باسمه تعالی

آزمون عملی پروژه درس بازیابی اطلاعات

نام و نام خانوادگی:

شماره دانشجویی:

سوال اول:

در این سوال می‌خواهیم کلیه پیش‌پردازش‌هایی که در پروژه شما بر روی یک متن ورودی انجام می‌شود را مرور کنیم. ممکن است شما یک پیش‌پردازش را در چند مرحله انجام دهید و یا چند مرحله از پیش‌پردازش‌های خواسته شده را در یک گام انجام دهید. متناسب با نوع پیاده‌سازی خود به شکل دقیق توضیح دهید که هر یک از پیش‌پردازش‌ها را به چه صورتی پیاده‌سازی کرده اید. برای هر مورد، توضیحات خودتان را به همراه تکه کد مربوط به آن بخش و همچنین خروجی آن بخش بر روی سندی که در ادامه مشخص شده را ارائه کنید. برای پاسخگویی به این سوال از سندی که شماره آن با سه رقم سمت راست شماره دانشجویی شما یکسان است استفاده کنید. مثلا اگر شماره دانشجویی شما ۴۰۳۳۱۰۶۸ است سند شماره ۶۸ (به ترتیبی که خبرها در فایل اسناد وارد شده اند) را پردازش کنید. شماره و متن سند را در زیر وارد کنید.

شماره سند:

متن سند:

استخراج توکن: بخش استخراج توکن مدل خود را در این بخش ارائه کنید.

توضیحات:

کد استخراج توکن:

توکن‌های استخراج شده برای سند مورد نظر (هر توکن به صورت جداگانه مشخص باشد):

نرمالسازی متن:

توضیحات:

کد نرمالسازی:

کلیه نرمالسازی‌های انجام شده برای سند مورد نظر (برای هر توکن به صورت جداگانه مشخص باشد):

حذف کلمات پر تکرار:

توضیحات:

کد حذف کلمات پرتکرار:

کلمات حذف شده برای سند مورد نظر:

ریشه یابی:

توضیحات:

کد ریشه یابی:

نتیجه ریشه‌یابی برای توکن‌های سند مورد نظر (معلوم باشد که هر کلمه ای به چه ریشه ای تبدیل شده است):

سوال دوم:

متن زیر را به عنوان یک خبر (سند) جدید به مجموعه اسناد پروژه اضافه کنید و بعد شاخص مکانی را برای همه اسناد تولید کنید و بازنمائی برداری این سند (یعنی وزن **tf-idf** کلیه کلمات موجود در متن داده شده) را گزارش نمایید. طبیعی است که وزن کلمات برای نسخه پیش‌پردازش شده سند محاسبه می شود که ممکن است بعضی کلمات در آن حذف شده باشند. برای نمایش بردار بازنمائی سند، هر کلمه از کلمات سند را به همراه وزن محاسبه شده برای آن کلمه گزارش کنید.

همچنین کلمات با بیشترین و کمترین وزن را مشخص و در مورد علت کم و زیاد بودن وزن آنها توضیح دهید.

دانشجویان دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر در مسابقات بین المللی برنامه سازی دانشجویی منطقه ای غرب آسیا به مقام سوم دست یافتند. به گزارش روابط عمومی دانشگاه صنعتی امیرکبیر: در بیست و چهارمین دوره از مسابقات بین‌المللی برنامه‌نویسی دانشجویی **ICPC** در منطقه‌ی غرب آسیا که با حضور ۸۰ تیم از دانشگاه‌های مختلف کشور در دانشگاه صنعتی شریف برگزار شد؛ تیم برنامه نویسی دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر موفق به کسب مقام سوم دانشگاهی و مقام دوم تیمی و مدال نقره در این دوره از مسابقات شد.

سوال سوم:

در شاخص مکانی ایجاد شده برای کل اسناد پروژه، لیست پست‌های مربوط به کلمه "دانشگاه" را بررسی و به سوالات زیر پاسخ دهید.

الف) شماره و عنوان خبرهایی که کلمه مذکور در آنها بیشترین و کمترین وزن را دارد را گزارش کنید.

ب) موقعیت مکانی کلمه مذکور برای ده سند اول موجود در لیست پست‌های این کلمه را گزارش کنید(شماره و عنوان سند و مکان‌های تکرار این کلمه را ذکر کنید). در صورت تکرار کلمه در یک سند، کلیه موقعیت های کلمه در آن سند ذکر شود. درستی موقعیت‌های محاسبه شده برای یک سند از سندهای مذکور را با گزارش متن سند بررسی کنید.

ج) عناصر لیست قهرمانان محاسبه شده برای این کلمه را گزارش کنید (شماره سند و وزن کلمه در این سندها را برای ۲۰ سند اول لیست قهرمانان ذکر کنید).

سوال چهارم:

در این سوال میخواهیم تاثیر **idf** در نتایج بازیابی شده را بررسی کنیم. برای این منظور شاخص جدیدی ایجاد کنید که در آن وزن کلمات در اسناد به جای رابطه **tf-idf** داده شده در تعریف پروژه صرفا بر اساس بخش **tf** محاسبه شود. با استفاده از چند عبارت جستجو و مقایسه نتایج حاصل از آنها در مدل بازیابی جدید و مدل بازیابی قبلی تاثیر این روش محاسبه بردار بازنمائی اسناد را بررسی کنید. کوئری های پیشنهادی خودتان و نتایج بازیابی شده در هر یک از دو مدل گزارش کنید و تحلیل خود را بیان نمایید. دقت داشته باشید که انتخاب عبارت جستجوی مناسب در این سوال مهم است.