به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۴

دی ماه ۱۴۰۰

*فهرست

شرح دادگان

برای انجام تمرین‌های بخش ۱ و ۲، قسمتی از مجموعه داده اخبار Reuters تعیین شده است. این مجموعه شامل ۵۵۰۱ سند خبری می‌باشد که هر کدام از اخبار در یکی از ۱۰ دسته مجموعه‌ی C قرار می‌گیرد.

C={acquisitions, corn, crude, earn, grain, interest, money-fx, ship, trade, wheat}

فرمت فایل داده شده به صورت CSV می‌باشد. هر سطر شامل شناسه خبر، متن خبر و دسته خبر می‌باشد.

پیش‌نیازها

- جهت انجام پیاده‌سازی‌ها حتما از زبان پایتون استفاده کنید.
- پیشنهاد می‌شود که از کتابخانه‌های NLTK و Scikit-learn کمک بگیرید. همچنین استفاده از کتابخانه های آماده برای خوشه‌بندی و محاسبه‌ی معیارهای ارزیابی خوشه‌بندی مجاز است.
- حتما پیش‌پردازش‌های لازم همچون tokenization، حذف stop-word ها، حذف علائم اختصاری و ... را روی داده ها اعمال نمایید.
- لطفا کدهای خود را تا حد امکان به‌صورت مرتب و همراه با کامنت بنویسید.
- **به صورت خلاصه نحوه پیاده‌سازیتان را بدون آوردن کد در گزارش توضیح دهید.**

بخش ۱ – Word Associations

در این تمرین هدف استخراج روابط syntagmatic و paradigmatic بین کلمات با استفاده از Mutual Information (MI) است. جهت پیاده‌سازی این روابط ابتدا MI (همراه با هموارسازی) بین کلمات را استخراج کنید. جهت شباهت‌سنجی برداری از شباهت کسینوسی استفاده نمایید.

نکته : برای ارتباط paradigmatic، بردار context کلمات را با استفاده از Mutual Information به دست آورید و سپس شباهت بردار context کلمات (شباهت کسینوسی) را به دست آورید.

۱- جهت بازنمایی بردار context کلمه کاندید W در ارتباط paradigmatic با استفاده از MI، فرمول محاسبه X_i (در اسلایدهای شماره ۱۲ و ۱۶به ترتیب برای روش EOWC و BM25 محاسبه شده است) را بازنویسی نمایید.

۲- ۱۰ کلمه‌ای که قویترین ارتباط syntagmatic با کلمات teacher و Iran دارند را به همراه امتیاز ارتباط به ترتیب گزارش نمایید.

۳- ۱۰ کلمه‌ای که قویترین ارتباط paradigmatic با کلمات Iran و teacher دارند را همراه با امتیاز ارتباط به ترتیب گزارش نمایید.

۴- با تحلیل نتایج به دست آمده در دو قسمت قبل تفاوت دو نوع ارتباط syntagmatic و paradigmatic بین کلمات را بیان کنید.

بخش ۲ – Clustering

در این بخش هدف آشنایی با انواع روش‌های خوشه‌بندی می‌باشد. جهت ایجاد نمایش برداری مجموعه اسناد از روش TF-IDF استفاده کنید. ماکزیمم اندازه واژگان را ۳۰۰۰ در نظر بگیرید.

۱- خوشه‌بندی K-Means را بر روی دادگان اخبار رویترز انجام دهید (تعداد ۱۰ خوشه) و معیارهای purity، F1،RI و NMI را برای این خوشه‌بندی محاسبه و گزارش نمایید. کدام کلاس‌ها باعث افزایش false-positives و false-negatives می‌شوند؟

۲- خوشه‌بندی سلسله مراتبی را با سه الگوریتم Average-Link، Complete-Link و Single-Link بر روی داده‌ی اخبار پیاده کنید. ۴ معیار ارزیابی فوق را برای این سه خوشه‌بندی گزارش کنید.

۳- نتایج به‌دست آمده برای انواع خوشه‌بندی‌ها را مقایسه و تحلیل نمایید (با توجه به دادگان).

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA4_StudendID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را به تفکیک هر بخش شامل شود.
- 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 به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد.
- 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 ۳۰درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

bazmi.parisa@gmail.com

مهلت تحویل بدون جریمه: **چهارشنبه ۸ دی ماه ۱۴۰۰**

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: چهارشنبه ۱۵ دی‌ماه ۱۴۰۰