تمرین برنامه‌نویسی شماره ۲

درس مبانی بازیابی اطلاعات و جستجوی وب

شاخص‌گذاری معکوس

مهلت ارسال پاسخ: ۱۳۹۶/۰۹/۲۹

۱ مجموعه دادگان[1]

دادگان در نظر گرفته شده برای این تمرین، مجموعه دادگان متنی Reuters-21578 می‌باشد. این مجموعه دادگان شامل متن تعداد ۲۱۵۷۸ مورد از خبرهاییست که در سال ۱۹۸۷ در مجله رویترز به زبان انگلیسی منتشر شده است. این مجموعه دادگان در فرمت SGML (فرمتی مشابه XML دارد) ارائه شده است، شامل تگهای متفاوت با معانی مختلف که بخشهای اطلاعاتی متفاوت را در فایل دادگان از هم مجزا نموده‌اند. جزئیات مربوط به این مجموعه دادگان و قالب و معنی تگهای مختلف آن، در فایل **readme.txt** به همراه فایلهای اصلی مجموعه دادگان قابل دسترس و مشاهده است. در اینجا برخی از تگها و خصائص مهم که به منظور حل تکلیف جاری و تکالیف بعدی لازم است، شرح داده می‌شود و در صورت نیاز به اطلاعات بیشتر می‌توانید جزئیات کاملتر را از فایل readme استخراج نمایید.

مجموعه دادگان Reuters-21578 متشکل از ۲۲ فایل با پسوند sgm می باشد که هر فایل شامل حدود ۱۰۰۰ قطعه متن متناظر با خبرهای جمع‌آوری‌شده است. در هر فایل sgm، ابتدا و انتهای قطعه مربوط به هر یک از خبرها توسط یک زوج تگ <REUTERS>....</REUTERS> شامل تعدادی خصیصه، مشخص شده است. از جمله خصائص مورد نیاز ما در این تمرین، NEWID و TOPICS است که به ترتیب بیانگر id سند (خبر) و اینکه آیا این خبر مرتبط با موضوع خاصی هست (YES) یا خیر (NO)، می‌باشد. لازم به ذکر است که خبرها ممکن است به یک یا چند موضوع (topic) مرتبط باشند.

تگ بعدی که مورد نیاز است <TOPICS>.....</TOPICS> است که بیانگر موضوع خبراست. هرکدام از موضوعاتی که خبر با آن مرتبط بوده است توسط محتوای یک زوج تگ <D>....</D> به صورت متداخل در این تگ بیان شده است.

[1] Dataset

تگ بعدی زوج تگ <TEXT>.....</TEXT> است که شامل اطلاعات متن اصلی خبر می‌باشد. داخل این تگ تعدادی تگ داخلی دیگر وجود دارد از جمله زوج تگ <TITLE>.....</TITLE> که محتوای آن شامل عنوان خبر است همچنین زوج تگ <BODY>.....</BODY> که محتوای آن بیانگر متن اصلی خبر می‌باشد.

۲ شرح تمرین

هدف این تمرین آشنایی دانشجویان با مفاهیمی همچون Term Frequency، Document Frequency و شاخص معکوس[۲] در فرایند بازیابی اطلاعات می باشد. امروزه کتابخانه‌های آماده‌ای همچون Lucene وجود دارند که بسیاری از محاسبات و عملیات موردنیاز در فرایند بازیابی اطلاعات را انجام داده و خروجی نهایی شامل inverted index را ارائه می‌دهند که قابلیت جستجو و بسیاری از عملیات دیگر را خواهد داشت. از این کتابخانه ها می‌توان به راحتی در پیاده‌سازی یک سیستم بازیابی اطلاعات استفاده نمود و از قابلیتهای آنها استفاده کرد اما هدف ما ملموس بودن بیشتر این مفاهیم برای دانشجویان می‌باشد. در نتیجه دانشجویان باید فرایند پارس نمودن اولیه محتوای مجموعه دادگان و عملیات پیش‌پردازش متون و تشکیل ساختار inverted index را خودشان به یک زبان برنامه‌نویسی دلخواه، پیاده‌سازی نمایند. مراحل کلی کار به صورت زیر می‌باشد:

۱. قدم اول تجزیه[۳] فایل‌های مجموعه دادگانی است که در اختیار دارید. برای این کار می‌توانید از کتابخانه-های موجود برای پارس نمودن فایل‌های XML مثل SAX Parser یا Lucene استفاده نمایید و یا اینکه پارسر خاص خودتان را پیاده‌سازی نمایید. لذا کارهایی که باید در مرحله تجزیه فایلهای خام مجموعه دادگان انجام دهید شامل موارد زیر است:
   - تمام فایلهای SGM را به ترتیب پارس نمایید.
   - قطعه اطلاعات متناظر با هر خبر را می‌توان به عنوان یک سند (document) متنی فرض نمود.
   - از اسناد (خبرهای) موجود در هر فایل، فقط اسنادی را پردازش کنید که مقدار مشخصه TOPICS در آنها YES باشد. از اسنادی که مرتبط با topic خاصی نیستند صرف نظر نمایید.
   - برای هر سند که شرط مورد قبل را داشت، شماره NEWID متناظر با آن و topicهایی را که با آنها در ارتباط است، استخراج نمایید.
   - عنوان و متن محتوای سند را نیز استخراج نمایید.
۲. پس از استخراج متن خام از فایل‌های دادگان نیاز است که برخی عملیات پیش‌پردازش[۴] روی آنها اعمال نمایید. عملیات پیش‌پردازش روی متون استخراج شده در عنوان و محتوای بدنه خبر اعمال می‌گردد. همانگونه که در مبحث مطرح شده در کلاس دیدیم، عملیات پیش‌پردازشی می‌تواند متنوع و به صلاحدید شما انجام شود ولی حداقل کارهایی که باید انجام دهید، شامل موارد زیر است:

