بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱-۰۰

استاد: احسان‌الدین عسگری

تمرین سری سوم — **مهلت تحویل: ۱۳ خرداد**

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

در این تمرین هریک از گروه‌ها بر روی پروژه‌ی خود به صوت جداگانه کار خواهند کرد. به طور کلی در هر پروژه فرآیندی شامل دریافت متن، پیش‌پردازش اولیه‌ی آن و در نهایت ساخت یک سامانه‌ی جستجو انجام می‌شود. در این تمرین در بسیاری از بخشها می‌توانید از حاصل کار عزیزان ترم گذشته که با زحمات تدریسیاران درس در قالب کتابخانه parsi.io ایجاد شده بهره ببرید. به امید خدا در ترم‌های آینده حاصل تجمیع زحمات شما عزیزان در قالب محصولات متن‌باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می‌گیرد تا در اثر این تلاشها محصولاتی ارزشمند برای پردازش متن‌های فارسی و بلکه زبان‌های ایرانی و فراتر از آن داشته باشیم. می‌توانید به این کتابخانه از طریق این لینک دسترسی داشته باشید

نحوه‌ی ارزیابی در تمرین سری سوم

برای ارزیابی در این تمرین، از معیار MRR استفاده کنید. این معیار نیاز به داشتن تعداد کوئری و همچنین تعدادی سند است که مرتبط بودن آن سندها با کوئری مورد نظر نیز مشخص شده باشد. دادگان تمرین سری سوم هیچ‌کدام دارای داده‌ی برچسب‌خورده نیستند و این بر عهده‌ی هریک از گروه‌هاست که دستکم ده کوئری بسازند. در گام بعد به‌ازای هر کوئری ورودی، دستکم ۱۰ خروجی سامانه‌ی بازیابی خود را مشخص کرده و در نهایت هر یک از اعضای گروه باید مشخص کنند که کدام‌یک از سندهای بازگردانده شده مرتبط یا نامرتبط است. این کار را هر یک از اعضای گروه به صورت جداگانه باید انجام دهند و در نهایت از معیار MRR برای ارزیابی سامانه‌ی خودتان استفاده کنید. لازم به ذکر است که همه‌ی گروه‌ها باید از قالب یکسانی برای تهیه‌ی داده‌ی ارزیابی استفاده کنند که قالب کلی در این لینک در دسترس قرار گرفته است. همچنین یادآوری می‌شود که معیار MRR در اولین جلسه‌ی آینده‌ی کلاس تدریس می‌گردد ولی اگر گروهی مایل است که زودتر با این معیار آشنا بشود می‌تواند برای نمونه این لینک را مطالعه کند.

همچنین روش‌های مختلفی که برای سامانه‌ی بازیابی اطلاعات در تمرین‌ها نوشته شده‌اند باید بر مبنای MRR و همچنین شهود خود اعضای گروه با یکدیگر مقایسه‌ی تحلیلی شود. فراموش نگردد که برای همه‌ی پروپوزال‌ها باید این معیار ارزیابی محاسبه شود. بسته به هر تمرین ممکن است ارزیابی دیگری نیز از شما خواسته شود.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. یک منبع مطمئن و جامع برای این کار وبگاه گنجور است. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انجام بگیرد. این پیش‌پردازش‌ها باید با ماهیت متن‌های مورد نظر که ساختار شعری دارند تطابق داشته باشد. در گام پایانی نیز باید قادر باشید تا با دریافت یک بیت یا شعر، بیت یا شعرهایی مرتبط با آن برگردانید

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. برای نمونه تاکنون بررسی صورت نگرفته که حذف ایست‌واژه‌ها در متن‌های شعری چه تاثیر بر تحلیل‌ها دارد. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد شعر

با دریافت یک شعر از کاربر، باید بتوان تعدادی شعر دیگر به وی پیشنهاد کرد که واضح است شعرهای پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از .tf-idf
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل شعر برای بازنمایی بردار تعبیه‌ی کل آن شعر استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی شعر ورودی با همه‌ی شعرهای موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه و پس ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. می‌توانید از مجموعه دادگانی که در cw در اختیارتان قرار می‌گیرد استفاده کنید و یا از یک منبع مطمئن و جامع مانند وبگاه خبرگزاری همشهری استفاده کرده و دیتای مورد نظر خود را کراول کنید. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انجام بگیرد. این پیش‌پردازش‌ها باید با ماهیت متن‌های مورد نظر که قالب مقالات خبری را دارند تطابق داشته باشد. در گام پایانی نیز باید قادر باشید تا با دریافت یک کوئری، اخبار مرتبط با آن برگردانید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم به ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها مانند گسترش لیست کلمات کلیدی یا حذف کلمات با فرکانس بالا یا پایین، ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. دریافت اخبار مرتبط با کوئری کاربر

