بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱-۰۰

استاد: احسان‌الدین عسگری

مهلت: ۱۴۰۱/۲/۳۱ پروپوزال

بوستان و گلستان سعدی

پیش‌گفتار

هدف از طراحی این بخش آن است که بتوان با استفاده از داده‌های متنی دو کتاب بزرگ سعدی به نام گلستان و بوستان، تحلیل‌های مختلفی به‌کار ببریم. در آغاز برای دریافت داده‌ها می‌توانید از هر روشی که می‌توانید متن این دو کتاب را تهیه کنید. یک راه مناسب برای این کار می‌تواند نوشتن یک خزنده‌ی ساده باشد که محتوای سایت گنجور ، بخش بوستان و گلستان سعدی را دریافت کند.
تحلیل‌های مختلفی که می‌تواند بر روی این مجموعه‌ی داده‌ها انجام بگیرد. برای نمونه:

۱. تحلیل لینک یا نود

بررسی این که کدام‌یک از مفاهیم (و یا شخصیت‌های) به کار رفته در بوستان و یا گلستان سعدی محوری است؟ به عبارتی مهم‌ترین دغدغه‌ی سعدی در حین نگارش این دو کتاب چه مفهومی بوده است؟

۲. پیشنهاد تک‌بیت‌های شعری

با دریافت یک یا چند بیت شعر از کاربر، باید بتوان تعدادی بیت شعر به وی پیشنهاد کرد که واضح است بیت‌های شعر پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد.

۳. پیشنهاد شعر

با دریافت یک یا چند شعر از کاربر، باید بتوان تعدادی شعر مختلف به کاربر پیشنهاد کرد که واضح است شعرهای پیشنهادی باید ارتباط تنگاتنگی با ورودی کاربر داشته باشد.

۴. داشتن یک موتور جستجوی شعری!

فرض کنید کاربری مایل است شعری از سعدی شیرازی بیابد که آن شعر پیرامون موضوعی مانند ریا و دروغ‌گویی باشد. ذر این مرحله باید شما قادر باشید با گرفتن یک یا چند مفهوم ورودی کاربری، یک یا چند شعر به او

پیشنهاد کنید. برای نمونه اگر کاربر دنبال شعری با مفهوم ریا بود، منطقی است که سامانه‌ی شما شعر زیر را پیشنهاد دهد:

شنیدم که نابالغی روزه داشت — به صد محنت آورد روزی به چاشت
پدر دیده بوسید و مادر سرش — فشاندند بادام و زر بر سرش
چو بر وی گذر کرد یک نیمه روز — فتاد اندر او ز آتش معده سوز
به دل گفت اگر لقمه چندی خورم — چه داند پدر غیب یا مادرم؟
کلید در دوزخ است آن نماز — که در چشم مردم گزاری دراز

۵. بررسی ذهنیت سعدی!

در این بخش هدف آن است که بتوانید با استفاده از روش‌های مختلف مانند embedding word نظر سعدی پیرامون گزاره‌های مختلف را به دست آوریم. برای نمونه شما باید با دریافت یک گزاره یا کلمه (مانند «مادر») و یا هر مفهوم دیگری، نزدیک‌ترین کلمات یا مفهوم‌های به‌کار رفته در شعر سعدی به آن را بیابید. از این طریق باید بتوانید تحلیل‌هایی نیز ارائه دهید. برای نمونه با بررسی واژه‌ی «مادر» و بررسی واژه‌های نزدیک به این واژه در بوستان و گلستان سعدی نظر سعدی پیرامون مادر مشخص خواهد شد. اگر کلمات نزدیک به «مادر»، مواردی مانند «خوبی»، «فداکاری» و ... باشد می‌تواند فهمید که نزدیک‌ترین مفهوم به کلمه‌ی مادر از دیدگاه سعدی چه چیزهایی است. نمونه‌ی دیگر می‌تواند بررسی کلمه‌ی «زن» باشد و از این طریق شاید بتوان فهمید که آیا واقعا سعدی بیش‌تر در شعرهایش «زن‌ستیز» بوده و یا خیر!

مقدمه

با توجه به نیاز جامعه علمی کشور به یک سامانه جستجو اختصاصی مقالات، در این پروژه قصد داریم تا یک سامانه جامع بازیابی اطلاعات برای مقالات حوزه‌های علمی مختلف را پیاده‌سازی کنیم. بخش‌های اصلی پروژه شامل بخش جستجو و رتبه‌بندی، نمایه‌سازی [1] ، آماده‌سازیCorpus ، و طراحی رابط کاربری می‌باشد.
همچنین امکاناتی مانند طبقه‌بندی و خوشه‌بندی مقالات بر اساس حوزه‌های علمی مختلف، ارزش‌گذاری مقالات و رتبه‌بندی نویسندگان در حوزه‌های مختلف، نیز در این سامانه پیاده‌سازی می‌گردند.
در ادامه به توضیح بخش‌های مختلف پروژه می‌پردازیم.

پیش‌نیاز آماده‌سازی Corpus

در ابتدا باید تعداد قابل قبولی (حداقل ۱۰۰۰۰) مقاله crawl شود و بخش‌های

- id مقاله
- عنوان مقاله
- چکیده
- سال انتشار
- نویسندگان مقاله
- موضوعات
- تعداد استنادهای[2] مقاله
- تعداد ارجاعات[3] مقاله
- عنوان ارجاعات مقاله (تنها ۱۰ مورد اول)

مقاله در corpus به طور مناسب ذخیره گردد. سایت مرجع پیشنهادی برای جمع‌آوری دادگان، سایت Semantic Scholar[4] می‌باشد.

