دانشکده مهندسی کامپیوتر

درس مبانی بازیابی اطلاعات و جستجوی وب

تمرین سری اول

استاد درس: دکتر رهائی

طراح تمرین: الهام غلامی

تاریخ انتشار: فروردین ۱۴۰۴

تاریخ تحویل: فروردین ۱۴۰۴

در رابطه با تمرین

- فایل های پاسخ خود را حتما به صورت یک فایل zip درآورده، به شکل HW1_StudentName_StudentID.zip نام گذاری کرده و ارسال کنید.
- به هیچ وجه تمرینی را از دیگران کپی نکنید. در صورت مشاهده تقلب و کپی در تمرینات، نمره هر دو طرف صفر در نظر گرفته می شود.

۱) پیاده سازی سیستم بازیابی اطلاعات

هدف این سوال آشنایی با بازیابی اطلاعات و جستجو در یک مجموعه داده است. یک دیتاست فارسی پیدا کنید. این دیتاست می‌تواند مجموعه داده‌های یک روزنامه، سایت و ... باشد. با استفاده از کتابخانه‌ای مانند hazm پیش پردازش را انجام دهید. مدل شما باید یک متن را به عنوان ورودی بگیرد و متن‌های مرتبط با متن ورودی را پیدا کند و بر اساس میزان شباهت رتبه‌بندی کند و خروجی دهد.

* راهنمایی: برای محاسبه میزان شباهت از تشابه کسینوسی استفاده کنید.

* پیش‌پردازش را به طور کامل انجام دهید(توکن‌سازی، حذف کلمات اضافه، stemming و lemmatization را انجام دهید.)

* برای افزایش سرعت جستجو چه تکنیکی به ذهنتان می‌رسد؟ پیاده سازی کنید.

* گزارشی از دقت در آزمایش‌های مختلف و کد خود تهیه کنید.

۲) بهبود سیستم بازیابی اطلاعات

در این سوال هدف آشنایی با مفهوم Relevance Feedback است. یکی از مشکلات اصلی در بازیابی اطلاعات این است که موتور جستجو نمیداند کاربر دقیقا چه چیزی را میخواهد. همانطور که می‌دانید هدف این است با تعامل با کاربر خروجی را بهتر کنیم. یکی از الگوریتم‌های کلاسیک این روش، الگوریتم Rocchio است.

(البته شما می‌توانید از هر الگوریتم دیگر نیز استفاده کنید.)

مراحل انجام:

همان سیستم سوال قبل را در نظر بگیرید (می‌توانید دیتاست کوچک‌تری را در نظر بگیرید). سپس یک ورودی به سیستم بدهید و مرتبط بودن ۵ خروجی برتر دیتاست را مشخص کنید.(درواقع شما باید نقش کاربر را بازی کنید) در مرحله بعد مطابق الگوریتم بالا(یا هر الگوریتم دیگر) سیستم را به‌روز رسانی کنید تا نتایج مرتبط تر بشوند.

در نهایت دقت را با دقت به دست آمده در سوال قبل مقایسه کنید و شهودات خود را گزارش کنید.

۳) بررسی شباهت معنایی در اصلاح املای عبارات (با استفاده از شاخص k-gram و مدل نویزی کانال)

در این سوال شما باید بخشی از یک موتور جستجو را طراحی کنید که وظیفه‌اش اصلاح املای عباراتی است که کاربران اشتباه تایپ می‌کنند. برخلاف بسیاری از پروژه‌های کلاسیک اصلاح املای واژه، در این پروژه تمرکز ما بر اصلاح املای "عبارات (multi-word phrases) "است، آن هم با درنظر گرفتن شباهت‌های آوایی، ساختاری و معنایی . مثلا فرض کنید کاربر ورودی ایی مانند machin lernng را تایپ کند. کد شما باید ۴ روش زیر را اجرا کند:

شباهت n-gram (Jaccard Similarity)

مدل نویزی کانال (Noisy Channel) مبتنی بر فاصله ویرایشی (Edit Distance)

شباهت آوایی (Soundex)

شباهت معنایی با استفاده از مدل‌های زبانی مثل spaCy

مثلا برای مثال گفته شده خروجی زیر مطلوب است:

query	final score	k-gram	noise	context	sound
machine learning	0.1960	0.3182	0.0498	-0.0838	0.5000
reinforcement learning	0.0027	0.0606	0.0000	-0.0496	0.0000
deep learning	0.0023	0.0833	0.0003	-0.0745	0.0000