بازیابی پیشرفته اطلاعات

نیم‌سال اوّل ۰۰-۹۹

دانشکده‌ی مهندسی کامپیوتر

مدرس: حمید بیگی

فاز اول پروژه (۱۰۰ نمره) زمان تحویل: ۲۳ آبان

هدف از فاز اول پروژه پیاده‌سازی یک سیستم بازیابی اطلاعات است. فاز اول پروژه از ۵ بخش کلی تشکیل شده است و دو مجموعه داده نیز در اختیار شما قرار دارد. مجموعه‌ی اول که به زبان انگلیسی است، شامل مجموعه اطلاعات بخشی از سخنرانی‌های Ted Talk است. توجه کنید که این مجموعه‌داده شامل تعدادی زیادی ستون است که شما در این پروژه تنها از ستون‌های title و description استفاده می‌کنید. مجموعه داده دوم به زبان فارسی است و بخشی از پیکره‌ی ویکی‌پدیای فارسی شامل چندین صفحه ویکی‌پدیا به فرمت xml است. توجه نمایید که تمامی صفحات در یک فایل xml قرار دارند و شما باید ابتدا با بررسی فایل و فهمیدن الگوی ذخیره‌سازی، هر صفحه را به صورت جدا استخراج نمایید زیرا هر صفحه یک مستند (document) مجزا است.

بخش اول به پیش‌پردازش متنی داده‌ها می‌پردازد که شامل یکسان‌سازی متن، جداسازی لغات، حذف لغات پرتکرار و ... است. بخش دوم نمایه‌سازی است. در بخش سوم باید روی این نمایه فشرده‌سازی صورت بگیرد. در ادامه، قسمت جستجو و بازیابی سیستم قرار دارد. در بخش چهارم باید پرسمان ورودی کاربر را تصحیح کرده و در بخش پنجم از نمایه‌های پیاده‌سازی شده برای جست‌وجو استفاده شود.

بخش ۱. پیش پردازش اولیه (۲۵ نمره)

در این بخش از پروژه ابتدا باید مجموعه فایل‌هایی که در اختیارتان قرار گرفته است را بخوانید. سپس به ترتیب مراحل پیش‌پردازش متنی که در ادامه آمده است را روی آن‌ها اعمال کنید. برای اعمال پیش رو می‌توانید از کتابخانه‌های آماده استفاده کنید. برای زبان پایتون کتابخانه هضم پیشنهاد می‌شود. برای یکسان‌سازی متون انگلیسی می‌توانید از کتابخانه NLTK استفاده کنید.

۱. **نرمال‌سازی متنی (normalization)**: برای یکسان سازی متون می‌توانید از توابع کتابخانه‌های معرفی شده استفاده کنید. اما در صورتی که می‌خواهید خودتان پیاده‌سازی کنید باید پیاده‌سازی‌تان شامل برگرداندن لغات به ریشه، case folding (برای یکسان سازی متون انگلیسی) و بقیه مواردی که در درس بیان شده است باشد.

۲. **جداسازی (tokenization)**: برای این کار می‌توانید از توابع کتابخانه‌های معرفی شده استفاده کنید.

۳. **حذف علائم نگارشی**: هر کدام از مجموعه متن‌ها یک سری علائم نگارشی مثل نقطه، ویرگول و ... دارند که آنها را باید حذف کنید.

۴. **یافتن و حذف لغات پرتکرار (stopwords)**: در این بخش، حذف درصد معقولی از لغات پرتکرار مورد نظر است. برای این منظور لازم است تا همه متن را پردازش کنید و نسبت به حجم متن، کلماتی که پرتکرار هستند را نمایش دهید. این نسبت را طوری در نظر بگیرید که کلمات پرتکرار به دست آمده، تا حد خوبی منطقی و کافی باشند.

۵. **بازگرداندن کلمات به ریشه (stemming)**: در نهایت افعال، اسامی و ... را به حالت ساده و پایه ای خود برگردانید.

نکات پیاده‌سازی

برای پیاده‌سازی این بخش یک تابع به نام prepare-text پیاده‌سازی کنید که متن خام را می‌گیرد و کلمات پیش‌پردازش شده را خروجی می‌دهد. برای نمایش لغات پرتکرار می‌توانید از هیستوگرام و یا لیست ساده‌ای از کلمات استفاده کنید.

بارم‌بندی

۱. گرفتن متن از کاربر و نمایش لغات آن بعد از پیش‌پردازش متنی (۱۵ نمره)

۲. نمایش لغات پرتکرار (از متون در اختیار قرار گرفته) (۱۰ نمره)

بخش ۲. نمایه‌سازی (۲۵ نمره)

در این بخش پیاده‌سازی نمایه جایگاهی (Positional) و نمایه Bigram مطلوب است. برای نمایه جایگاهی باید به ازای هر لغت، لیستی از اسناد شامل آن لغت و جایگاه(ها) هر لغت در آن سند را داشته باشید و برای نمایه Bigram نیز ترکیب‌های دو حرفی تمامی کلمات موجود در لغت‌نامه که این ترکیب در آنها موجود است را ذخیره کنید. این نمایه برای قسمت اصلاح پرسمان مورد استفاده قرار خواهد گرفت. نمایه شما باید پویا باشد یعنی با حذف سند از نمایه نیز حذف شده و با اضافه کردن سند در طول اجرای برنامه به نمایه اضافه شود. همچنین بعد از نمایه‌سازی باید قادر باشید نمایه را در فایلی ذخیره کرده و از آن بخوانید. پویا بودن نمایه و ذخیره‌سازی آن را برای هر دو نوع نمایه در نظر بگیرید.

