دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه (فاز اول، دوم و سوم)

پاییز ۱۴۰۰

۲- فاز دوم

در این مرحله می‌خواهیم مدل بازیابی اطلاعات را گسترش و بازنمایی اسناد را به صورت برداری انجام دهیم تا بتوانیم نتایج جستجو را بر اساس ارتباط آن‌ها با پرسمان کاربر رتبه‌بندی کنیم. به این صورت که برای هر سند یک بردار عددی استخراج می‌شود که بازنمایی آن سند در فضای برداری است و این بردارها ذخیره می‌شوند. در زمان دریافت پرسمان، ابتدا بردار متناظر با آن پرسمان در همان فضای برداری ساخته و سپس با استفاده از یک معیار شباهت مناسب، شباهت بردار عددی پرسمان با بردار تمام اسناد در فضای برداری محاسبه می‌شود و در نهایت نتایج خروجی بر اساس میزان شباهت مرتب‌سازی می‌شوند. برای افزایش سرعت پاسخگویی مدل بازیابی اطلاعات می‌توان روش‌های مختلفی را به کار گرفت که به تفصیل در ادامه بیان می‌شود.

۲-۱ مدل‌سازی اسناد در فضای برداری

در مرحله قبل پس از استخراج توکن‌ها اطلاعات به صورت یک دیکشنری و شاخص مکانی ذخیره شدند. در این بخش هدف آن است که اسناد در فضای برداری بازنمایی شوند. با استفاده از روش وزن‌دهی $tf-idf$ بردار عددی برای هر سند محاسبه خواهد شد و درنهایت هر سند به صورت یک بردار شامل وزن‌های تمام کلمات آن سند بازنمایی می‌شود. محاسبه‌ی وزن هر کلمه t در یک سند d با داشتن مجموعه‌ی تمام اسناد D با استفاده از معادله‌ی زیر محاسبه می‌شود:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) = \left(1 + \log\left(f_{t,d}\right)\right) \times \log\left(\frac{N}{n_t}\right)$$

که در آن $f_{t,d}$ تعداد تکرار کلمه‌ی t در سند d و n_t تعداد سندهایی است که کلمه‌ی t در آنها ظاهر شده است. توضیحات بیشتر این روش در فصل ۶ کتاب مرجع درس آمده است.

در نمایش برداری فوق برای کلمه‌ای که در یک سند وجود نداشته باشد وزن صفر در نظر گفته می‌شود و از این جهت بسیاری از عناصر بردارهای محاسبه شده صفر خواهد بود. برای صرفه جویی در مصرف حافظه به جای آن که برای هر سند یک بردار عددی کامل در نظر بگیرید که بسیاری از عناصر آن صفر هستند می‌توانید وزن کلمات در اسناد مختلف را در همان لیست‌های پست‌ها ذخیره کنید. در زمان پاسخ‌گویی به پرسمان کاربر که در ادامه توضیح داده می‌شود نیز همزمان با جستجوی کلمات در لیست‌های پست‌ها می‌توانید وزن کلمات در اسناد مختلف را نیز واکشی کنید و به این شکل تنها عناصر غیر صفر بردارهای اسناد ذخیره و پردازش می‌شوند.

۲-۲ پاسخ‌دهی به پرسمان در فضای برداری

با داشتن پرسمان کاربر، بردار مخصوص پرسمان را استخراج کنید (وزن کلمات موجود در پرسمان را محاسبه کنید). سپس با استفاده از معیار شباهت سعی کنید اسنادی را که بیشترین شباهت (کمترین فاصله) را به پرسمان ورودی دارند پیدا کنید. سپس نتایج را به ترتیب شباهت نمایش دهید. معیارهای فاصله‌ی مختلفی می‌تواند برای این کار در نظر گرفته شود که ساده‌ترین آنها شباهت کسینوسی بین بردارها است که زاویه‌ی بین دو بردار را محاسبه می‌کند. این معیار به صورت زیر تعریف می‌شود:

$$similarity(a,b) = \cos(\theta) = \frac{a.b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

توجه کنید که برای افزایش سرعت می‌توانید با استفاده از تکنیک $Index\ elimination$ شباهت کسینوسی را با اسنادی که امتیاز صفر خواهند گرفت محاسبه نکنید. در انتهای کار برای نمایش یک صفحه از نتایج پرسمان تنها کافیست K سندی انتخاب شوند که بیشترین شباهت را به پرسمان دارند.

۲-۳ افزایش سرعت پردازش پرسمان

با استفاده از تکنیک $Index\ elimination$ تا حدودی مشکل زیاد بودن زمان در مرحله قبل حل می‌شود اما همچنان زمان پاسخگویی برای بسیاری از کاربردها قابل قبول نمی‌باشد. برای آنکه سرعت پردازش و پاسخگویی افزایش یابد می‌توانید از $Champion\ lists$ استفاده کنید که قبل از آنکه پرسمانی مطرح شود و در مرحله پردازش اسناد، یک لیست از مرتبط‌ترین اسناد مربوط به هر $term$ در لیست جداگانه‌ای نگهداری شود. برای پیاده‌سازی این بخش پس از ساخت شاخص معکوس مکانی، $Champion\ list$ را ایجاد کنید و تنها بردار پرسمان را با بردار اسنادی که از طریق جستجو در $Champion\ list$ به دست آورده‌اید مقایسه کنید و K سند مرتبط را به نمایش بگذارید. توضیحات بیشتر این روش در فصل ۷ کتاب آمده است.

توجه: می‌توانید وزن دهی $tf-idf$ و ایجاد لیست $Champion$ را با استفاده از شاخص مکانی که در مرحله قبل پیاده‌سازی کردید، انجام دهید.

