بسم الله الرحمن الرحیم

فاز اول پروژه درس بازیابی پیشرفته اطلاعات

سیستم بازیابی اطلاعات روزنامه‌ی همشهری

مدرس: دکتر سلیمانی

نیم‌سال اول سال تحصیلی ۹۶ - ۹۷

دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف

مقدمه

در این پروژه شما می‌بایست یک سیستم بازیابی اطلاعات بر روی مجموعه‌ی روزنامه‌ی همشهری پیاده‌سازی کنید. این پروژه از چهار بخش اصلی تشکیل شده است. بخش اول آماده‌سازی اولیه داده‌هاست که شامل یکسان‌سازی متن، جدا کردن لغات و بازگرداندن به ریشه می‌شود که برای این اعمال می‌توانید از کتابخانه‌ی هضم استفاده کنید. هم‌چنین در این بخش باید لغات پرکاربرد اسناد را بدست آورید و آنها را حذف نمایید. بخش دوم پیاده‌سازی یک نمایه‌ساز است که با دریافت آدرس اسناد نمایه‌های مورد نیاز برای جستجو را می‌سازد. بخش سوم امکان جست‌وجو بر روی اسناد را می‌دهد و در بخش چهارم سیستم ارزیابی می‌شود. در ادامه هر کدام از این بخش‌ها به طور کامل توضیح داده‌می‌شوند.

مجموعه اسناد

در این پروژه از مجموعه اسناد برگرفته از روزنامه همشهری از سال ۲۰۰۳ تا ۲۰۰۷ استفاده می‌شود که از مجموعه‌ی همشهری[1] انتخاب شده‌اند. این مجموعه‌ی اسناد شامل سه بخش سندها، پرسمان‌ها 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. در پوشه پرسمان‌ها، ۵۰ پرسمان وجود دارد که برای ارزیابی در اختیار شما قرار داده‌شده‌اند و در پوشه داوری فایلی قرار دارد که شامل لیست اسناد مرتبط با هر پرسمان است. در هر سطر از فایل داوری شماره‌ی پرسمان و نام سند مرتبط با آن قرار دارد.

---

[1] Abolfazl AleAhmad , Hadi Amiri , Ehsan Darrudi , Masoud Rahgozar , Farhad Oroumchian, **Hamshahri**: **A standard Persian text collection**, Journal of Knowledge-Based Systems, Vol. 22 No.5, p.382-387, Elsevier, July 2009.

بخش اول: آماده‌سازی اولیه‌ی داده‌ها

این بخش با هدف آماده‌سازی لغات برای قرارگرفتن در نمایه انجام می‌شود. برای تسهیل کار شما می‌توانید از توابع کتابخانه‌ی هضم[2] که امکان استفاده از آن در زبان پایتون وجود دارد استفاده کنید. عملیات‌های مورد نیاز به طور دقیق‌تر در زیر توضیح داده شده‌اند:

۱ـ یکسان‌سازی متن (Normalization): برای این کار می‌توانید از کلاس Normalizer در کتابخانه‌ی هضم استفاده کنید.

۲ـ جداسازی لغات (Tokenization): برای این کار می‌توانید از تابع word_tokenize در کتابخانه هضم استفاده کنید.

۳ـ یافتن و حذف لغات پرکاربرد (Stop Words): برای پیدا کردن لغات پرکاربرد می‌توانید با بررسی تعداد تکرار هر لغت در تمامی اسناد، لغات با بیشترین تکرار را انتخاب کنید و سپس آن‌ها را حذف نمایید.
دقت کنید که در جستجوی دقیق (phrasal search) که در بخش سوم توضیح داده می‌شود، این کلمات به عنوان یک کلمه به حساب خواهند آمد.

۴ـ بازگردانی به ریشه (Stemming):برای این کار می‌توانید از کلاس Stemmer در کتابخانه هضم استفاده کنید.

در صورتی که می‌خواهید از نسخه جاوا این کتاب‌خانه استفاده کنید، می‌توانید توضیحات صورت پروژه‌ی سال گذشته[3] را مطالعه بفرمایید.

بخش دوم: ساخت نمایه

در این بخش باید نمایه‌ی مورد نیاز برای استفاده در بخش جست‌وجو 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.
- داده‌ساختار استفاده شده برای نمایه‌ها باید مطابق با داده‌ساختارهای تدریس شده در درس باشد.

