**به نام خدا**

**بازیابی پیشرفته اطلاعات**
**نیم‌سال اول ۹۷-۹۸**
**مدرس: دکتر حمید بیگی**

---

**فاز اول پروژه** **تاریخ تحویل: ۹۷/۰۸/۰۴**

---

هدف از این پروژه پیاده‌سازی یک سیستم بازیابی اطلاعات است. پروژه از ۵ بخش تشکیل شده است. برای انجام این پروژه دو مجموعه داده در اختیارتان قرار گرفته است. مجموعه اول، اسناد برگرفته از روزنامه همشهری است. مجموعه دوم، اسناد برگرفته از این صفحه است.(این داده‌ها به زبان انگلیسی هستند.)
بخش اول پروژه به پیش پردازش متنی داده‌ها می‌پردازد، که شامل یکسان سازی متن، جدا سازی لغات و حذف لغات پرتکرار و ... است. بخش دوم نمایه‌سازی است. در بخش سوم نیز باید روی این نمایه فشرده‌سازی صورت بگیرد. در ادامه، قسمت جستجو و بازیابی سیستم قرار دارد که در بخش چهارم پرسمان ورودی کاربر را باید تصحیح کرده و در بخش پنجم از نمایه‌های پیاده‌سازی شده برای جست‌وجو استفاده می‌شود.

**بخش اول:** پیش پردازش اولیه
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 پیش‌رو می‌توانید از کتابخانه‌های آماده استفاده کنید. برای زبان پایتون کتابخانه هضم و جاوا jhazm 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 و ویرگول و ... دارند که آنها را باید حذف کنید.
۴. **یافتن و حذف لغات پرتکرار (stop words):** در این بخش، حذف درصد معقولی از لغات پرتکرار مورد نظر است.
۵. **بازگرداندن کلمات به ریشه (stemming):** در نهایت افعال و اسامی و ... را به حالت ساده و پایه ای خود برگردانید.

**بارم بندی**
گرفتن متن از کاربر و نمایش لغات آن بعد از پیش‌پردازش متنی (۱۵ نمره)
نمایش لغات پرتکرار (از متون در اختیار قرار گرفته) (۱۰ نمره)

**بخش دوم:** نمایه‌سازی

در این بخش پیاده‌سازی نمایه جایگاهی (positional index) و نمایه Bigram مطلوب است. برای نمایه جایگاهیی 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 نگه‌دارید. این نمایه برای قسمت اصلاح پرسمان که در بخش بعد توضیح داده خواهد شد، 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.

**بارم بندی**

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

نمایش posting list کلمه ورودی توسط کاربر (۵ نمره)

نمایش جایگاه کلمه وارد شده توسط کاربر در هر سند (۵ نمره)

**بخش سوم:** فشرده‌سازی نمایه‌ها

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 ادامه دهید.)

**بارم بندی**

نمایش میزان حافظه اشغال شده قبل و بعد از اعمال variable bytes.

(۵ نمره)

نمایش میزان حافظه اشغال شده قبل و بعد از اعمال gamma code.

(۵ نمره)

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

**بخش چهارم:** اصلاح پرسمان

در صورتی که پرسمان ورودی دارای غلط املایی باشد، یا به عبارتی لغت(هایی) از آن در دیکشنری 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.

**بارم‌بندی**

نمایش پرسمان اصلاح شده (۲۰ نمره)

**بخش پنجم:** جستجو و بازیابی اسناد

در این بخش دو روش جستجو باید به صورتی که در ادامه توضیح داده شده است، پیاده شوند.

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.

توجه: برای هر دو نوع جستجو نمایش ۲۰ سند در صورت موجود بودن کافی می باشد.

**بارم‌بندی**

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

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

نکات

1. 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.
2. امکان تغییر بارم‌بندی وجود دارد.
3. برای ارسال پروژه، کد خود را به صورت zip شده در سایت کوئرا بارگذاری نمایید.
4. می‌توانید سوالات و اشکالات خود را در سایت پیاتزا زیر پست مربوط به این تمرین بپرسید.
5. تمام قسمت‌های پروژه را خودتان به تنهایی پیاده‌سازی کنید. در صورت مشاهده تقلب، طبق قوانین دانشکده با شما برخورد خواهد شد.

**موفق باشید :)**