«بسمه تعالی»

پروژه درس بازیابی اطلاعات

مقدمه

هدف از این پروژه، طراحی و پیاده‌سازی سیستم بازیابی اطلاعات برای مجموعه دادگان تعیین شده می‌باشد.

اولین مرحله، پیش‌پردازش مجموعه دادگان است. پس از آن نمایه‌ها با ویژگی‌های خواسته شده پیاده‌سازی می‌شود. در گام بعدی به ذخیره و بازخوانی نمایه‌ها به همراه روش‌های فشرده‌سازی پرداخته می‌شود. پس از آن تکنیک‌های اصلاح پرسمان پیاده‌سازی شده و در نهایت جستجو روی دادگان صورت می‌گیرد.

قسمت اجباری پروژه 4 نمره و قسمت اختیاری 4 نمره دارد.

توضیحات مربوط به هر بخش در ادامه آمده است که اهداف، محدودیت‌ها و خواسته‌های آن بخش را مشخص می‌کند.

مجموعه دادگان

مجموعه دادگان (Dataset) مورد بررسی در این پروژه از سایت kaggle فراهم شده است. این مجموعه شامل اطلاعات ٦٠٠٠ فیلم سینمایی از سال ١٩٠٤ تا ٢٠١٧ است. داده‌ها در قالب فایل csv دارای ستون‌هایplot ، title، id می‌باشد id .یک شناسه یکتا برای هر فیلم است که برای ارزیابی بهتر عملکرد شما به داده‌ها اضافه شده است و در مجموعه دادگان اصلی وجود نداشته است. همانطور که می‌دانید برای پیاده‌سازی نمایه باید به هر سند یک شناسه اختصاص بدهید. شناسه مربوط به هر سند باید id ذکر شده برای آن در مجموعه دادگان باشد. هر فیلم از دو بخش title و plot تشکیل شده است که از این دو بخش در ساخت نمایه و جستجو استفاده می‌شود. plot .خلاصه‌ای از طرح داستان فیلم است.

پیش‌پردازش و آماده‌سازی داده‌ها (اجباری 1 نمره)

در این بخش ابتدا داده‌ها را از فایل بخوانید. برای آماده‌سازی متن می‌توانید از کتاب‌خانه‌های آماده استفاده کنید. یکی از کتاب‌خانه‌های معروف برای این کار در زبان پایتون NLTK است اما در انتخاب روش پیاده‌سازی این بخش مختارید. این تابع یک متن انگلیسی ورودی گرفته و توکن‌های مربوط به آن‌را در قالب یک لیست خروجی می‌دهد. متن ورودی در عمل عنوان یا طرح داستان هر فیلم

است. دقت کنید که لیست خروجی شامل تعدادی توکن است که عملیاتcase folding ، stemmingو lemmatization روی آن‌ها اجرا شده است. در ضمن علائم نگارشی نباید به عنوان توکن در نظر گرفته شود. با توجه به نحوه پیاده‌سازی انواع بازگردانی به ریشه قابل قبول است.

شناسایی و حذف stop-words (اختیاری 1 نمره)

این بخش باید توسط خودتان و بدون استفاده از کد آماده پیاده‌سازی شود. termهای موجود در مجموعه دادگان را بر اساس تکرار آن‌ها مرتب کرده و پرتکرارترین آن‌ها را به عنوان stop-words در نظر بگیرید. اینکه چند term را به عنوان stop-words در نظر بگیرید به عهده خودتان است.

termهای به دست آمده از این بخش نباید در نمایه حضور داشته باشند.

نمایه‌سازی (اجباری 1 نمره)

در این بخش باید برای سامانه positional index بسازید. برای هر term باید مشخص باشد که آن term در عنوان چه فیلم‌هایی و در چه جایگاهی از عنوان هر فیلم قرار گرفته است. همچنین برای هر term باید مشخص باشد که آن term در طرح داستان چه فیلم‌هایی و در چه جایگاهی از طرح داستان هر فیلم قرار گرفته است.