با دریافت یک کوئری شامل عنوان خبر یا دسته‌ی آن (ورزشی، سیاسی و ...) از کاربر، باید بتوان تعدادی مقاله‌ی خبری به وی پیشنهاد کرد که با آن کوئری ارتباط داشته باشد.. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از .tf-idf
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در متن یک مقاله‌ی خبری برای بازنمایی بردار تعبیه‌ی کل آن متن استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی کوئری ورودی با همه‌ی مقالات خبری موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه و پس ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. یک منبع مطمئن و جامع برای این کار وبگاه قائمیه است. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انجام بگیرد. این پیش‌پردازش‌ها باید با ماهیت متن‌های مورد نظر که ساختار شعری دارند تطابق داشته باشد. در گام پایانی نیز باید قادر باشید تا با دریافت یک بیت یا شعر، بیت یا شعرهایی مرتبط با آن برگردانید

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌بارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

در این مرحله لازم است تا متن به طور کلی تمیز (انجام امور پیش پردازشی) و طبقه بندی (بر اساس داستان ها و عنوان های فصول) شوند. بسته به ایده خودتان، نحوه ای برای ذخیره سازی فصول و بیت ها اتخاذ کنید.

برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. برای نمونه تاکنون بررسی صورت نگرفته که حذف ایست‌واژه‌ها در متن‌های شعری چه تاثیر بر تحلیل‌ها دارد. این مرحله به صورت آزمایش و خطا انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد شعر و رتبه بندی

با دریافت یک کوئری از کاربر، باید بتوان تعدادی از ابیات دیگر(برای از بین نرفتن انسجام پاسخ ها میتوان به جای یک بیت، یک یا دو بیت قبلی و بعدی آن را نیز به عنوان یک خروجی به حساب آورد) به وی پیشنهاد کرد که واضح است ابیات پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از .tf-idf
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار(برحسب وزن های tf-idf) بردارهای تعبیه، برای نمونه fasttext

با توجه به نتایج به دست آمده از ۴ روش بالا ، ذکر کنید که به نظر کدام روش برای مسئله ما بهتر عمل می کند.
در حالتی که از روش word embedding یا همان بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین

بردارهای تعبیه‌ی واژه‌های موجود در کل شعر برای بازنمایی بردار تعبیه‌ی کل آن شعر استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی بیت ورودی با همه‌ی بیت شعرهای موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را با ترتیب رتبه نزدیکی گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز k بیت بیشتر مرتبط را با استفاده از روشی که خودتان نیاز است تصمیم بگیرید ، رتبه بندی کرده و نمایش دهید. برای دستکم ده داستان شاهنامه ورودی و خروجی‌های خود را نشان دهید.
همچنین نیاز است تا این پیشنهادات شعری را برای حالتی که ایست واژه ها حذف شده و یا نشده اند، به دست آورید و بررسی کنید که کدام یک بهتر خروجی می دهد.

پیش‌گفتار

در این تمرین قصد داریم تا یک سامانه بازیابی اطلاعات ساده بر اساس مقالات عملی یک حوزه خاص را پیاده‌سازی نماییم. در ابتدا لازم است که اطلاعات تعدادی مقاله جمع‌آوری گردد. سپس در گام دوم لازم است که تعداد پیش‌پردازش بر روی داده‌ها انجام شود تا داده برای مراحل بعدی آماده گردد. در گام پایانی نیز باید بر اساس الگوریتم‌های گفته شده، یک سیستم جستجو ساده بر روی داده‌ها پیاده‌سازی شود.

۱. دریافت داده‌ها

در ابتدای کار لازم است که اطلاعات تعدادی مقاله با استفاده از روش‌های crwal گردآوری شود. این اطلاعات موارد زیر می‌باشند:

- id مقاله
- عنوان
- چکیده
- نام نویسندگان

توجه داشته باشید که لازم است اطلاعات حداقل ۱۰۰۰ مقاله در یک حوزه خاص (برای مثال مقاله‌های مربوط به مدل‌های transformer و کاربردهای آن در NLP) جمع‌آوری شود. توجه شود که در صورت جمع آوری داده بیشتر، سیستم پیاده‌سازی شده، عملکرد بهتری خواهد داشت. برای جمع‌آوری اطلاعات مقالات می‌توانید از سایت Semantic Scholar استفاده کنید.