[2] Inverted Index
[3] parse
[4] Preprocessing

- Tokenizing: استخراج کلمات موجود در متن عنوان و محتوای هر سند و حذف کاراکترهای ناخواسته مثل علائم و ...
- نرمال‌سازی نمایش تمام کلمات به صورت حروف کوچک انگلیسی
- اِعمال عملیات حذف stop wordها با استفاده از لیست stop wordهای داده‌شده در فایل **stopwords.txt** به همراه دیگر فایلهای مربوط به این تمرین.
- عملیات stemming: Stemming را می‌توانید با استفاده از الگوریتمها و پیاده‌سازیهای موجود از آنها مثل الگوریتم Porter انجام دهید (پیاده‌سازیهای مختلف از الگوریتم Porter به زبانهای مختلف برنامه‌نویسی وجود دارد که به راحتی قابل جستجو، دسترسی و بهره‌برداریست و نیازی نیست خودتان آن را پیاده‌سازی نمایید).

۳. ساخت شاخص معکوس

پس از انجام عملیات پیش‌پردازش و استخراج token های نهایی، می‌توان ساختار شاخص معکوس را تشکیل داد. برای انجام این مرحله به ساختار داده‌ای که قابلیت جستجو در آن و نمایش نتیجه جستجو را داشته باشد، نیاز می‌باشد. در قالب کلی ساختار داده شاخص معکوس، باید علاوه بر ID اسناد حاوی یک کلمه، به ازای هر سند، تعداد تکرار آن کلمه در سند (TF) نیز محاسبه شده و وجود داشته باشد. همچنین برای هر کلمه موجود در شاخص، تعداد اسناد حاوی آن کلمه (DF) نیز محاسبه و در شاخص معکوس ذخیره شود. ساختار داده‌ای که در پیاده‌سازی از آن استفاده می‌کنید دلخواه است ولی بهتر است از ساختارهای لیست یا map موجود در زبانهای برنامه‌نویسی باشد که مدیریت آنها ساده‌تر صورت گیرد.

۳ گزارش

۱. معرفی کتابخانه‌های استفاده شده در بخش‌های مختلف کار و توضیح مختصری درمورد نحوه استفاده از این کتابخانه‌ها مثلاً

برای parse کردن فایلهای مجموعه دادگان و یا عملیات stemming.

۲. تهیه آمار اولیه در قالب یک جدول از تعداد اسناد مفید استخراج شده از مجموعه دادگان و تعداد کل موضوعات متمایز مرتبط با این اسناد.

۳. تشریح مراحل انجام عملیات پیش‌پردازش متن و جزئیات مختصری از آن. بدیهیست اگر کار اضافه‌ای نسبت به حداقلهای بیان شده در بخش ۲-۲ از مستند جاری انجام داده اید باید در این بخش شرح دهید تا در ارزیابی مدنظر قرار گیرد. درنهایت توضیح دهید که فرایند پیش‌پردازشی که انجام داده‌اید چه تاثیری بر فرایند بازیابی خواهد داشت.

۴. تهیه آماری از تعداد کلمات متمایز موجود در کل مجموعه اسناد پردازش شده در قالب جدولی مقایسه‌ای شامل تعداد کل tokenها قبل و بعد از حذف علائم ناخواسته، تعداد tokenها پس از حذف stop wordها، تعداد tokenها پس از stemming.
۵. تشریح ساختارداده نهایی شاخص معکوس و جزئیات مربوط به آن (نحوه پیاده‌سازی، عملیات قابل انجام روی این ساختار داده، اطلاعات موجود در آن، تعداد کلمات موجود در شاخص و ...)

۴ موارد لازم جهت ارزیابی

،

- کدهای پیاده‌سازی همراه با توضیحات مفید مرتبط با کلاسها و متدها به صورت comment (فقط فایلهای کد برنامه‌ها را ارسال نمایید تا حجم بسته خیلی زیاد نشود).
- فایل گزارش کار.
- یک فایل متنی مستخرج از شاخص معکوس به این صورت که به ازای هر token موجود در شاخص، یک سطر شامل tokenID, tokenString, DocumentFrequency نوشته شده باشد.
- یک فایل متنی متناظر با شاخص معکوس به این صورت که به ازای هر token موجود در شاخص، به ازای اسنادی که این کلمه در آنها ظاهر شده است سطرهایی داشته و در هر سطر اطلاعاتی به فرمت tokenID, docID, TermFrequency نوشته شده باشد. به عنوان مثال اگر یکی از tokenها کلمه Memphis با ID برابر با ۷۸۶۵ باشد که در اسناد شماره ۲۳، ۴۵۵ و ۱۳۸۵ به ترتیب به تعداد ۵، ۳ و ۱۲.