پویاسازی نمایه (اجباری 1 نمره)

نمایه ایجاد شده باید قابلیت حذف و اضافه تک سند را داشته باشد. برای اضافه شدن سند، یک رشته داده می‌شود که اطلاعات مربوط به سند شامل id و plot و title در آن با کاما جدا شده است. برای حذف سند نیز id آن داده می‌شود.

تضمین می‌شود که شرط یکتا بودن id اسناد نقض نشود. برای مثال دو سند با شناسه یکسان به مجموعه اضافه نخواهد شد. البته ممکن است حذف شده و دوباره اضافه شود.

ذخیره و فشرده‌سازی نمایه (اختیاری 1 نمره)

در این بخش باید توانایی ذخیره کردن نمایه و بارگذاری مجدد آن را به سامانه اضافه کنید. ذخیره‌سازی به ۳ روش صورت می‌گیرد. بدون فشرده‌سازی، فشرده‌سازی از روش gamma-code و فشرده‌سازی از روش .variable-byte روش‌های فشرده‌سازی باید توسط خودتان پیاده‌سازی شود.

بخشی از نمره شما در این قسمت به میزان فشرده‌سازی نمایه اختصاص داده شده است. بنابراین پیاده‌سازی بهینه روش‌های فشرده‌سازی مهم است.

اصلاح پرسمان (اختیاری 1 نمره)

در صورتی که پرسمان ورودی دارای غلط املایی باشد یا به عبارتی لغاتی 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 با روش bigramو معیار jaccard نزدیک‌ترین لغات به لغت با غلط املایی را پیدا کنید. سپس با استفاده از معیار edit distance بهترین لغت را از میان آن‌ها بیابید.

نیازی به ذخیره‌سازی و فشرده‌سازی نمایه بایگرم نیست. همچنین می‌توانید از کد آماده برای محاسبه edit distance استفاده کنید.

جستجو و بازیابی اسناد (اجباری 1 نمره)

در این بخش جستجو ترتیب‌دار در فضای برداری tf-idf به روش lnn-ltn انجام می‌شود. یک پرسمان title و یک پرسمان plot ورودی گرفته شده و هر کدام در بخش مربوطه از اسناد جستجو می‌شوند. امتیاز نهایی هر سند برابر با جمع وزن‌دار امتیاز به دست آمده از جستجو در بخش title و plot است. به این صورت که وزن plot واحد در نظر گرفته شده و وزن عنوان به عنوان ورودی داده می‌شود.

در نهایت اسناد برتر را نمایش دهید. تعداد حداکثر اسناد برتر نیز به عنوان ورودی داده می‌شود.

نمایش اسناد (اختیاری 1 نمره)

نکته بسیار مهم نحوه نمایش اسناد انتخابی است. برای نمایش هر سند علاوه بر استفاده از شناسه و عنوان، یک هایلایت برای آن درست کنید. به این معنا که کلمات موجود در پرسمان را که باعث انتخاب سند شده‌اند به همراه ۳-۲ term قبل و بعد از آن به عنوان هایلایت آن سند نمایش دهید. اینگونه کاربر می‌تواند خیلی سریع دلیل بازیابی اسناد توسط سامانه را متوجه شود. مشابه کاری که سرچ گوگل انجام می‌دهد و ۳-۲ خط مربوطه را زیر وبسایت‌های پیشنهادی نمایش می‌دهد. طبیعتا راه حل بهینه برای این‌کار استفاده از قابلیت‌های نمایه جایگاهی می‌باشد.

نکات پایانی و مهم

سیستم را بهینه پیاده‌سازی کنید تا در زمان کمتری بارگذاری و نمایه‌سازی انجام شود.

پروژه را می توانید در گروه های دو نفره انجام دهید.

حداکثر تا 15 اردیبهشت فرصت دارید تا اعضای انتخابی گروه خود را برای بنده در ایتا بفرستید.

اگر دانشجویی قصد دارد پروژه را به تنهایی انجام دهد نیز اسم خود را برای بنده بفرستد.

هیچ محدودیتی در زبان برنامه نویسی انتخابی وجود ندارد.

مجموعه دادگان مربوطه در سامانه بارگذاری خواهد شد.

نحوه تحویل پروژه به صورت حضوری خواهد بود.

موفق و پیروز باشید.