۲. پیش‌پردازش اولیه متن

در این مرحله باید متون گردآوری شده را برای مراحل بعد آماده‌سازی و تمیز نمایید. در این بخش با توجه به عملکرد هر پیش‌پردازش و تاثیر آن بر خروجی نهایی، می‌توانید روش‌های مدنظر خود را اعمال کنید.
البته توجه داشته باشید که برای بخش عنوان و نام نویسندگان بهتر است که از برخی از پیش‌پردازش‌ها (مانند تقلیل فرم) استفاده نشود!

۳. دریافت مقاله‌های مرتبط با کوئری

در این بخش (که بخش اصلی تمرین می‌باشد)، یک سامانه جستجوی ساده طراحی می‌گردد. عملیات جستجو بر اساس بخش‌های مختلف مقاله (عنوان، چکیده، و نویسنده) می‌باشد. برای جستجو در ابتدا بخش مدنظر انتخاب می‌گردد و سپس کوئری وارد می‌شود. سپس عنوان k مقاله برتر (از نظر شباهت به کوئری) به کاربر برگردانده می‌شود.
برای پیاده‌سازی این بخش باید از ۴ الگوریتم زیر استفاده شود:

- روش بازیابی boolean برای بخش‌های نویسنده و عنوان

- روش بازیابی مبتنی بر tf-idf برای بخش‌های عنوان و چکیده
- استفاده از یک روش مبتنی بر transformer برای بخش چکیده
- استفاده از میانگین وزن‌دار (بر حسب tf-idf) بردارهای تعبیه[1] برای بخش چکیده

در مورد روش سوم، صرفا با استفاده از یک مدل transformer آماده[2]، و استفاده از آن به عنوان یک feature extractor، شباهت میان بردار کوئری و بردار خروجی چکیده مقاله باید بررسی شود و شبیه‌ترین موارد باید به عنوان خروجی سیستم ارائه گردند.
در مورد روش boolean مفهوم شباهت میان کوئری و مطرح نمی‌شود و لازم نیست که حتما k مورد به عنوان خروجی ارائه شود.

[1] word embedding
[2] pretrained

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انجام بگیرد. این پیش‌پردازش‌ها باید با ماهیت متن‌های مورد نظر که نوشتارهای سلامت و مقالات پزشکی را شامل می‌شوند تطابق داشته باشد. در گام پایانی نیز باید قادر باشید تا با دریافت یک موضوع یا بیماری، نوشته‌های مرتبط با آن را برگردانید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به از منابعی مانند بخش سلامت و درمان سایت نمناک ، اخبار مجله پزشکی دکتر سلام ، نمونه دادگان این مخزن یا هر منبع و روش مطلوب خودتان دریافت کنید. لازم به ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد مطلب

با دریافت یک موضوع یا عنوان از کاربر، باید بتوان تعدادی نوشته‌ی مرتبط به کاربر پیشنهاد کرد. واضح است که مطالب پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از چهار روش مختلف باید استفاده کنید:

(آ) استفاده از روش بازیابی boolean

(ب) استفاده از tf-idf

(ج) استفاده از یک مدل بر پایه‌ی Transformer ها

(د) استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل متن برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی مقالات موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انچام بگیرد. در گام پایانی نیز باید قادر باشید تا با دریافت یک موضوع، فایل‌های Readme و لینک‌های گیتهاب مرتبط با آن را برگردانید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم به ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند حذف Readme های تمپلیت بی‌محتوا، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد مطلب

با دریافت یک موضوع یا عنوان از کاربر، باید بتوان تعدادی نوشته‌ی مرتبط به کاربر پیشنهاد کرد. واضح است که مطالب پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از ۴ روش مختلف باید استفاده کنید:

(آ) استفاده از روش بازیابی boolean

(ب) استفاده از tf-idf

(ج) استفاده از یک مدل بر پایه‌ی Transformer ها

(د) استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل متن برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی فایل‌های Readme موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. سپس پیش‌پردازش روی داده‌ها انجام گیرد و در نهایت یک موتور جستجوی ساده ارائه دهید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم به ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها مانند گسترش لیست کلمات کلیدی یا حذف کلمات با فرکانس بالا یا پایین، ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. دریافت صفحات مرتبط با کوئری

در این بخش شما باید یک موتور جستجوی ساده طراحی کنید که با استفاده از کوئری کاربر، صفحات مرتبط را برگرداند. سپس تعدادی از نتایج را با رنک‌بندی مناسب به کاربر برگردانید. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از .tf-idf
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