پردازش متن

در این بخش، یک سری پردازش‌های مختلف برای یکسان سازی متن و query انجام می گردد. تعدادی از این موارد به صورت زیر می‌باشند:

- نرمال‌سازی (normalization)
- جداسازی (tokenization)

[1] Indexing
[2] citation
[3] reference
[4] www.semanticscholar.org

- یافتن و حذف stopwords
- Lemmatization یا Stemming
- حذف علائم نگارشی
- تصحیح پرسمان

مواردی مانند امکان استفاده از جستجوی wildcard و جستجوی phrase در این بخش پیاده سازی می‌گردند.
توجه داشته باشید که جستجو بر اساس فیلدهای مختلف(نام، عنوان، ...) انجام می‌شود. در نتیجه ممکن است که برخی از پیش‌پردازش‌ها برای تعدادی از فیلدها مناسب نباشد و هر فیلد باید عملیات پردازشی مخصوص به خود را داشته باشد. برای مثال، استفاده از lemmatization برای جستجو نام مناسب نیست.

جستجو و رتبه‌بندی

در ابتدا با استفاده از روش‌های مناسب، بخش‌های ذخیره‌شده مقاله (مانند چکیده) باید نمایه‌سازی شوند و به‌صورت بهینه در corpus ذخیره گردند. سپس با استفاده از الگوریتم مناسب و با داشتن query و نمایه‌های مربوط به corpus ، عملیات جستجو و رتبه‌بندی انجام می‌شود و مرتبط‌ترین مقالات به‌صورت نزولی به کاربر برگردانده می‌شود. به همراه هر مقاله نیز، نام سه نویسنده اول، متن نمونه[5]، سال انتشار، تعداد citation ارائه می‌شود.

رابط کاربری

برای راحت بودن استفاده از این سیستم، لازم است که یک رابط کاربری برای سیستم ایجاد شود.

خوشه‌بندی مقالات

در این بخش باید با استفاده از فیلدهای مفید مقاله‌های crawl شده، یک مدل خوشه‌بندی و دسته‌بندی پیاده‌سازی شود.
هدف

ارزش‌گذاری مقالات

با اجرای الگوریتم مناسب بر روی ارجاعات مقالات، مقالات ارزش‌گذاری می‌شوند. برای ارزش‌گذاری مقالات، صرفاً ارجاعات مقالات به یکدیگر بررسی می‌شوند و گراف ارجاعات ایجاد می‌شود و با استفاده از روش مناسب به ارزش‌گذاری مقالات پرداخته می‌شود.

رتبه‌بندی نویسندگان

برای رتبه‌بندی نویسندگان، مفهوم ارجاع نویسندگان به یکدیگر مطرح می‌شود. زمانی که نویسنده A در مقاله خود به مقاله P که نویسنده B جزو نویسندگان آن مقاله (P) می‌باشد، ارجاع دهد، می‌گوییم که نویسنده A به نویسنده B ارجاع داده است. با توجه به این رابطه، می‌توان گراف ارجاعات بین نویسندگان را ایجاد و سپس با استفاده از الگوریتم مناسب ، نویسندگان را رتبه‌بندی کرد.

[5] snippet

مقدمه

در این پروژه قصد داریم تا یک موتور جستجو برای مقالات خبری پیاده‌سازی کنیم. بخش‌های اصلی پروژه شامل بخش پردازش متن، نمایه‌سازی، Corpus، جستجو و رتبه‌بندی می‌باشد.
در ادامه به توضیح بخش‌های مختلف پروژه می‌پردازیم.

پیش‌نیاز (آماده‌سازی Corpus)

در ابتدا باید تعداد قابل قبولی (برای مثال ۵۰۰۰) مقاله‌ی خبری crawl شود و بخش‌های

- تاریخ انتشار خبر
- عنوان خبر
- خلاصه‌ی خبر
- برچسب‌های خبر
- موضوع

به صورت مناسب ذخیره گردد.

پردازش متن

در این مرحله از پروژه می‌بایست اقدامات لازم جهت پردازش متون اخبار را انجام دهید. این اقدامات شامل:

- واکشی خبر
- نرمال‌سازی (normalization)
- جداسازی (tokenization)
- حذف کلمات پرتکرار (stopwords)
- ریشه‌یابی کلمات (stemming یا lemmatization)
- حذف علائم نگارشی

توجه داشته باشید که جستجو بر اساس فیلدهای مختلف (عنوان خبر، برچسب‌ها، ...) انجام می‌شود. در نتیجه ممکن است که برخی از پیش‌پردازش‌ها برای تعدادی از فیلدها مناسب نباشد و هر فیلد باید علمیات پردازشی مخصوص به خود را داشته باشد.

نمایه‌سازی

با استفاده از روش مناسب، بخش‌های ذخیره شده اخبار (مانند خلاصه‌ی خبر) نمایه‌سازی شوند
همچنین فشرده‌سازی نمایه نیز باید اعمال شود.

جستجو و رتبه‌بندی

با استفاده از الگوریتم مناسب و با داشتن query و نمایه‌های مربوط به ،corpus عملیات جستجو و رتبه‌بندی انجام می‌شود و مرتبط‌ترین اخبار به‌صورت نزولی و به همراه خلاصه‌ی خبر به کاربر برگردانده شوند. همچنین می‌توانید درصد شباهت خبر بازگردانده شده به کوئری کاربر را نمایش دهید.

ساخت گراف ارتباط

در این بخش نیاز است تا با در نظر گرفتن لوکشین‌های مرتبط با یک خبر و دسته‌بندی آن (ورزشی، سیاسی و ...)، گراف ارتباط لوکشین خبر و نوع آن را پیدا کنید و الگوریتم پیج رنک روی این گراف پیاده سازی کنید.

