۱ مقدمه

در این پروژه می‌خواهیم یک سیستم بازیابی اطلاعات متشکل از ۴ بخش روی پیکره‌ی روزنامه‌ی همشهری [۱] پیاده‌سازی نماییم. مجموعه اسناد مورد استفاده برگرفته از روزنامه‌ی همشهری از سال ۲۰۰۳ تا ۲۰۰۷ می‌باشد. این پیکره به صورت سه بخش مجموعه اسناد، پرسمان‌ها و اسناد مرتبط با هر پرسمان در اختیار شما قرار می‌گیرد.

۲ بخش اول: آماده‌سازی داده‌ها

هدف از انجام این بخش آماده‌سازی لغات برای قرار گرفتن در نمایه می‌باشد. شما می‌توانید از کتابخانه‌ی هضم [۱] (قابل استفاده در زبان پایتون و جاوا) برای انجام موارد زیر استفاده نمایید:

۱. جداسازی لغات (Tokenization): برای این کار می‌توانید از تابع word_tokenize کتابخانه هضم استفاده نمایید.

۲. یکسان‌سازی متن (Normalization): برای این کار می‌توانید از کلاس Normalizer این کتابخانه استفاده نمایید.

۳. یافتن و حذف لغات پرکاربرد (Stop Words): برای یافتن لغات پرکاربرد می‌توانید تعداد تکرار هر لغت در تمامی اسناد را محاسبه و لغات با بیشترین تکرار را حذف نمایید. البته برای جستجوی دقیق (phrasal search) لازم است این کلمات استفاده شوند.

۴. بن‌واژه‌یابی (Stemming): برای انجام این کار نیز می‌توانید از کلاس Stemmer کتابخانه هضم استفاده بکنید.

نمره‌دهی

- دریافت متون فارسی از کاربر و نمایش هر یک از کلمات آن پس از انجام عملیات‌های ذکر شده (**۷ نمره**)
- امکان مشاهده‌ی لیست لغات پرکاربرد (**۳ نمره**)

۳ بخش دوم: ایجاد شاخص‌ها

در این قسمت لازم است تا شاخص‌های مورد نیاز برای استفاده در بخش جستجو را پیاده‌سازی نمایید.
شاخص‌های مورد انتظار برای پیاده‌سازی:

- شاخص مکانی (positional index): شاخص ساخته شده می‌بایست علاوه بر شماره اسناد جایگاه کلمات را نیز بیابد.
- شاخص برای عبارات خاص (wildcards): برای این حالت لازم است از داده‌ساختاری استفاده کنید که از پرسمان‌های دارای * نیز پشتیبانی کند.

توجه داشته باشید که شاخص ساخته شده می‌بایست پویا بوده و امکان اضافه کردن و حذف اسناد وجود داشته باشد. همچنین می‌بایست بتوان شاخص را ذخیره و یا بارگیری کرد.

[۱] http://www.sobhe.ir/hazm/

نمره‌دهی

- ساخت شاخص‌ها (**۱۵ نمره**)
- امکان اضافه کردن و حذف اسناد (**۷ نمره**)
- امکان ذخیره و بارگیری (**۵ نمره**)
- نمایش posting list کلمات :اسناد شامل کلمه و موقعیت‌های مکانی کلمه در آن سند (**۵ نمره**)
- نمایش تمامی کلمات مطابق با یک wildcard (**۳ نمره**)

۴ بخش سوم: جستجو و بازیابی اطلاعات

در این قسمت انتظار می‌رود دو نوع جستجوی ترتیب‌دار و دقیق پیاده‌سازی شوند.

۱. جستجوی ترتیب‌دار در فضای برداری tf-idf به دو روش lnn-ltn و lnc-ltc (البته می‌توانید روش‌های بیشتری را نیز در نظر بگیرید): پس از دریافت پرسمان و نوع جستجو و نحوه‌ی امتیازدهی، لیستی از اسناد مرتبط (حداکثر ۲۰ سند مرتبط) را به ترتیب امتاز خروجی می‌دهد. ممکن است پرسمان ورودی شامل یک یا چند لغت wildcard باشد که در این صورت می‌بایست تمام ترکیب‌های معادل با هر wildcard جایگزین شده و در نهایت اسناد بر اساس پارامترهای مربوطه امتیازدهی شوند.

۲. جستجوی دقیق (phrasal search): در این حالت پرسمان ورودی شامل تعدادی لغت و عبارات داخل گیومه است. در این حالت فرض می‌کنیم که این نوع پرسمان‌ها شامل لغات wildcard نمی‌باشند. اسناد بازیابی شده می‌بایست دقیقا شامل عبارات مربوطه باشند. توجه داشته باشید که بازیابی این قسمت نیز بصورت ترتیبی می‌باشد.

نمره‌دهی

- نمایش لیست اسناد مرتبط (**۱۵ نمره ترتیب‌دار و ۱۵ نمره دقیق**)
- امکان انتخاب سند یافته شده و نمایش محتوای آن (**۵ نمره**)

۵ بخش چهارم: تحلیل و ارزیابی نتایج

با استفاده از مجموعه‌ی پرسمان و نتایج مربوطه که در پیکره گنجانده شده‌اند می‌توانیم سیستم ساخته شده و نتایج آن را ارزیابی نماییم. در این قسمت شما می‌بایست مجموعه پرسمان و پاسخ‌های درست را دریافت نموده و ارزیابی لازم را با مقایسه آن‌ها با پاسخ‌های سیستم پیاده‌سازی شده انجام دهید. برای ارزیابی از ۲ معیار F-Measure و MAP استفاده نمایید.

نمره‌دهی

- دریافت شماره پرسمان و نام معیار از کاربر و نمایش دقیق مقدار محاسبه شده (**۱۵ نمره**)
- امکان ارزیابی همه‌ی پرسمان‌های موجود با توجه معیار انتخاب شده و نمایش مقدار محاسبه شده (**۵ نمره**)

مراجع

[1] AleAhmad, Abolfazl, Amiri, Hadi, Darrudi, Ehsan, Rahgozar, Masoud, and Oroumchian, Farhad. Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382 – 387, 2009.