نکات پیاده‌سازی

برای سادگی پیاده‌سازی برای هر کارکرد خواسته شده در توضیحات بالا یک تابع پیاده‌سازی کنید. برای مثال دو تابع برای حذف و اضافه کردن سند به نمایه، توابعی برای ذخیره‌سازی و لود کردن نمایه و غیره.

بارم‌بندی

۱. نمایه‌سازی از روی پوشه‌های در اختیار قرار داده شده (۱۵ نمره)

۲. نمایش posting list کلمه ورودی توسط کاربر (۴ نمره)

۳. نمایش جایگاه کلمه وارد شده توسط کاربر در هر سند (۳ نمره)

۴. مشاهده‌ی تمام کلماتی که دارای یک bigram خاص درون خود هستند (۳ نمره)

بخش ۳. فشرده‌سازی نمایه‌ها (۱۵ نمره)

در این بخش هدف فشرده‌سازی نمایه‌های ساخته‌شده به دو روش variable byte و gamma code است. (برای ذخیره‌سازی در فایل و بخش‌های بعدی می‌توانید فقط یکی از این دو روش را به انتخاب خود ادامه دهید.)

نکات پیاده‌سازی

برای هر دو نوع فشرده‌سازی یک تابع پیاده‌سازی کنید. نهایتا از توابعی که در بخش قبل برای ذخیره‌سازی پیاده‌سازی کردید برای بخش سوم بارم‌بندی استفاده کنید.

بارم‌بندی

۱. پیاده‌سازی و نمایش میزان حافظه اشغال شده قبل و بعد از اعمال variable byte (۵نمره)

۲. پیاده‌سازی و نمایش میزان حافظه اشغال شده قبل و بعد از اعمال gamma code . (۵ نمره)

۳. ذخیره سازی نمایه‌ها در فایل و بارگذاری از آن (۵ نمره)

بخش ۴. اصلاح پرسمان (۱۰ نمره)

در صورتی که پرسمان ورودی دارای غلط املایی باشد، یا به عبارتی لغت(هایی) از آن در لغت‌نامه موجود نباشد، لازم است که با جستجوی لغت‌های احتمالی و انتخاب بهترین لغت به ادامه‌ی جستجو با پرسمان اصلاح شده پرداخته شود. برای اینکار ابتدا باید به وسیله‌ی روش bigram و معیار jaccard نزدیک‌ترین لغات به لغت با غلط املایی را پیدا کنید. سپس بهترین لغت از میان آن‌ها را با استفاده از معیار edit distance بیابید.

نکات پیاده‌سازی

برای این بخش یک تابع پیاده‌سازی کنید که ورودی خام را گرفته و متن تصحیح شده‌ی آن را نمایش دهد. دقت کنید ورودی و خروجی هر دو رشته‌ی متنی هستند و نه لیستی از کلمات.

بارم‌بندی

۱. نمایش پرسمان اصلاح شده (۵ نمره)

۲. محاسبه‌ی فاصله‌ی جاکارد دو کلمه (۲ نمره)

۳. محاسبه‌ی فاصله‌ی ویرایش دو کلمه بدون استفاده از کتاب‌خانه‌های آماده (۳ نمره)

بخش ۵. جستجو و بازیابی اسناد (۲۵ نمره)

در این بخش دو روش جستجو باید به صورتی که در ادامه توضیح داده شده است، پیاده‌سازی شوند. البته توجه نمایید که روش دوم جستجو امتیازی است.

۱. جستجوی ترتیب‌دار در فضای برداری tf-idf به روش lnc-ltc : در این روش جستجو بعد از دریافت پرسمان ورودی، باید لیستی از اسناد مرتبط به ترتیب امتیاز نمایش داده شود.

۲. جستجوی proximity با اندازه‌ی پنجره‌ی وارد شده در ورودی: در این روش جستجو ابتدا باید اسنادی که تمام کلمات پرسمان در یک بازه‌ای به اندازه‌ی پنجره‌ی داده شده، در آن سند وجود داشته باشند، پیدا شوند. سپس از بین آن ها به ترتیب امتیازشان براساس جستجوی ترتیب دار در فضای بردار tf-idf به روش lnc-ltc داک ها نمایش داده شوند.

نکات پیاده‌سازی

برای هر دو نوع جستجو نمایش ۱۰ سند در صورت موجود بودن کافی می باشد.

بارم‌بندی

۱. نمایش لیست اسناد مرتبط به ترتیب شباهت در جستجوی ترتیب‌دار در فضای برداری tf-idf به روش lnc-ltc (۱۰ نمره)

۲. نمایش لیست اسناد مطابق با پرسمان و اندازه پنجره ورودی در جستجوی proximity (۱۵ نمره)

بخش ۶. نکات

۱. باید یک واسط کاربری برای تست موارد مختلف مشخص شده در قسمت بارم‌بندی هر بخش، برای تحویل حضوری وجود داشته باشد. واسط کاربری می‌تواند تحت کنسول پیاده سازی شود.

۲. علاوه بر واسط کاربری که بخش‌های مختلف پروژه‌ی شما را اجرا می‌کند، لازم است گزارشی از پیاده‌سازی هر بخش تهیه کنید و آن را به همراه فایل‌های پروژه آپلود کنید. در این گزارش نتیجه‌ی اجرای هر بخش را به صورت اسکرین شات یا توضیح بیاورید. پروژه‌های بدون گزارش برای تصحیح در نظر گرفته نمی‌شوند.

۳. تمامی فایل‌های پروژه به همراه گزارش را در یک فایل زیپ در سامانه‌ی کوئرا آپلود نمایید.

۴. امکان تغییر بارم‌بندی وجود دارد.

۵. تنها زبان برنامه‌سازی مجاز پایتون است.

۶. کدها از نظر شباهت بررسی خواهد شد و با موارد تقلب طبق آیین‌نامه‌ی تمارین درسی برخورد می‌شود.