پیش‌گفتار

با توجه به افزایش اهمیت جایگاه سلامت و اخبار پزشکی، نیاز به حضور یک سامانه‌ی توان‌مند برای یافتن منابع معتبر و مرتبط با کم‌ترین تاخیر وجود دارد. هدف از این پروژه ایجاد یک سیستم بازیابی اطلاعات در زمینه‌های مرتبط با سلامت، پزشکی و مطالعات پیرامون بیماری‌ها و به‌کار بردن تحلیل‌های مختلف خواهد بود.

در آغاز برای دریافت داده‌ها می‌توانید از هر روشی که می‌توانید متن این دو کتاب را تهیه کنید. جهت جمع‌آوری دادگان مرتبط می‌توانید از سایت ResearchGate با ذخیره‌ی چکیده‌های مقالات مرتبط، دانش سیستم در این حوزه را تقویت کنید. می‌توانید به‌صورت دستی یک لیست اولیه از مقالات مورد تایید خود تهیه کنید و با شروع از آن‌ها و در نظر گرفتن reference ها و citation ها و پیش رفتن تا چند مرحله مقالات زیادی پیدا کنید.
همچنین برای جمع‌آوری دادگان می‌توانید از بخش سلامت و درمان سایت نمناک و اخبار مجله پزشکی دکتر سلام استفاده کنید (نمونه دادگان crawl شده در این مخزن قابل مشاهده است).

پروژه می‌تواند شامل قسمت‌های مختلفی باشد؛ برای نمونه:

۱. پردازش متن

این بخش از پروژه شامل نرمال‌سازی متن، استخراج توکن‌ها و حذف لغات پرتکرار و علائم نگارشی، و stem - ming و lemmatization خواهد بود. دقت شود که روی هر field مختلف از مقاله (مانند عنوان، چکیده، و ...) پیش‌پردازش مناسب آن صورت گیرد تا در جست‌وجو عملکرد سیستم مناسب‌تر باشد.

۲. نمایه‌سازی (Indexing) و فشرده‌سازی نمایه

پس از پردازش متن و با کمک توکن‌های موجود، نمایه‌سازی دادگان باید صورت گیرد (برای مثال نمایه‌های BiGram و Positional). همچنین، لازم است تا امکان ذخیره‌ی نمایه به‌صورت فشرده و بارگذاری مجدد آن وجود داشته باشد (می‌توانید عملکرد روش‌هایی مانند variable code و gamma code را مقایسه و سپس گزینه‌ی مطلوب را انتخاب کنید).

۳. تصحیح پرسمان و پیشنهاد هنگام جست‌وجو

در سیستم بازیابی اطلاعات، لازم است تا در صورت مشاهده غلط املایی در پرسمان کاربر، با توجه به دانش لغات و عبارات سیستم، واژه‌ی صحیح جایگزین شود. همچنین می‌توان هنگام تایپ عبارت مورد جست‌وجو پیشنهاداتی (مرتب شده براساس مرتبط بودن) به‌منظور تکمیل پرسمان کاربر ارائه شود. همچنین می‌توانید با ایجاد یک زیرسیستم توصیه‌گر، مقالات و نوشتارهای مرتبط را به کاربر پیشنهاد دهید.

۴. طبقه‌بندی و خوشه‌بندی

برای طبقه‌بندی و خوشه‌بندی مقالات، روش‌های متعددی به‌کار گرفته می‌شوند. در این بخش می‌توانید با اعمال برخی از این روش‌ها می‌توان تحلیل‌ها و دسته‌بندی‌های مناسبی از دادگان موجود ارائه دهید.

۵. رتبه‌بندی مقالات و نویسندگان

در این بخش می‌توانید اعتبار مقالات و نویسندگان را (با الگوریتم‌های مناسب) براساس اعتبار آن‌ها و میزان ارجاع‌هایی که داده/گرفته‌اند محاسبه و رتبه‌بندی کنید.

پیش‌گفتار

در این پروژه قصد داریم با بررسی دیتای مناسب و استخراج داده‌ی لازم، فهرستی قابل جستجو از ریاضی‌دانان انجمن ریاضی آمریکا استخراج کنیم، بطوریکه رابطه‌ی بین ریاضی‌دانان مشخص باشد. این رابطه عمدتاً به صورت استادی‌ـ شاگردی‌ست. یعنی باید برای هر ریاضی‌دان، لیستی از اساتید و شاگردانش قابل مشاهده باشد؛ هم‌چنین امکان مشاهده سلسه‌مراتبی نیز باشد، یعنی شاگردان و شاگردانِ شاگردان و ... تا فاصله‌ای معین را بتوان به شکل درختی مشاهده کرد. رابطه‌ی بین ریاضی‌دانان می‌تواند در قالب‌های دیگری از جمله استاد پایان‌نامه و ... نیز باشد. هم‌چنین برای هر ریاضی‌دان باید امکان مشاهده بیوگرافی مختصر و جامعی باشد.

نهایتاً خروجی شما باید یک سیستم برای بازیابی اطلاعات با رابط کاربری مناسب بر اساس نام ریاضی‌دانان، موضوعات تحقیقاتی آنها، کشور و زمان زندگی‌هایشان باشد. مواردی مانند تصحیح اشتباهات املایی و پیشنهاد ورودی بر اساس حروف تایپ شده نیز در این پروژه مد نظر است.

پیشنهادها

در این شجره‌نامه می‌توانید به استخراج موضوعات گوناگونی بپردازید؛ برای نمونه:

۱. خلاصه بیوگرافی

برای نمایش نتیجه‌ی جستجو، شما می‌توانید از بیوگرافی هر ریاضی‌دان بخش‌های مهم‌تر (مانند تولد، علت معروفیت، تئوری‌ها و شاگردان برجسته، مرگ و ...) را استخراج کرده و نمایش دهید. بدیهی‌ست در این بخش نیاز به پیش‌پردازش متن خواهید داشت.

