به نام او

پروژه‌ی درس ذخیره و بازیابی اطلاعات **ترم دوم سال تحصیلی۹۵-۹۴**

تاریخ تحویل: در پرتال اعلام خواهد شد. لطفا به فاصله زمانی های مناسبی پرتال دانشجویی خود را چک کنید.

گروه کامپیوتر دانشگاه کاشان

- هدف از این پروژه تسلط به مفاهیم و چگونگی کارکرد سیستم های بازیابی اطلاعات تحت وب است.
- برای پیاده‌سازی از هر زبان برنامه‌نویسی که مایلید استفاده کنید ولی ترجیحا از C++، C# یا MATLAB استفاده کنید. در صورت زمانگیربودن اجرای کدتان، گزارش تهیه کنید.
- این پروژه را به صورت **تک** یا **دو** نفره ارایه می‌کنید.
- این پروژه **پنج** نمره از بیست نمره‌ی شما را تشکیل می‌دهد.

دیتاست Cranfield یک دیتاست استاندارد در زمینه‌ی بازیابی اطلاعات است که حاوی ۱۴۰۰ سند (Document) می‌باشد که برای پروژه‌ی شما تنها ۱۰۰ سند اول آن منظور گردیده است. مطالب این دیتاست از علم aerodynamic آمده است! برنامه‌ای بنویسید که حاوی قسمت‌های زیر باشد و جواب تمام سوال‌های زیر را در خروجی اعلام نماید:

* حتما حتما؛ هر یک از کارهای زیر را به ترتیب روی دیتاست انجام دهید و خروجی را در فایلی جداگانه ذخیره کنید تا خروجی پس از هر مرحله در اختیار ما باشد! (این نکته بسیار مهم است و نمره‌ی بالایی دارد!!)

۱) تمام کاراکترهای بی‌مصرف آمده در فایل delimiters.txt و کلمه‌های بی‌مصرف آمده در فایل stopwords.txt را از سندهای دیتاست حذف کند.
۲) تعداد کلمات موجود در دامنه‌ی لغات (vocabulary) چند تا است؟
۳) ده کلمه ای که بیشترین تکرار را دارند چه کلماتی هستند؟

اگر فرض کنیم تمام کلمات موجود در دیتاست Cranfield (پس از ریشه‌یابی به وسیله‌ی الگوریتم Porter) به همراه تعداد تکرارشان و تعداد حضورشان در سندهای مختلف (دقیقا با همین ترتیب) در فایل words.txt آمده است؛ برای پاسخ به سوالات زیر برنامه بنویسید:

۴) بردار وزن را برای هر سند به روش TF-IDF ایجاد کنید.
۵) کاربر برای جست‌وجو در این اسناد درخواستی (query) را تایپ کرده است. بردار وزن را برای درخواست آمده در فایل query.txt به روش TF-IDF ایجاد کنید.
۶) شباهت سندها را با درخواست آمده به وسیله‌ی روش کسینوسی محاسبه و ۱۰ سند را به عنوان جواب به کاربر بازگردانید.

موفق باشید.

علی محمد نیک فرجام