2 http://www.sobhe.ir/hazm/

3 http://ce.sharif.edu/courses/95-96/1/ce324-1/assignments/files/assignDir/MIR_Project1.pdf

- امکان ذخیره‌سازی و بارگیری نمایه نیز باید فراهم باشد.

نمایه‌های مورد انتظار برای پیاده‌سازی:

- نمایه جای‌گاهی (positional index):

برای این قسمت بایستی نمایه‌ای بسازید که با استفاده از آن بتوان شماره تمامی اسنادی که یک کلمه در آن آمده است و همچنین همه جایگاه‌های این کلمه در هر سند را پیدا کرد.

- نمایه برای تشخیص عبارات wildcard

برای پشتیبانی کردن سیستم از عبارات wildcard در جستجو، نیازمند داشتن داده‌ساختاری مناسب هستید تا بتوانید کلمات موجود در لغت‌نامه را که در پرسمان صدق می کنند بیابید. برای سادگی تصمیم گرفته شده که تنها از پرسمانهایی که * در انتهای آنها قرار دارد، پشتیبانی کنید. پشتیبانی از پرسمان‌هایی که * در وسط آن‌هاست امتیازی محسوب می‌شود.

بخش سوم: جست‌وجو و بازیابی اسناد

در این بخش انتظار می‌رود شما دو نوع جستجوی ترتیب‌دار و دقیق را که در زیر توضیح داده می‌شوند، پیاده سازی نمایید:

- جستجوی ترتیب‌دار در فضای برداری tf-idf به دو روش lnn-ltn و lnc-ltc : پس از دریافت پرسمان ورودی و نوع جستجو، لیستی از اسناد مرتبط به ترتیب امتیاز خروجی می‌دهد. ممکن است پرسمان ورودی شامل یک یا چندین لغت wildcard باشد. هر ترکیب از لغات نمایه معادل با این لغات باید یک بار با آنها جایگزین شوند و در نهایت بین همه اسناد بازگردانی شده به ازای ترکیب‌های مختلف لغات معادل با لغات wildcard، اسنادی که بیشترین امتیاز را کسب کرده‌اند بازگردانی شوند. به عنوان مثال اگر لغات معادل روز* = {روزنامه، روزگار} و هم* = {همشهری، همسایه} باشند، برای پرسمان روز* هم*، باید چهار پرسمان رونامه همشهری، روزنامه همسایه، روزگار همشهری و روزگار همسایه در نظر گرفته‌شوند و بین کل این چهار مجموعه سند، اسنادی که بیشترین امتیاز را کسب کرده‌اند به عنوان اسناد مرتبط با پرسمان اولیه بازگردانی شوند.
- جستجوی دقیق(phrasal search): پرسمان ورودی این نوع جستجو شامل تعدادی لغت و عبارات داخل گیومه است. برای سادگی پرسمان‌های 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 پرسمان باشد. به عنوان نمونه برای پرسمان "q4 q5" q3 "q1 q2" سند q4 q5 q1 q2 q3 مرتبط محسوب می‌شود. دقت نمایید که خروجی این قسمت نیز می‌بایست ترتیب‌دار باشد. به این صورت که ابتدا مجموعه‌ی تمامی اسناد دارای عبارات داخل گیومه را پیدا می‌کنید سپس با استفاده از تمام لغات داخل پرسمان (شامل لغات داخل گیومه) بازیابی ترتیب‌دار را بر روی این مجموعه از اسناد انجام می‌دهید.
  - مثال:

پرسمان : "اقتصاد جهان" خصوصي سازي سابقه
سند مرتبط : خصوصي سازي در اقتصاد جهان بيش از سي سال سابقه دارد
سند غير مرتبط : خصوصي سازي در اقتصاد ايران و جهان بيش از سي سال سابقه دارد

بخش چهارم: ارزیابی سیستم

در مجموعه اسناد موجود علاوه بر فایل اسناد، تعدادی پرسمان و نتیجه آنها در اختیار شما قرار گرفته است، در این بخش سیستم شما باید مجموعه پرسمان‌ها و پاسخ‌های درست برای هر پرسمان 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 پیاده سازی کنید.