۲. کلیدی‌ترین ریاضی دانان

با بررسی تعداد شاگردان، مقالات، یا پایان‌نامه‌های هر استاد و یا به کمک تحلیل لینک، فهرستی مرتب از کلیدی‌ترین ریاضی‌دانان ارائه دهید.

۳. کلیدی‌ترین سرزمین‌ها

با بررسی زادگاه ریاضی‌دانان، فهرستی مرتب از کشورهایی که بیشترین تعداد ریاضی‌دان عضو انجمن ریاضی آمریکا را دارند فراهم کنید و تغییرات این کشورها را در گذر زمان گزارش دهید.

۴. محبوب‌ترین مراکز علمی

با بررسی تعداد شاگردان، مقالات، یا پایان‌نامه‌های هردر سده‌ها یا دهه‌های گوناگون فهرستی مرتب از محبوب‌ترین دانشگاه‌ها یا مراکز علمی، با توجه به تعداد ریاضی‌دانان هر یک، ارائه دهید.

۵. سامانه توصیه‌گر

طراحی یک Recommender System ساده، به طوریکه برای هر نتیجه‌ی جستجو شده فهرستی از ریاضی‌دانان که موضوع پایان‌نامه‌ی مشابهی دارند نمایش داده شود، یا برای موضوع مشخصی، فهرستی از ریاضی‌دانان مشغول به کار در آن زمینه فراهم گردد.

جمع‌آوری داده

رای استخراج ریاضی‌دانان انجمن ریاضی آمریکا و روابط بین آنها می‌توانید از اطلاعات موجود در این وبسایت استفاده کنید. برای مشاهده‌ی پروفایل و استخراج بیوگرافی هر راضی‌دان می‌توانید از صفحه‌ی ریاضی‌دانان در ویکی‌پدیا، و یا صفحه‌ی بیوگرافی هر ریاضی‌دان در وبسایت MacTutor استفاده کنید. مثلاً این صفحه‌ی بیوگرافی دیریکله است. برای دسترسی به این داده‌ها باید از crawl کردن کمک بگیرید. در این زمینه، استفاده از لینک۱، لینک۲ و لینک۳ می‌تواند به شما کمک کند.

پیش‌گفتار

در این پروژه قصد داریم تا علاوه بر بررسی تفصیلی شاهنامه، برای آن یک موتور جستجو پیاده سازی کنیم. شاهنامه اثر حکیم ابوالقاسم فردوسی توسی، حماسه‌ای منظوم، بر حسب دست نوشته‌های موجود دربرگیرنده نزدیک به ۵۰,۰۰۰ بیت تا نزدیک به ۶۱,۰۰۰ بیت و یکی از بزرگ‌ترین و برجسته‌ترین سروده‌های حماسی جهان است که سرایش آن دست‌آوردِ دست‌کم سی سال کارِ پیوستهٔ این سخن‌سرای نامدار ایرانی است. موضوع این شاهکار ادبی، افسانه‌ها و تاریخ ایران از آغاز تا حملهٔ عرب‌ها به ایران در سدهٔ هفتم میلادی است (شاهنامه از سه بخش اسطوره‌ای، پهلوانی و تاریخی تشکیل شده‌است) که در چهار دودمان پادشاهیِ پیشدادیان، کیانیان، اشکانیان و ساسانیان گنجانده می‌شود.

در این پروژه لازم است تا موارد زیر انجام شود:

۱. ۱ مجموعه داده

ابتدا لازم است تا متن شاهنامه فردوسی را دریافت کنید. برای این منظور می توانید به خواست خود از هر منبع معتبری استفاده کنید. برای سهولت بیشتر شما، می توانید از سایت قائمیه شاهنامه فردوسی را دریافت کنید.

۲. ۲ انجام مراحل پیش پردازشی

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.

۳. سامانه جست و جو و رتبه بندی

در این بخش لازم است به یکی از روش های tf-idf و یا embedding word از متن شاهنامه ابیات مرتبط با کوئری کاربر را استخراج و رتبه بندی کنید.

۴. سامانه توصیه گر

در این قسمت، باید با یک روش دلخواه خود، کوئری کاربر را اگر دارای خطای نوشتاری است، تصحیح نمایید. همچنین لازم است تا ابیات شعری مرتبط با بیت وارد شده را از شاهنامه و اشعار شاعران دیگر (شاعران پس از فردوسی که از داستان ها و متون شاهنامه استفاده و یا اشاره کرده اند) بیابید و توصیه نمایید.(بسته به متون و شاعران انتخاب شده ممکن است برای این بخش انجام مراحل پیش پردازشی برای متون باقی شعرا لازم باشد)

۵. خوشه بندی

در این بخش قصد داریم تا بفهمیم به طور کلی شاهنامه دارای چه موضوعات و مباحثی است. به این صورت که لازم است با در نظر گرفتن امبدینگ ابیات (یا روش های مشابه یا بهتر) ابیات را خوشه بندی کنیم. برای مثال ممکن است در انتهای این بخش به این برسیم که دسته های ابیات شاهنامه عبارتند از ابیات مربوط به جنگ،...و...

۶. ساخت گراف ارتباط

در این بخش نیاز است تا با در نظر گرفتن اشاره به اسامی مختلف در نزدیکی هم در شاهنامه، گراف ارتباط شخصیت های افراد موجود در شاهنامه را پیدا کنید و الگوریتم پیج رنک روی این گراف پیاده سازی کنید و به این ترتیب رتبه اهمیت افراد در شاهنامه را بیابید.

۷. ساخت رابط کاربری

در این بخش نیاز است با استفاده از Elastic Search و یا Milvus یک رابط کاربری برای استفاده راحت از تمامی موارد موجود در پروژه بسازید.

