فاز اول: استخراج اطلاعات ساختاریافته از یک سند

در فاز اول این پروژه، هدف ما استخراج اطلاعات از یک سند متنی به صورت ساختاریافته (Structured) و در قالب JSON است. این فرآیند به ما این امکان را خواهد داد تا اطلاعات غیرساختاریافته‌ی موجود در متن را به یک فرمت قابل‌استفاده و قابل‌تجزیه و تحلیل تبدیل کنیم.

اهداف فاز اول

- **استخراج اطلاعات:** شناسایی و استخراج معیارها، مشخصه‌ها و اطلاعات مرتبط از متن غیرساختاری.
- **ساختاردهی اطلاعات:** تبدیل اطلاعات استخراج شده به یک فرمت JSON ساختار یافته که شامل عناوین، معیارها و مشخصه‌های مربوطه باشد.
- **خودکارسازی فرآیند:** طراحی یک الگوریتم یا برنامه که به صورت خودکار قادر به شناسایی و استخراج اطلاعات از متن باشد، با توجه به این که عناوین و ساختار متن ممکن است (برای موضوعات مختلف) متفاوت باشند.

در این فاز به صورت خلاصه، مراحل زیر دنبال خواهند شد:

۱) فرمت فایل ورودی به صورت PDF می‌باشد.
۲) به طور خودکار بایستی متن ورودی با استفاده از یک کتابخانه از PDF استخراج‌شده و در نهایت براساس عناوین موجود در متن به JSON تبدیل شود.
۳) نیاز به یک الگوریتم هیورستیک برای یافتن سلسله مراتب در فرمت JSON داریم.
۴) یک ابزار یکپارچه برای تمامی مراحل فوق می‌بایست ایجاد شود.

فاز دوم: مدل‌سازی بازیابی اطلاعات از یک مجموعه اسناد

در این فاز شما قرار است که برروی یک‌سری سند که مجموعاً شامل تمامیِ اطلاعات یک سند اصلی هستند، مدل‌های بازیابی مختلف را پیاده‌سازی کرده و از query زدن استفاده کنید.

می‌توانید از تکنیک‌های مختلف برای پیاده‌سازی مدل‌های بازیابی اطلاعات استفاده کنید. این مدل‌ها شامل:

- **مدل بولین:** استفاده از جستجوی بولین برای بازیابی اسناد بر اساس عبارات کلیدی و شرایط منطقی.
- **مدل برداری:** پیاده‌سازی مدل‌های برداری مانند TF-IDF برای محاسبه شباهت بین اسناد و جستجوی اطلاعات.

***دقت کنید که مجاز به استفاده از کتابخانه‌های مختلف مانند Scikit-learn، NLTK، یا SpaCy هستید تا به بهینه‌سازی و تسریع فرآیند بازیابی کمک کنند.**

در این فاز به صورت خلاصه، مراحل زیر دنبال خواهند شد:

۱) شروع این فاز بعد از اتمام فاز اول خواهد بود.
۲) ورودی این فاز پس از تحویل فاز اول در اختیار دانشجویان قرار می‌گیرد.
۳) دانشجویان بایستی مدل‌های بولین، برداری و یک مدل هیورستیک را پیاده‌سازی و بردار مورد نظر را در RAM بسازند.
۴) تعدادی کوئری به همراه نتیجه‌ی معیار به دانشجویان داده خواهد شد که بایستی خروجی را را براساس precision و recall آماده کنند.
۵) همچنین در نهایت می‌بایست با بررسی و انجام تنظیمات مختلف روی مراحل اجرای کار، نتایج خروجی کوئری‌های داده‌شده را گزارش کرده و آن‌ها را بایکدیگر مقایسه کنید.

نمره اضافه:

۱) برای حداکثر ۲گروه که بر اساس کل اسناد، تعدادی کوئری تعریف کرده و ۲۰ جواب اول را به‌صورت مرتب آماده کنند (حداقل ۳۰ کوئری در زمینه‌های[1] مختلف).

[1] context

۲) برای گروه‌هایی که از ایده‌ای غیر از TF-IDF و پیاده‌سازی معمول مدل بولین و به طور کل ایده‌های خلاقانه استفاده کنند.

فاز سوم: ساخت تزاروس و گسترش کوئری[۲]

هدف ما ساخت یک تزاروس (Thesaurus) برروی اسناد داده‌شده فاز قبل و استفاده از آن برای گسترش کوئری‌هاست. این فرآیند به ما این امکان را می‌دهد که دقت و کارایی بازیابی اطلاعات را افزایش دهیم و تأثیر این گسترش را بر روی نتایج بازیابی بررسی کنیم.

اهداف:

- **ساخت تزاروس:** ایجاد یک تزاروس که شامل واژه‌ها و عبارات مرتبط با اطلاعات موجود در اسناد باشد.
- **گسترش کوئری:** استفاده از تزاروس برای گسترش کوئری‌های داده شده از فاز قبلی و بهبود نتایج بازیابی.
- **تحلیل تأثیر:** بررسی و تحلیل تأثیر مثبت و منفی گسترش کوئری بر روی نتایج بازیابی و ارائه گزارش‌های مربوطه.

در این فاز به صورت خلاصه، مراحل زیر دنبال خواهد شد:

۱) **تحلیل اسناد:** بررسی اسناد برای شناسایی واژه‌ها و عبارات کلیدی و ایجاد روابط معنایی بین آن‌ها.

۲) **ساخت تزاروس:** ایجاد یک ساختار تزاروس که شامل واژه‌های اصلی و مترادف‌ها، هم‌معناها و عبارات مرتبط باشد.

۳) **گسترش کوئری:** با استفاده از تزاروس، کوئری‌های ورودی را گسترش داده و واژه‌ها و عبارات جدیدی به آن‌ها اضافه کنید.

۴) **تحلیل نتایج:** بررسی نتایج بازیابی قبل و بعد از گسترش کوئری و ارزیابی تأثیر آن بر دقت و جامعیت نتایج.

*** این تحلیل می‌تواند شامل محاسبه معیارهای مختلف مانند Precision، Recall و F1-Score باشد.**

[2] Query expansion