فاز اول پروژه

در انتهای این پروژه دانشجو باید یک سیستم بازیابی اطلاعات ساده را پیاده‌سازی کند. این پروژه شامل 3 فاز مختلف است. فاز اول پروژه از دو بخش اصلی تشکیل شده است.
بخش اول آماده‌سازی اولیه‌ی داده‌ها است که شامل: یکسان‌سازی متن، جد ا کردن لغات، بازگرداندن به ریشه، حذف کلمات پرتکرار و ... می‌شود.
بخش دوم پیاده‌سازی یک **رابط کاربری** است.

مجموعه اسناد

در این پروژه برای مجموعه داده می‌توانید دو رویکرد زیر را داشته باشید:

- ✓ انتخاب مجموعه داده‌ی فارسی (در قسمت منابع درس در سس قرار داده شده است) برگرفته از مجموعه اسناد روزنامه همشهری از سال ۲۰۰۳ تا ۲۰۰۷. این مجموعه‌ی اسناد شامل سه بخش سندها، پرسمان‌ها (Queries) و اسناد مرتبط با هر پرسمان می‌شود که در پوشه‌های جداگانه در اختیار شما قرار می‌گیرد. در پوشه‌ی سندها، پوشه‌ای برای هر سال وجود دارد که در آن اسناد روزنامه‌های مرتبط با آن سال قرار دارد.

- ✓ از مجموعه داده‌ای که در درس داده کاوی ساخته‌اید استفاده کنید و آن را شبیه مجموعه داده بالا تکمیل کنید. توضیحاتی از نحوه ساخت مجموعه داده باید ارایه شود.

بخش اول: آماده‌سازی اولیه‌ی داده‌ها

این بخش با هدف آماده‌سازی لغات برای قرارگرفتن در نمایه انجام می‌شود. برای تسهیل کار شما می‌توانید از کتابخانه‌های آماده استفاده کنید. برای زبان پایتون توابع کتابخانه‌ی هضم پیشنهاد می‌شود و برای یکسان‌سازی متون انگلیسی می‌توانید از کتابخانه NLTK استفاده کنید. عملیات مورد نیاز به‌طور دقیق‌تر در زیر توضیح داده شده‌اند:

1- Normalization
2- Tokenization
3- Stop Words
4- Stemming
5- حذف علائم نگارشی
6- حذف اعداد
7- ...

نکات پیاده‌سازی

برای پیاده‌سازی این بخش یک تابع به نام Preprocess-text پیاده‌سازی کنید که متن خام را گرفته و کلمات پیش‌پردازش شده را در خروجی نشان می‌دهد.

بخش دوم: رابط کاربری

پیاده‌سازی یک واسط کاربری ساده برای اجرای تعاملی بخش‌های مختلف سیستم و همچنین مشاهده نتایج آنها ضروری است. با اجرای برنامه باید گزینه‌هایی برای اجرای بخش‌های مختلف قبلی مختلف در اختیار کاربر قرار گیرد. با انتخاب هر بخش از سمت کاربر، باید گزینه‌هایی برای اجرای زیربخش‌های هر بخش در اختیار کاربر قرار گیرد (این رابط در فازهای بعدی نیز باید تکمیل شود).

در انتهای این پروژه دانشجو باید یک سیستم بازیابی اطلاعات ساده را پیاده‌سازی کند. این پروژه شامل 3 فاز مختلف است که فاز دوم پروژه شامل دو بخش زیر است:

بخش اول: ساخت نمایه Index Construction
در این بخش باید نمایه‌ی مورد نیاز برای استفاده در بخش جستجو را بسازید. در ساخت نمایه به نکات زیر توجه فرمایید:

- نمایه‌ی شما باید پویا باشد به این معنی که امکان حذف یا افزودن سند به آن وجود داشته باشد.
- امکان ذخیره‌سازی و بارگیری نمایه نیز باید فراهم باشد.

نمایه‌های مورد انتظار برای پیاده‌سازی:

- Non-positional index
- positional index
- نمایه برای تشخیص عبارات wildcard

بخش دوم: فشرده‌سازی نمایه Index Compression
در این بخش هدف فشرده‌سازی نمایه‌های ساخته شده به دو روش Variable byte و Gamma code است. میزان حافظه اشغالی **قبل و بعد** از اعمال هر دو فشرده‌سازی باید مشخص شود.

این فاز نیز شامل دو بخش زیر است:

بخش اول: جستجو و بازیابی اسناد

در این بخش انتظار می‌رود که دانشجو دو نوع جستجوی ترتیب‌دار و دقیق را که در زیر توضیح داده می‌شوند، پیاده‌سازی نماید:

- جستجوی ترتیب‌دار در فضای برداری tf-idf: پس از دریافت کوئری ورودی و نوع جستجو، فهرستی از اسناد مرتبط به ترتیب امتیاز خروجی می‌دهد. ممکن است کوئری ورودی شامل یک یا چندین لغت wildcard باشد. هر ترکیب از لغات نمایه معادل با این لغات باید یک‌بار با آنها جایگزین شوند و درنهایت بین همه اسناد بازگردانی شده به ازای ترکیب‌های مختلف لغات معادل با لغاتwildcard اسنادی که بیشترین امتیاز را کسب کرده‌اند بازگردانی شوند.
- جستجوی دقیق:phrasal search : کوئری ورودی این نوع جستجو شامل تعدادی لغت و عبارات داخل گیومه است. برای سادگی کوئری‌های این قسمت شامل لغات wildcard نیستند. اسناد بازیابی شده می‌بایست شامل عبارات داخل گیومه باشند و در لغات داخل این عبارات ترتیب لغات نیز حفظ شود ولی ترتیب عبارات نسبت به هم لزومی ندارد مطابق کوئری باشد. به‌عنوان نمونه برای کوئری "q4 q3" "q1 q2" سند "q5 q1 q2 q3 q4 q5" مرتبط محسوب می‌شود. دقت نمایید که خروجی این قسمت نیز باید ترتیب‌دار باشد. به این صورت که ابتدا مجموعه‌ی تمامی اسناد دارای عبارات داخل گیومه را پیدا کرده و سپس با استفاده از تمام لغات داخل کوئری (شامل لغات داخل گیومه) بازیابی ترتیب‌دار را بر روی این مجموعه از اسناد انجام می‌دهید.

بخش دوم: ارزیابی سیستم

در مجموعه اسناد موجود علاوه بر فایل اسناد، تعدادی کوئری و نتیجه آنها در اختیار شما قرار گرفته است، در این بخش سیستم شما باید مجموعه کوئری‌ها و پاسخ‌های درست برای هر کوئری را دریافت کند و با مقایسه پاسخ سیستم با نتایج درست سیستم شما را ارزیابی کند. برای ارزیابی باید ۲ معیار F-measure و MAP را پیاده‌سازی کنید.

توجه داشته باشید که سیستم شما باید قابلیت محاسبه هرکدام از این معیارها را بر روی روش‌های متفاوتی که برای بازیابی اسناد پیاده‌سازی کردید به‌طور جداگانه داشته باشد. برای مدل‌های بازیابی ترتیب‌دار حداکثر سند بازیابی شده را برابر با ۵ قرار دهید.