موتور جستجوی صفحات وب

مقدمه

اینترنت یکی از منابع اولیه اطلاعات برای میلیون ها نفر است که می‌توان اطلاعات مربوط به همه مسائل را در آن پیدا کرد. علاوه بر این، اگر بخواهیم اطلاعاتی در مورد یک موضوع خاص بازیابی کنیم، ممکن است هزاران صفحه وب مرتبط با آن موضوع پیدا کنیم. اما دغدغه اصلی ما یافتن صفحات وب مرتبط از میان آن مجموعه است. اینترنت در چند سال گذشته رشد تصاعدی داشته است. تقریباً ۱۵ تا ۲۰ میلیارد صفحه در وب وجود دارد و اخیراً این تعداد به مرز ۱ تریلیون رسیده. طبق مطالعات انجام شده ۲۰ تا ۱۵ درصد صفحات وب اکثراً تکراری از صفحات اصلی و برخی از آنها صفحات کاملاً بی‌ربط هستند. بنابراین، انفجار وب بسیاری از مشکلات جدید را برای سیستم های بازیابی اطلاعات ارائه می‌دهد. سیستم‌های بازیابی اطلاعات به کاربران کمک می‌کنند تا کارهای جستجو را با یافتن تعداد محدودی از اسناد مرتبط در میان هزاران صفحه متن با سازماندهی ساختاری کمی انجام دهند. در عین حال، توسعه دهندگان سیستم های بازیابی باید بتوانند اثربخشی کلی این سیستم‌ها را ارزیابی کنند.
هدف انجام این پروژه دستیابی به یک موتور جستجوی ساده است که بر اساس عوامل مختلف کاربر را به نتیجه‌ی مطلوب برساند. بدیهی است که انتظار ما از شما پیاده‌سازی یک موتور جستجوی کامل مثل گوگل نیست.

مراحل انجام تمرین

واکشی داده‌ها

برای تشکیل یک مجموعه داده، داده‌های متنی مورد نظر را در ابتدا واکشی کنید.

پیش‌پردازش

داده‌های جمع‌آوری شده را با اعمال جداسازی، نرمال‌سازی، حذف علائم نگارشی، ریشه‌یابی کلمات و حذف کلمات پرتکرار، پیش‌پردازش کنید. در ادامه روی توکن‌های استخراج شده، نمایه‌های مناسب بسازید.

تصحیح اشکالات املایی

اگر پرسمان ورودی توسط کاربر دارای غلط املایی است، باید در این بخش به اصلاح آن بپردازید.

بازیابی اطلاعات

سیستم شما باید با جستجوی پرسمان یا کوئری ورودی، نتایج مورد نظر را به ترتیب رنک به کاربر نمایش دهد.

پیشنهادها

برای سیستم جستجوی خود، می‌توانید موضوعات مختلفی را استخراج کنید. چند نمونه برای مثال ذکر شده؛

۱. سیستم پیشنهاددهنده

در این قسمت، با توجه به پرسمان ورودی کاربر، چند نمونه از پرسش‌های مشابه را که اشتراک کلمات بیشینه با پرسمان کاربر دارند، برگرداند.

۲. استخراج عناصر کلیدی

در این قسمت می‌توانید یک عنصر کلیدی برای هر صفحه انتخاب کنید، مثلا یک موضوع یا یک شخص. سپس فهرستی از مرتبط‌ترین نتایج را با رنک‌بندی مناسب خروجی دهید.

۳. خوشه‌بندی پرسش‌ها

با استفاده از پرسش‌های استخراج شده، خوشه‌های مربوط به پرسش‌های مختلف را با استفاده از الگوریتم‌های مناسب، خروجی دهید.

۴. گراف ارتباط

می‌توانید گراف ارتباط موجود بین عناصر کلیدی هر صفحه را بسازید و الگوریتم پیج‌رنک را پیاده‌سازی کنید.

۵. sentiment-analysis

برای داده‌های واکشی شده و پرسمان ورودی، این قابلیت وجود داشته باشد.

نجوم

مقدمه

علم نجوم یکی از قدیمی‌ترین دانش‌های بشری است. اخترشناسان در تمدن‌های اولیه‌ی بشری به‌دقت آسمان شب را بررسی می‌کردند و با استفاده از ابزارهای ساده‌ی اخترشناسی که از همان ابتدا شناخته‌شده بود، به مطالعه‌ی آسمان‌ها می‌پرداختند .
در این پروژه قصد داریم با بررسی داده‌های مناسب، یک سیستم بازیابی اطلاعات پیاده‌سازی کنیم که قابلیت جست‌وجو در متن و بازیابی مطالب مناسب را داشته باشد. داده‌های مورد استفاده در این پروژه داده‌های نجومی مربوط به کتاب التفهیم نوشته ابوریحان بیرونی است. این کتاب شامل پنج باب مختلف به همراه پانصدو سی پرسخ به همراه پاسخ آن است. این کتاب در این لینک قابل مشاهده است.

پیش‌پردازش‌های اولیه

در این بخش ابتدا داده‌ها را از لینک داده شده استخراج کنید. سپس موارد زیر را برای پیش پردازش داده‌های استخراج شده انجام دهید.

- ابعد از استخراج توکن‌ها عملیات lemmatization و stemming را روی آنها انجام دهید.
- در مرحله‌بعد باید توکن‌ها را بر اساس تعداد تکرارشان مرتب کرده و با تعیین یک treshold مناسب stop word ها را حذف کنید.
- چدر این بخش باید عملیات ساخت نمایه روی توکن‌های استخراج شده انجام شود.