اگر از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل متن برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصبه‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی صفحات کرال شده را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

در این تمرین، یک سامانه بازیابی اطلاعات برای شبکه اجتماعی‌ای که در پروپوزال پروژه انتخاب کرده‌اید، طراحی خواهید کرد. در گام اول داده‌ها را با استفاده از روش دلخواه خود جمع‌آوری می‌کنید. در گام دوم پیش‌پردازش‌های مورد نیاز را در صورت لزوم بر روی داده جمع‌آوری شده اعمال می‌کنید، به نحوی که با متون جمع‌آوری شده از شبکه اجتماعی مطابقت داشته باشند. در نهایت نیز باید بتوانید متون مرتبط به یک کوئری ورودی را در سامانه خود بازیابی کنید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترس برای تدریسیارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

در این گام متون جمع‌آوری شده را پیش‌پردازش می‌کنید. برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. در نتیجه این گام در ترکیب با گام بعدی و به صورت آزمایش و خطا و مشاهده نتایج انجام می‌شود. پیش‌پردازش‌ها می‌تواند شامل مواردی مانند حذف ایست‌واژه‌ها، نشانه‌های نگارشی، رعایت نیم‌فاصله‌ها، بازگردانی به ریشه و ... باشد.

۳. پیشنهاد متن مرتبط با کوئری

با دریافت یک کوئری از کاربر، باید بتوان متون مرتبط (مثلا در توییتر، توییت‌های مرتبط) به آن کوئری را به کاربر نمایش داد، به‌گونه‌ای که متون بازیابی شده ارتباط تنگاتنگی با کوئری داده شده داشته باشند. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از tf-idf.
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در یک متن موجود در شبکه اجتماعی برای بازنمایی بردار تعبیه‌ی کل آن متن استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی کوئری ورودی با همه‌ی متون موجود در داده جمع‌آوری شده خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه، k تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارشتان حتما نشان دهید.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها، یعنی ریاضی‌دانان و اطلاعات آنها به هر روش دل‌خواه دریافت گردند. روش پیشنهادی، crawl کردن وبسایت‌های ذکر شده در پروپوزال این موضوع پروژه است. در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن بیوگرافی‌ها انجام بگیرد. این پیش‌پردازش‌ها باید به گونه‌ای باشند که با ماهیت این متون که در مورد ریاضیات است متناسب باشند. در گام پایانی نیز باید قادر باشید تا با دریافت نام یک ریاضی‌دان، بیوگرافی و نام شاگردانش را نمایش دهید. هم‌چنین با دریافت یک موضوع، مثلاً هندسه اقلیدسی، از کاربر، لیستی از ریاضی‌دانان آن حوزه نمایش دهید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد مطلب

با دریافت یک موضوع یا عنوان از کاربر، باید بتوان تعدادی نوشته‌ی مرتبط به کاربر پیشنهاد کرد. واضح است که مطالب پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از ۴ روش مختلف باید استفاده کنید:

(آ) استفاده از روش بازیابی boolean

(ب) استفاده از tf-idf

(ج) استفاده از یک مدل بر پایه‌ی Transformer ها

(د) استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل متن برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی داده‌های موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

۴. موتورهای جستجو

نهایتاً باید بتوانید دو موتور جستجوی مستقل از هم به عنوان خروجی این تمرین ارائه دهید. یکی از این موتور جستجوها با گرفتن ورودی نام یک ریاضی‌دان، فهرست شاگردان او را خروجی می‌دهد. ورودی موتور جستجوی دیگر موضوعات مختلف حوزه ریاضی است؛ به عنوان مثال هندسه، جبر، ترکیبیات و ... . خروجی این موتور جستجو ریاضی‌دانانی هستند که در این حوزه کار کرده‌اند، که با توجه به اطلاعات موجود در صفحه‌ی هر ریاضی‌دان می‌توانید آن را استخراج نمایید.

پیش‌گفتار

هدف از این تمرین، پیاده‌سازی یک سامانه بازیابی اطلاعات برای داده‌های قرآنی است. در گام اول باید کتاب قرآن مجید دریافت شده و محتوای آن به فرمت قابل استفاده درآید. سپس پیش پردازش‌های لازم روی داده انجام شود تا در نهایت بتوانید آیه‌های مرتبط به یک کوئری ورودی را در سامانه خود بازیابی کنید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترس برای تدریسیارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

در این گام متون جمع‌آوری شده را برای مرحله‌ی بعد آماده و تمیز می‌کنید. برخی از پیش‌پردازش‌ها ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند پس باید بتوانید به یک تعادل در این زمینه برسید. در نتیجه این گام در ترکیب با گام بعدی و به صورت آزمایش و خطا و مشاهده نتایج انجام می‌شود. پیش‌پردازش‌ها می‌تواند شامل مواردی مانند ساخت توکن‌ها، حذف ایست‌واژه‌ها، حذف اعراب، بازگردانی به ریشه و ... باشد.