توجه داشته باشید که سیستم شما باید قابلیت محاسبه هر کدام از این معیارها را بر روی روشهای متفاوتی که برای بازیابی اسناد پیاده سازی کردید به طور جداگانه داشته باشد. برای مدل‌های بازیابی ترتیب‌دار حداکثر سند بازیابی شده را برابر با ۲۰ قرار دهید.

رابط کاربری

پیاده‌سازی یک واسط کاربری ساده تحت کنسول برای اجرای تعاملی بخشهای مختلف سیستم و همچنین مشاهده نتایج آنها ضروری است. با اجرای برنامه می‌بایست چهار گزینه برای اجرای چهار بخش مختلف در اختیار کاربر قرار گیرد. با انتخاب هر بخش از سمت کاربر، می‌بایست گزینه‌هایی برای اجرای زیربخش‌های هر بخش در اختیار کاربر قرار گیرد.

بخش اول (۱۰ نمره)

- دریافت متن فارسی از کاربر و نمایش هر یک از کلمات آن پس از اعمال عملیات‌های مربوط به آمادگی اولیه داده‌ها بر روی آنها (۶ نمره)
- امکان مشاهده‌ی لیست لغات پرکاربرد (۴ نمره)

بخش دوم (۳۷ نمره + ۱۰ نمره امتیازی)

- ساخت نمایه‌ها از روی پوشه‌ی اسناد (۱۵ نمره)
- اضافه کردن یک سند با وارد کردن نام آن توسط کاربر (۵ نمره)

- حذف یک سند با واردکردن نام آن از توسط کاربر (۵ نمره)
- ذخیره‌سازی نمایه در یک فایل با فرمت دلخواه (۵ نمره)
- بارگیری نمایه از فایل ذخیره‌شده (۲ نمره)
- مشاهده posting list یک کلمه شامل تمام اسنادی که کلمه در آن‌ها ظاهر شده‌است و جایگاه کلمه در هر سند
- نمایش تمامی کلمات مطابق با یک لغت wildcard (۵ نمره + ۱۰ نمره امتیازی)

بخش سوم (۳۸ نمره)

انجام عملیات جستجو که شامل گام‌های زیر است:

- انتخاب نوع جستجو (ترتیب‌دار یا دقیق)
- دریافت پرسمان
- انتخاب نوع بازیابی (lnn-ltn یا lnc-ltc)
- نمایش لیست اسناد مرتبط (۱۵ نمره برای بازیابی ترتیب‌دار + ۲۰ نمره برای بازیابی دقیق)
- امکان انتخاب سند و نمایش محتویات آن (۳ نمره)

بخش چهارم (۱۵ نمره)

- دریافت شماره‌ی پرسمان و نام معیار از کاربر
- نمایش دقیق مقدار محاسبه شده (۱۲ نمره)
- ارزیابی سیستم روی همه‌ی پرسمان‌های موجود در صورت وارد شدن کلمه all از سمت کاربر (۳ نمره)

نکات پایانی

- برای پارامترهایی که مقدار آن‌ها در صورت پروژه ذکر نشده است می‌توانید مقداری دلخواه انتخاب نمایید و یا از کاربر دریافت نمایید.
- در صورت عدم تطابق داده‌ساختارهای پیاده‌سازی‌شده با داده‌ساختارهای تدریس شده در کلاس، نمره‌ای به آن‌ها تعلق نمی‌گیرد.
- برای ارسال پروژه، صرفا کد خود را به صورت zip شده در سایت کوئرا آپلود نمایید.
- سوالات خود را در قسمت پرسش و پاسخ سایت پیاتزا مطرح فرمایید.
- موعد تحویل پروژه ساعت ۲۳:۹۵ روز جمعه ۱۲ آبان است و جریمه‌ی تأخیر مطابق با قوانینی که در سایت درس آپلود شده‌است[4]، خواهد بود.
- لطفا تمام قسمتهای پروژه را خودتان و به تنهایی پیاده‌سازی فرمایید. در صورت تخلف برخورد جدی صورت خواهد گرفت.

موفق باشید

[4] http://ce.sharif.edu/courses/96-97/1/ce324-1/resources/root/regulations.pdf