سیستم اصلاح اشکالات املایی

در صورتی که پرسمان ورودی دارای غلط املایی باشد، لازم است یک جستجو بین لغات احتمالی انجام شده و بهترین لغت به عنوان پرسمان جایگزین پیشنهاد شود.

سیستم بازیابی اطلاعات

در این قسمت کاربر باید بتواند با وارد کردن یکی از پانصد و سی پرسش، پاسخ های مرتبط با آن پرسش را به ترتیب رتبه‌ای که با توجه به الگوریتم پیاده‌سازی شده دریافت می‌کند بازیابی کند.

سیستم پیشنهاد دهنده

در این قسمت باید با استفاده از پرسش انجام شده توسط یک کاربر بتوانیم چند نمونه از پرسش‌های مشابه را نیز پیدا کنیم. (پرسش مشابه پرسشی است که بیشترین اشتراک کلمات را داشته باشد.)

خوشه‌بندی پرسش‌ها

در این قسمت باید با استفاده از پرسش‌های استخراج شده از متن کتاب و استفاده از الگوریتم‌های تبدیل متن به وکتور، خوشه های مربوط به پرسش‌های مختلف، که هر خوشه مربوط به هر باب است را پیدا کنید.

تحلیل لینک

صورت فلکی مجموعه ای از ستاره‌ها است که با یکدیگر شکلی را در آسمان پدید می آورند. هر صورت فلکی از تعدادی ستارگان ثابت تشکیل یافته است، که چگونگی قرار گرفتن هر گروه از آنها نسبت به یکدیگر، به یک جسم یا حیوان شباهت دارد و نام آن جسم یا جانور را روی آن دسته گذاشته اند.
در این لینک اشکال مربوط به ۸۸ صور فلکی اصلی آمده است. این تصاویر شامل ستاره‌های موجود در هر صورت فلکی و ارتباط بین این ستاره‌ها است. هر ارتباط بین دو ستاره را به صورت یک لینک دو طرفه در نظر گرفته و با استفاده از الگوریتم‌های مربوط به تحلیل لینک، ستاره‌های پر اهمیت‌تر را مشخص کنید.

مقدمه

هدف از این پروژه، استخراج و بررسی داده‌هایی قابل جست‌وجو از یک شبکه اجتماعی مبتنی بر متن (نظیر twitter) است. یعنی در ابتدا باید یک شبکه اجتماعی در نظر گرفته شده و داده‌های متنی از آن استخراج شوند و سپس بر اساس روش‌هایی که در درس معرفی شده‌اند، قابلیت بازیابی متون بر اساس پرسمان ورودی فراهم شود.
با توجه به گسترده بودن ابعاد شبکه‌های اجتماعی و دلخواه بودن آن شبکه، برای استخراج داده‌ها از crawling استفاده می‌شود.

مراحل انجام پروژه

در این پروژه مراحل زیر باید انجام بشوند:

۱. استخراج داده

به این منظور باید از شبکه اجتماعی گفته شده، داده‌های متنی crawl شوند.

۲. پیش‌پردازش داده‌ها

در این مرحله داده‌های جمع‌آوری شده باید پیش‌پردازش شوند. اعمالی نظیر جداسازی، نرمال‌سازی، حذف علائم نگارشی،stemming و lemmatization و حذف کلمات پرتکرار در این مرحله انجام می‌شوند.

۳. نمایه‌سازی و فشرده‌سازی

در این بخش به انتخاب خود نمایه‌های مورد نیاز در پروژه نظیر نمایه positional، bigram و ... را پیاده‌سازی می‌کنید. همچنین این نمایه‌ها باید قابل ذخیره و بارگذاری باشند، در نتیجه باید فشرده‌سازی نمایه‌ها نیز مد نظر قرار بگیرد.

۴. تصحیح و تکمیل خودکار پرسمان ورودی

در صورتی که پرسمان ورودی دارای غلط املایی باشد، باید امکان تصحیح آن وجود داشته باشد. همچنین امکان تکمیل خودکار پرسمان نیز در نظر گرفته شود.

۵. جستجو و بازیابی پرسمان

در نهایت سیستم باید قابلیت جستجوی پرسمان ورودی را در داده‌های جمع‌آوری شده داشته باشد و نتایج باید به ترتیب رتبه به کاربر بازگردانده شوند.

پیشنهادها

در این پروژه موارد زیر نیز قابل بررسی هستند.

۱. طراحی سیستم توصیه‌گر

سیستم توصیه‌گر با ورودی گرفتن نام کاربری یک کاربر بتواند افرادی مشابه را به او معرفی کند. به این منظور می‌توان از تحلیل لینک (در شبکه اجتماعی twitter مواردی نظیر retweet، hashtag و ...) در کنار بررسی شباهت متون استفاده کرد.

۲. Sentiment Analysis

برای هر یک از داده‌های استخراج شده و همچنین پرسمان ورودی قابلیت Sentiment Analysis وجود داشته باشد.

بازیابی بر روی فایل‌های Readme سایت Github

مقدمه

گیتهاب یکی از بهترین سایت‌های موجود برای برنامه‌نویسان است که در آن می‌توانند کدهای خود را به صورت گروهی بزنند و تغییرات آن را در هر مرحله مشاهده کنند. هر پروژه برنامه‌نویسی توضیحاتی در قالب فایل‌های Readme دارد که برای مشخص کردن موضوع کلی پروژه بسیار کارآمد می‌باشد. هدف از این پروژه زدن یک موتور جست‌وجو است به این صورت که با وارد کردن موضوع تمامی لینک‌های گیتهاب موجود مرتبط با آن یافت شوند. همچنین رتبه‌بندی بین لینک‌ها برای ما حائز اهمیت است. این رتبه‌بندی می‌تواند علاوه بر میزان مرتبط بودن موضوع فایل وابسته به امتیازدهی کاربران به آن صفحه گیتهاب باشد.