۲-۴ گزارش

۱. پاسخ به پرسمان در حالت‌های زیر:

الف) یک پرسمان از کلمات ساده و متداول تک کلمه‌ای

ب) یک پرسمان از عبارات ساده و متداول چند کلمه‌ای

پ) یک پرسمان دشوار و کم تکرار تک کلمه‌ای

ت) یک پرسمان دشوار و کم تکرار چند کلمه‌ای

در هر مورد، تیتر خبر بازیابی شده را به همراه جمله(هایی) که حاوی عبارت پرسمان بوده‌اند، گزارش کنید. همچنین در هر مورد با ذکر جزئیات شرح دهید که آیا سند بازیابی شده به پرسمان کاربر مرتبط هست یا خیر؟
۲. موارد ب و ت را با روش مکانی فاز یک نیز تکرار کنید و نتایج دو حالت را با هم مقایسه و تحلیل کنید.

۲-۵ بازنمایی با استفاده از تعبیه‌گذاری کلمه[3] (اختیاری)

در بخش قبل مشاهده کردید که برای نگهداری بردارهای اسناد به صورت *tf-idf* با چالش حافظه روبرو هستید. همچین در حالت *tf-idf* به دلیل طول بسیار زیاد بردارها، در زمان بازیابی چالش زمان نیز مطرح است. از آنجا که در کارهای صنعتی و تحقیقاتی نیز با حجم قابل توجهی داده روبه‌رو هستیم، می‌خواهیم با روش‌های نوین بازنمایی اسناد آشنا شویم که فرم فشرده‌تری از بازنمایی را ارائه می‌دهند. هدف از این بخش، بازنمایی اسناد با استفاده از تعبیه‌گذاری کلمه است.در این دسته از روش‌های بازنمایی برای هر کلمه یک بردار با طول حدودا ۲۰۰ یا ۳۰۰ بُعد بدست می‌آید، این بردارها با توجه به مجاورت کلمات در اسناد آموزشی ساخته می‌شوند بنابراین می‌توانند تا حدی (با توجه به روش‌های مختلف) بافت متن را در ساخت بردار کلمه دخیل کنند. در این روش بعد از بدست آوردن بردار کلمه، بردار کل متن را بدست می‌آوریم.

۲-۵-۱ بازنمایی اسناد

در این قسمت لازم است با استفاده از word2vec مدل skip-gram بازنمایی اسناد را به دست آورید. برای این کار می‌توانید از کتابخانه‌های آماده استفاده کنید. (راهنمایی: کتابخانه‌ی gensim). پس از آموزش مدل word2vec، به ازای هر کلمه یک بردار ۳۰۰ بُعدی که بیانگر کلمه در فضای برداری است، خروجی داده می‌شود. برای بازنمایی سند لازم است دو روش زیر پیاده‌سازی شود:
۱. آموزش مدل بر روی مجموعه دادگان فاز اول و محاسبه‌ی بردار بازنمایی هر سند با استفاده از میانگین وزن‌دار کلمات آن سند به صورتی که وزن هر کلمه معادل با *tf-idf* متناظر با آن کلمه باشد.
۲. استفاده از بازنمایی کلمات موجود در مجموعه بردارهای از پیش آموزش داده شده با استفاده از word2vec بر روی حجم زیادی از مجموعه داده اخبار و سپس استفاده از میانگین وزن‌دار بازنمایی کلمات سند به منظور محاسبه بازنمایی هر سند. (مجموعه بردارهای از پیش آموزش داده شده ذکرشده در فایل فشرده شده new.fa.text.300.vec.zip موجود است.)

۲-۵-۲ بازنمایی پرسمان

با دریافت پرسمان کاربر لازم است بردار متناظر با آن ساخته و سپس مشابه با مرحله‌ی دوم پروژه، شباهت کسینوسی بردار پرسمان با تمام اسناد محاسبه شود. در نهایت K سند مرتبط بصورت رتبه‌بندی شده نمایش

[3] Word Embedding

داده شود. لازم به ذکر است روش استفاده شده برای ساخت بردار پرسمان باید مشابه با روش بازنمایی اسناد باشد.

۲-۵-۳ تحلیل عملکرد مدل بازیابی اطلاعات و گزارش

معیارهای mean reciprocal rank و mean average precision و precision@k (به ازای kهای ۱و۵) را برای حالت‌های مختلف پرسمان (اعم از پرسمان کوتاه، طولانی، عبارت پرسشی با کلمات رایج و عبارت پرسشی با کلمات نادر) محاسبه کنید. به ازای هر حالت از پرسمان، عملکرد مدل در حالت بازنمایی *tf-idf* را با بازنمایی word2vec (در هر دو حالت بازنمایی با مدل از پیش آموزش داده شده و مدلی که خودتان آموزش داده‌اید) مقایسه و نتایج را تحلیل کنید. در تحلیل‌های خود لازم است دلیل بهتر بودن عملکرد هر بازنمایی برای هر نوع از پرسمان‌ها ذکر کنید.

توجه: برای برچسب گذاری باینری اسناد بازیابی‌شده لازم است محتوای سند بازیابی‌شده به پرسمان مد نظر شما پاسخ دهد. بطور مثال اگر شما می‌خواهید در مورد نرخ مسکن در محدوده‌ی میدان آزادی بدانید، سند بازیابی شده‌ای که به این سوال شما پاسخ می‌دهد برچسب "صحیح" و سندی که کلمات پرسمان را دارد اما به سوال شما پاسخ نمی‌دهد، برچسب "اشتباه" می‌گیرد.

توجه: در هر آزمایش لازم است عبارت پرسمان، عنوان اخبار بازیابی‌شده، برچسب هر خبر و نحوه‌ی محاسبه‌ی معیارها گزارش شود.