به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین سوم

آبان‌ماه ۱۴۰۲

*فهرست

بخش ۱- سوالات عملی

شرح دادگان

مجموعه دادگان مورد استفاده در این بخش، نظرات کاربران راجع به اپلیکیشن اسپاتیفای است و فایل آن از سامانه دروس قابل دریافت است. این مجموعه شامل ۶۱۵۹۴ نظر است که شامل متن نظر، زمان انتشار نظر، امتیاز به اپلیکیشن، تعداد کاربرانی که نظر را پسندیدند و پاسخ اسپاتیفای به نظر است که در این سوال فقط به متن نظر کار داریم.

فرمت فایل داده شده به صورت CSV است.

پیش‌نیازها

- جهت انجام پیاده‌سازی‌ها حتما از زبان پایتون استفاده کنید.
- پیش‌پردازش‌های لازم از قبیل حذف stopwordها، tokenization، حذف علائم نگارشی و ... را روی مجموعه داده خود اعمال کنید.
- پیشنهاد می‌شود که از کتابخانه NLTK کمک بگیرید.
- لطفا کدهای خود را تا حد امکان به صورت مرتب و به همراه کامنت بنویسید.

سؤال ۱- Word Association Mining

هدف این تمرین استخراج روابط جانشینی (Paradigmatic) و هم‌نشینی (Syntagmatic) است. برای استخراج روابط هم‌نشینی از معیار Mutual Information استفاده کنید. برای استخراج روابط جانشینی از نمایش محاسبه X_i برای هر دو روش EOWC و BM25 استفاده کنید. جهت شباهت‌سنجی برداری از شباهت کسینوسی استفاده کنید.

پس از استخراج روابط بین کلمات به سوالات زیر پاسخ دهید:

الف) ۱۰ کلمه که بیشترین رابطه جانشینی را با هر کدام از کلمات Fix و Like را دارند را به همراه امتیاز به ترتیب گزارش نمایید. دقت کنید که برای هر دو روش BM25 و EOWC این‌کار را انجام دهید.

ب) ۱۰ کلمه که بیشترین رابطه هم‌نشینی را با هر کدام از کلمات Fix و Like را دارند را به همراه امتیاز به ترتیب گزارش نمایید.

ج) با تحلیل نتایج به دست آمده در دو قسمت قبل، تفاوت دو نوع ارتباط جانشینی و هم‌نشینی بین کلمات را بیان کنید.

بخش ۲- سؤالات تئوری

سؤال ۱- Indexing

الف)

کد گاما و دلتای معادل اعداد ۹ و ۱۰۰را بنویسید. محاسبات را به طور کامل شرح بدهید.

ب)

بزرگترین بازه متوالی‌ای از اعداد طبیعی را بیابید که در این بازه، طول رشته کد دلتا و گامای هر کدام از این اعداد با هم برابر باشد.

ج)

بازه قسمت قبل را در نظر بگیرید. ثابت کنید که به ازای هر عدد بزرگتر از این بازه، طول رشته کد دلتای آن از کد گامای آن کوچک‌تر است.

سؤال ۲: Axiomatic

Lower Bounding Constraint 1 به صورت زیر تعریف می‌شود:

LB1:

Let Q be a query. Assume D_1 and D_2 are two documents such that $S(Q,D_1) = S(Q,D_2)$. If we reformulate the query by adding another term $q \notin Q$ into Q, where $c(q,D_1) = 0$ and $c(q,D_2) > 0$, then $S(Q \cup \{q\},D_1) < S(Q \cup \{q\},D_2)$.

الف)

روش Okapi مرتب‌سازی اسناد را به صورت زیر انجام می‌دهد. آیا LB1 برای این فرمول برقرار است؟ به صورت تحلیلی پاسخ خود را ثابت کنید.

$$S(Q.D) = \sum_{w \in Q \cap D} ln \frac{N - df(w) + \frac{1}{2}}{df(w) + \frac{1}{2}} \cdot \frac{(k_1 + 1) \times c(w.D)}{k_1 \left((1-b) + b \frac{|D|}{avdl}\right) + c(w.D)} \cdot \frac{(k_3 + 1) \times c(w.Q)}{k_3 + c(w.Q)}$$

ب)

بنظر شما چرا محدودیت TF-LNC تحت هر شرایطی برای روش مرتب‌سازی اسناد Pivoted Normalization Method برقرار نیست. در حد دو یا ۳ خط توضیح دهید.

$$S(Q.D) = \sum_{w \in Q.D} \frac{1 + \ln(1 + \ln(c(w.D))}{(1-s) + s \frac{|D|}{avdl}} \cdot c(w.Q) \cdot ln \frac{N+1}{df(w)}$$

ج)

در مورد محدودیت LB2 تحقیق کنید و توضیح بدهید که این محدودیت چه مفهومی را توضیح می‌دهد (در حد دو یا سه خط).

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA3_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 به طور 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: adibiali.76@gmail.com

مهلت تحویل بدون جریمه: **۵ آذرماه ۱۴۰۲**

مهلت تحویل با تأخیر، با جریمه ۱۵ درصد: **۱۲ آذرماه ۱۴۰۲**