مراحل انجام تمرین

واکشی داده‌ها

برای تشکیل یک مجموعه داده، داده‌های متنی مورد نظر را در ابتدا واکشی کنید.

پیش‌پردازش

داده‌های جمع‌آوری شده را با اعمال جداسازی، نرمال‌سازی، حذف علائم نگارشی، ریشه‌یابی کلمات و حذف کلمات پرتکرار، پیش‌پردازش کنید. در ادامه روی توکن‌های استخراج شده، نمایه‌های مناسب بسازید.

تصحیح اشکالات املایی

اگر پرسمان ورودی توسط کاربر دارای غلط املایی است، باید در این بخش به اصلاح آن بپردازید.

بازیابی اطلاعات

سیستم شما باید با جستجوی پرسمان یا کوئری ورودی، نتایج مورد نظر را به ترتیب رتبه‌ای که قبل‌تر توضیح داده شد به کاربر نمایش دهد.

پیشنهادها

برای سیستم جستجوی خود، می‌توانید موضوعات مختلفی را استخراج کنید. چند نمونه برای مثال ذکر شده؛

۱. طراحی سیستم توصیه‌گر

در این قسمت، با توجه به پرسمان ورودی کاربر، چند نمونه از پرسش‌های مشابه را که بیشترین شباهت با پرسمان کاربر دارند، برگرداند.

۲. استخراج کلیدواژه‌ها

در این قسمت می‌توانید یک کلیدواژه برای هر صفحه انتخاب کنید، سپس فهرستی از مرتبط‌ترین نتایج را با رتبه‌بندی مناسب خروجی دهید.

۳. خوشه‌بندی پرسش‌ها

با استفاده از پرسش‌های استخراج شده، خوشه‌های مربوط به پرسش‌های مختلف را با استفاده از الگوریتم‌های مناسب، خروجی دهید.

۴. گراف ارتباط

می‌توانید گراف ارتباط موجود بین کلیدواژه هر صفحه را بسازید و بر اساس آن تحلیلی بر روی داده‌ها انجام دهید.

مقدمه

همانطور که می‌دانید تهیه‌ی غذا از روش‌های مختلف و با دستورات مختلفی امکان‌پذیر است؛ بخصوص در مورد آشپزهای مختلف، نحوه‌ی پخت یک غذا و یا حتی مواد اولیه بکار رفته در آن غذا می‌تواند متفاوت باشد. در این پروژه قصد داریم تا یک سیستم جستجو برای بازیابی اطلاعات مختلف راجع به دستورات غذایی پیاده‌سازی کنیم.
بخش‌های اصلی این پروژه شامل پردازش متن، نمایه‌سازی، جستجو و رتبه‌بندی اطلاعات بازیابی‌شده می‌باشد. در نهایت هم از شما می‌خواهیم که یک خوشه‌بندی مناسب برای دستورات غذایی ارائه دهید.
در ادامه به توضیح بخش‌های مختلف پروژه می‌پردازیم.

واکشی و دریافت دادگان مسئله

در این بخش شما باید به تعداد مناسبی دستور غذایی دسترسی داشته‌باشید؛ خوب است که دستورات غذایی شما شامل اطلاعات زیر باشد.

- نام غذا
- مدت زمان پخت
- مواد اولیه مورد استفاده در آن غذا
- مراحل و یا توضیحات پخت

بدین منظور شما می‌توانید اطلاعات مدنظر خود را از سایت دلخواه crawl کنید همچنین می‌توانید از دادگان موجود در این لینک استفاده کنید؛ ممکن است در طی این مسیر، نیاز به ثبت‌نام و ورود به حساب kaggle داشته‌باشید.

پردازش متن

در این بخش از شما می‌خواهیم تا برای یکسان‌سازی متن و query، اعمالی مانند جداسازی، نرمال‌سازی، حذف علائم نگارشی، ریشه‌یابی کلمات و حذف کلمات پرتکرار را بر روی داده‌های جمع‌آوری‌شده انجام دهید.
دقت کنید که استفاده زیاد یا کم از هریک از اعمال پیشنهادیِ بالا ممکن است منجر به تضعیف عملکرد سیستم بازیابی شما شود، پس برقراری تعادل در میزان استفاده از هریک از آن‌ها امری مهم است که برعهده خودتان می‌باشد.

نمایه‌سازی

با استفاده از روش مناسب، متن دستورات غذایی و مواد اولیه، نمایه‌سازی می‌گردند؛ نمایه‌های bigram و positional باید مورد استفاده قرار گیرند. همچنین خوب است که از فشرده‌سازی نمایه نیز استفاده کنید.

تصحیح پرسمان و پیشنهاد در هنگام جست‌وجو

در سیستم بازیابی اطلاعات، لازم است تا در صورت مشاهده غلط املایی در پرسمان کاربر، با توجه به دانش لغات سیستم، واژه‌ی صحیح جایگزین شود. همچنین می‌توان هنگام تایپ عبارت مورد جست‌وجو پیشنهاداتی (مرتب شده براساس مرتبط بودن) به‌منظور تکمیل پرسمان کاربر ارائه شود.

بازیابی و رتبه‌بندی

با استفاده از الگوریتم مناسب و با داشتن پرسمان می‌توان عملیات جستجو و رتبه‌بندی را انجام داد و باید مرتبط‌ترین نتایج را بصورت نزولی به کاربر برگردانید.

خوشه‌بندی

یک موضوع مهم در مورد دستورات مختلف غذایی دسته‌بندی مناسب آن‌هاست و با این کار می‌توان فرآیند نگه‌داری و بازیابی آن‌ها را بهبود بخشید. مهم‌ترین داده‌ای که شما می‌توانید از آن در راستای خوشه‌بندی دستورات غذایی استفاده کنید، مواد اولیه مورد استفاده در دستورات غذایی جمع‌آوری شده است؛ پس با با استفاده از الگوریتم و راهکار مناسب دستورات غذایی موجود را بر اساس مواد اولیه مورد استفاده در آن‌ها خوشه‌بندی کنید.

