پروژه درس مبانی بازیابی و جستجوی اطلاعات

نیمسال دوم ۹۶-۹۷، دانشگاه کاشان

این پروژه توسط تیم‌های دانشجویی با **حداکثر ظرفیت سه نفر** انجام می‌شود. هر تیم باید این پروژه را با Solr یا Elastic انجام دهد. هر دوی این زبانها در کلاس درس و توسط خود دانشجویان بررسی و معرفی شده است.

خروجی پروژه برای هر تیم شامل دو قسمت است: گزارش تکمیل شده و فایل برنامه. گزارش تکمیل شده، فایلی است که به تمامی سوالات پاسخ داده است. همچنین روند به دست آوردن پاسخ را با زبان انتخاب شده، توضیح داده و به تصویر می‌کشد. فایل(های) برنامه نوشته شده را نیز به طور کامل و با توضیحات مناسب هر قسمت به همراه گزارش ارسال می‌شود. درستی کار انجام شده، توضیحات واضح و شفاف، توضیف بهتر روند انجام کار مبنای نمره دهی خواهد بود.

زمان تحویل پروژه ۲۱ تیرماه ۱۳۹۷ است. فایل‌های گزارش را به صورت PDF آماده کرده و تمامی آنها را فشرده کرده و تنها یک فایل فشرده ایمیل نمایید. حتما اسم اعضای گروه در ابتدا گزارش باشد. برای ایمیل خود نام مناسبی قرار دهید.

برای انجام این تکلیف پیکره داده همشهری را از آدرس goo.gl/FexCiS دانلود نمایید.

سوالات: علاوه بر پاسخ هر سوال، چگونگی به دست آوردن جواب را در برنامه خود را توضیح دهید؛ برای هر مرحله تصاویر مناسب قرار دهید.

۱- این پیکره چند سند دارد؟

۲- یک مجموعه stop word برای خود تعریف کنید. این مجموعه کلمات چیست؟

۳- یک پیش پردازش، tokenization بر روی متن انجام دهید. (نیازی به stemming نیست). حذف Stop word مطلوب است. به گونه ای که "پردازش زبان، در جستجوی عبارات؟" به صورت زیر دربیاید:

پردازش	زبان	جستجوی	عبارات

تعداد کل کلمات منحصر بفرد را گزارش دهید.

۴- فرکانس سند (Document Frequency) برای تمام کلمات را پیدا و سپس IDF آنها را محاسبه کنید.

۵- تمام اسناد را به صورت TF-IDF گزارش دهید.

۶- فرض کنید کوئری "بازار بزرگ تهران" را داریم. تمام اسناد را بر اساس این کوئری بازیابی کنید و لیست مرتب شده‌ی ۲۰ سند اول را به همراه شماره سند DOCNO و امتیاز آنها (شباهت کسینوسی) گزارش کنید.

موفق باشید