml قابل تجزیه نیست.)

فایل پرس و جوها[2] :

این فایل شامل پرس‌وجوها می‌باشد.

فایل قضاوتهای مرتبط[3] :

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

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. جهت شاخص‌گذاری می‌توانید از دستورات موجود در ابزار گالاگو استفاده کنید.

---

[1] corpus
[2] Queries
[3] Relevance Judgment

سوال ۱- بررسی روش‌های هموارسازی

ابزار گالاگو، به صورت پیش فرض بازیابی را به روش Query-Likelihood انجام می‌دهد. هدف از این سوال آشنایی با روش‌های هموارسازی و مقایسه تاثیر هر یک بر روی کیفیت رتبه‌بندی می‌باشد.

یکی از مشکلات مطرح در حوزه بازیابی اطلاعات، وجود احتمال‌های صفر است که محاسبات را در عمل دچار مشکل می‌کند. روش‌های هموارسازی برای حل این مشکل مطرح شدند تا احتمال رخداد کلمات دیده نشده پرس‌وجو در اسناد را تخمین بزنند.

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

- روش JM با پارامتر λ
- روش Dirichlet prior با پارامتر μ
- روش Additive Smoothing

راهنمایی:

- با توجه به اینکه روش JM و Dirichlet prior در ابزار گالاگو پیاده‌سازی شده است، می‌توانید از توابع پیش‌فرض گالاگو استفاده کنید.
- با استفاده از نمودار مناسب برای هر یک از روش‌های خواسته شده مقادیر مختلف λ و μ را بررسی کنید و مقدار بهینه را مشخص نمایید.
- در مقداردهی برای پارامترها بهتر است ابتدا گام‌های بلند و سپس گام‌های کوچک آزمایش گردند تا منابع محاسباتی تلف نشود.
- تعداد اسناد بازیابی شده را 1000 و برای ریشه‌یابی از porter stemmer استفاده کنید.
- بازیابی برای پرس و جوهای ۵۱ –۱۰۰ انجام شود.

سوال ۲- هموارسازی دو مرحله‌ای

مشاهده شد که هریک از روش‌های هموارسازی مزایایی داشتند، هدف از این سوال بررسی هموارسازی دو مرحله‌ای می‌باشد که از ترکیب دو روش هموارسازی Dirichlet و JM بدست می‌آید. در این تمرین قصد داریم به بررسی این روش هموارسازی بپردازیم و نتایج بدست آمده را با سوال قبل مقایسه کرده و تحلیل کنید.

معادله این تابع هموارساز به صورت زیر می‌باشد:

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

راهنمایی:

- می‌توانید با تغییر توابع کلاس DirichletScoringIterator این روش را پیاده‌سازی نمایید.
- با استفاده از نمودار مناسب برای هر یک از روش‌های خواسته شده مقادیر مختلف λ و μ را بررسی نمایید و مقدار بهینه را شناسایی و مشخص نمایید.
- در مقداردهی برای پارامترها بهتر است ابتدا گام‌های بلند و سپس گام‌های کوچک آزمایش گردند تا منابع محاسباتی تلف نشود.
- تعداد اسناد بازیابی شده را 1000 و برای ریشه‌یابی از porter stemmer استفاده کنید.
- بازیابی برای پرس و جوهای ۵۱ –۱۰۰ انجام شود.

سوال ۳- پیاده‌سازی تابع وزن‌دهی با استفاده از Pseudo Relevance Feedback

در این سوال قصد داریم به پیاده سازی 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 سوال ابتدا فرمت پرس‌وجوهای ۵۱ تا ۱۰۰ را به فرمت زیر با پسوند tsv ذخیره کنید:

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

سپس یک پروژه جاوا ایجاد نمایید و api گالاگو را به dependency پروژه اضافه کنید، سپس فایل‌های فوق را به پروژه ایجاد شده اضافه نمایید.

1. در کلاس fbMixtureModel.java تابعی با نام computeWeights() ایجاد شده است، می‌بایست این تابع را به نحوی ویرایش کنید که با استفاده از روش EM وزن‌دهی به کلمات استخراج شده از سندهای منتخب فراهم شود.
2. با استفاده از نمودار مناسب رابطه بین تعداد سندهای منتخب برای بازخورد به ازای مقادیر بزرگتر از ۱ و معیار ارزیابی را نمایش دهید و به تحلیل نتایج بپردازید، مقدار بهینه را شناسایی و گزارش کنید.
3. با توجه به تعداد اسناد منتخب، رابطه بین تعداد کلمات استخراج شده برای بازخورد و معیار ارزیابی را با نمودار مناسب نمایش دهید و مقدار بهینه را گزارش کرده و در نهایت به تحلیل نتایج بدست آمده بپردازید.

ملاحظات (حتما مطالعه شود)

- تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA2_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 ارئه 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 بررسی ها 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: 1400/09/15
- مهلت تحویل با تاخیر و جریمه 30 درصد: 1400/09/22
- در صورت بروز هر گونه ابهام و سوال با دستیاران آموزشی طراح سوال تماس بگیرید.

beheshti.7676@gmail.com

Hoomanshirvani@ut.ac.ir