مقدمه

هدف از این تمرین به صورت کلی بررسی و تحلیل و ساخت سامانه‌های بازیابی اسناد برای داده‌های قرآنی است. دریافت کتاب قرآن مجید بر عهده‌ی خود گروه‌هاست که آن را تهیه کنند.

مراحل انجام تمرین

پیش‌پردازش

داده‌های جمع‌آوری شده را با اعمال جداسازی، نرمال‌سازی، حذف علائم نگارشی، ریشه‌یابی کلمات و حذف ایست‌واژه‌ها و یا هر کار مورد ضرورت دیگر، می‌توان پیش‌پردازش کرد.

تصحیح اشکالات املایی

اگر پرسمان ورودی توسط کاربر دارای غلط املایی است، باید در این بخش به اصلاح آن بپردازید.

بازیابی اطلاعات

سیستم شما باید با جستجوی یک آیه‌ی قرآنی بتواند نتایج مورد نظر را به ترتیب رنک به کاربر نمایش دهد.

پیشنهادها

برای سیستم جستجوی خود، می‌توانید موضوعات مختلفی را استخراج کنید. چند نمونه برای مثال ذکر شده؛

۱. سیستم پیشنهاددهنده

در این قسمت سیستم شما با توجه به حدیث ورودی کاربر، چند حدیث مرتبط و مشابه را باید بتواند برگرداند.

۲. استخراج عناصر کلیدی

مشخص کردن واژگان / گزاره‌های کلیدی در احادیث

۳. خوشه‌بندی احادیث

در این بخش می‌توانید مجموعه‌ی احادیث را با روش‌های مختلف خوشه‌بندی کنید و از روی نتایج به‌دست آمده تحلیل‌هایی ارائه دهید.

۴. گراف ارتباط

می‌توانید گراف ارتباط موجود بین عناصر کلیدی احادیث را بسازید.

۵.

موتور جستجوی حدیث

مقدمه

هدف از این تمرین به صورت کلی بررسی و تحلیل و ساخت سامانه‌های بازیابی اسناد برای داده‌های حدیث است. منبع احادیث با تمرکز بر کتاب بحارالانوار یا کتاب اصول کافی است که در اختیار گروه‌ها قرار می‌گیرد.

مراحل انجام تمرین

پیش‌پردازش

داده‌های جمع‌آوری شده را با اعمال جداسازی، نرمال‌سازی، حذف علائم نگارشی، ریشه‌یابی کلمات و حذف ایست‌واژه‌ها و یا هر کار مورد ضرورت دیگر، می‌توان پیش‌پردازش کرد.

تصحیح اشکالات املایی

اگر پرسمان ورودی توسط کاربر دارای غلط املایی است، باید در این بخش به اصلاح آن بپردازید.

بازیابی اطلاعات

سیستم شما باید با جستجوی یک آیه‌ی قرآنی/ حدیثی بتواند نتایج مورد نظر را به ترتیب رنک به کاربر نمایش دهد.

پیشنهادها

برای سیستم جستجوی خود، می‌توانید موضوعات مختلفی را استخراج کنید. چند نمونه برای مثال ذکر شده؛

۱. سیستم پیشنهاددهنده

در این قسمت سیستم شما با توجه به پرس‌وجوی ورودی کاربر، چند سند مرتبط و مشابه را برمی‌گرداند.

۲. استخراج عناصر کلیدی

گزاره‌های کلیدی در سوره‌های مختلف قرآنی چه بوده‌اند؟ استخراج واژگان/گزاره‌های کلیدی شاید بتواند مشخص‌کننده‌ی به نوعی خلاصه‌ی یک سوره باشد!

۳. خوشه‌بندی سوره‌ها

در این بخش می‌توانید سوره‌های قرانی را با روش‌های مختلف خوشه‌بندی کنید و از روی نتایج به‌دست آمده تحلیل‌هایی ارائه داد.

۴. گراف ارتباط

می‌توانید گراف ارتباط موجود بین عناصر کلیدی هر سوره را بسازید.

۵.

موتور جستجوی کتاب‌های تاریخی

مقدمه

هدف از این تمرین به صورت کلی بررسی و تحلیل و ساخت سامانه‌های بازیابی اسناد برای داده‌های تاریخی است. مجموعه‌ی داده‌ی مورد استفاده می‌تواتد برای نمونه استفاده از کتاب تاریخ تمدن ویلدورانت یا کتاب‌های تاریخی موجود دیگر باشد در ادامه برخی از کارهای رایج که می‌تواند بر روی این مجموعه‌ی داده صورت بگیرد را معرفی می‌کنیم:

۱. سیستم پیشنهاددهنده

در این قسمت سیستم شما با توجه به دریافت یک سند ورودی از کاربر، چند سند مرتبط با ورودی را باید بتواند برگرداند.

۲. استخراج عناصر یا شخصیت‌های کلیدی

مشخص کردن شخصیت‌ها/ واژگان / گزاره‌های کلیدی در این کتاب‌ها

۳. خوشه‌بندی متن کتاب بر حسب فصل یا صفحه یا ...

در این بخش می‌توانید کل کتاب را با روش‌های مختلف خوشه‌بندی کنید و از روی نتایج به‌دست آمده تحلیل‌هایی ارائه دهید.

۴. گراف ارتباط

می‌توانید گراف ارتباط موجود بین عناصر (شخصیت‌ها) کلیدی این کتاب‌ها را بسازید.