۳. پیشنهاد متن مرتبط با کوئری

با دریافت یک کوئری از کاربر، باید بتوان آیه‌های مرتبط به آن کوئری را به کاربر نمایش داد، به‌گونه‌ای که متون بازیابی شده ارتباط تنگاتنگی با کوئری داده شده داشته باشند. برای این کار باید از ۴ روش مختلف استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از tf-idf
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

اگر از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل متن برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی آیات موجود در پایگاه داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه ۱۰ مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.

پیش‌گفتار

در این بخش لازم است که در ابتدا داده‌ها به هر روش دل‌خواه دریافت گردند. یک منبع مطمئن و جامع برای این کار وبگاه ویکی‌پدیا است. دقت کنید که از دیتاست معرفی‌شده در پروپوزال مربوط به این قسمت در داک پروژه هم می‌توانید استفاده کنید و در انتخاب دادگان مسئله کاملا دست شما باز است.
در گام دوم در صورت لزوم بایستی پیش‌پردازش‌های لازم بر روی متن‌ها انجام بگیرد. این پیش‌پردازش‌ها باید با ماهیت متن‌های مورد نظر که قالب دستورات غذایی را دارند تطابق داشته‌باشد. در گام پایانی نیز باید قادر باشید تا با دریافت یک کوئری، دستورات غذایی مربوط به آن را برگردانید.

۱. دریافت داده‌ها

در این مرحله نیاز دارید تا داده‌ها را به هر روش مطلوب خودتان دریافت کنید. لازم ذکر است که داده‌های نهایی شما نیز مانند کدها باید قابل دسترسی برای تدریس‌یارها باشد.

۲. پیش‌پردازش اولیه‌ی متن

برخی از پیش‌پردازش‌ها مانند گسترش لیست کلمات کلیدی یا حذف کلمات با فرکانس بالا یا پایین، ممکن است بر روی خروجی سامانه‌ی شما تاثیر مستقیم مثبت و یا حتی منفی داشته باشند. باید بتوانید به یک تعادل در این زمینه برسید. این مرحله به صورت آزمایش خطا و انجام می‌شود و تاثیر این مرحله بر روی خروجی سامانه در مرحله‌ی بعد مشخص می‌شود. پیش‌پردازش‌ها شامل مواردی مانند رعایت نیم‌فاصله‌ها، برگرداندن به ریشه، حذف ایست‌واژه‌ها و نشانه‌های نگارشی و ... می‌تواند باشد.

۳. پیشنهاد دستور غذایی مرتبط با کوئری

با دریافت یک کوئری از کاربر باید بتوان دستوریات غذایی مرتبط به آن کوئری را به وی پیشنهاد کرد که واضح است دستورات غذایی بازیابی‌شده باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد. برای این کار از ۴ روش مختلف باید استفاده کنید.

۱. استفاده از روش boolean
۲. استفاده از tf-idf.
۳. استفاده از یک مدل بر پایه‌ی Trasnformer ها
۴. استفاده از میانگین وزن‌دار بردارهای تعبیه، برای نمونه fasttext

در حالتی که از روش بردارهای تعبیه استفاده می‌کنید، می‌توانید از میانگین بردارهای تعبیه‌ی واژه‌های موجود در کل دستور غذایی برای بازنمایی بردار تعبیه‌ی کل آن نوشته استفاده کنید. در گام پایانی فاصله‌ی کسینوسی میان بردار تعبیه‌ی نوشته‌ی ورودی با همه‌ی دستورات غذایی موجود در پایگاه‌داده‌ی خود را بسنجید و k نزدیک‌ترین خروجی را گزارش کنید. عدد k برابر با ۱۰ است اما این پارامتر باید قابل تغییر باشد. در حالتی که از روش‌های

دیگر مانند boolean و یا tf-idf استفاده می‌کنید نیز به طریق مشابه و پس ده تا از مرتبط‌ترین خروجی‌ها را گزارش کنید. ورودی و خروجی‌های خود را در گزارش‌تان حتما نشان دهید.