درس:

بازیابی اطلاعات

تعریف پروژه (فاز اول، دوم و سوم)

پاییز ۱۴۰۰

مقدمه

هدف از این پروژه ایجاد یک موتور جستجو برای بازیابی اسناد متنی است به گونه‌ای که کاربر پرسمان خود را وارد نموده و سامانه اسناد مرتبط را بازنمایی می‌کند. پروژه در سه مرحله تعریف شده است که عبارتند از:

مرحله‌ی اول: ایجاد یک مدل بازیابی اطلاعات ساده

مرحله‌ی دوم: تکمیل مدل بازیابی اطلاعات و ارائه‌ی قابلیت‌های کارکردی پیشرفته‌تر

مرحله‌ی سوم: پیاده‌سازی الگوریتم خوشه‌بندی و دسته‌بندی و بازیابی بر اساس خوشه/ دسته

در انجام پروژه به نکات زیر توجه فرمایید:

- پروژه انفرادی است.
- تنها در موارد ذکرشده در تمرین مجاز به استفاده از کتابخانه‌های آماده هستید.
- کدهای خود را در کوئرا بارگذاری نمایید (آدرس مربوطه در سایت درس قرار داده می‌شود).
- کدهای شما (به همراه کدهای دانشجویان ترم‌های گذشته) توسط کوئرا بررسی می‌شود. در صورت وجود شباهت، نمره‌ی تمام فازهای پروژه صفر خواهد شد.
- ملاک اصلی انجام فعالیت ارائه گزارش مربوطه است و ارسال کد بدون گزارش فاقد ارزش است. سعی کنید گزارش شما دقیقا در راستای موارد خواسته‌شده باشد و از طرح موارد اضافی خودداری کنید.
- مهلت ارسال فاز اول پروژه تا پایان روز ۲۸ آبان‌ماه، فاز دوم تا پایان روز ۱۹ آذرماه و فاز سوم تا پایان روز ۱ دی‌ماه می‌باشد.
- فازهای یک، دو، سه و بخش امتیازی به ترتیب ۳۰، ۴۰، ۳۰ و ۳۰ درصد از نمره‌ی پروژه را به خود اختصاص می‌دهند.
- به ازای هر روز تاخیر در فاز اول و دوم ۵ درصد از نمره‌ی فاز مربوطه کسر می‌شود.
- ارسال فاز سوم با تاخیر امکان پذیر نخواهد بود.
- موعد تحویل متعاقبا از طریق سایت درس اعلام خواهد شد.**راهنمایی:**

در صورت نیاز می‌توانید سوالات خود در خصوص پروژه را از تدریسیاران درس، از طریق ایمیل زیر بپرسید.

IR.course1400@gmail.com

۱- فاز اول

در این فاز از پروژه به منظور ایجاد یک مدل بازیابی اطلاعات ساده نیاز است تا اسناد شاخص‌گذاری شوند تا در زمان دریافت پرسمان از شاخص مکانی برای بازیابی اسناد مرتبط استفاده شود. به طور خلاصه مواردی که در این فاز انجام شوند به شرح زیر می‌باشد.

- پیش‌پردازش داده‌ها
- ساخت شاخص مکانی
- پاسخ‌دهی به پرسمان کاربر

در ادامه هر مورد به صورت کامل شرح داده می‌شود.

۱-۱ پیش‌پردازش اسناد

قبل از ساخت شاخص مکانی لازم است متون را پیش‌پردازش کنید. گام‌های لازم در این قسمت به صورت زیر می‌باشد.

- استخراج توکن
- نرمال‌سازی متون
- حذف کلمات پر تکرار[1]
- ریشه‌یابی

برای انجام پیش‌پردازش‌های لازم می‌توانید با صلاح‌دید خود یکی از کتابخانه‌های آماده را انتخاب و از آن استفاده کنید (راهنمایی: کتابخانه۱ و کتابخانه۲) و یا پیاده‌سازی شخصی خود را داشته باشید.

توجه: برای پیاده‌سازی شخصی بخش‌های مربوط به پیش‌پردازش اسناد نمره‌ی ارفاقی لحاظ نمی‌شود.

۱-۲ ساخت شاخص مکانی

با استفاده از اسناد پیش‌پردازش‌شده در گام قبل، شاخص مکانی را بسازید. در شاخص مکانی ساخته شده علاوه بر جایگاه کلمات در اسناد، باید به ازای هر کلمه از دیکشنری مشخص باشد که تعداد تکرار آن کلمه در کل اسناد چقدر است. همچنین باید مشخص باشد که در هر سند تعداد تکرار یک کلمه‌ی مشخص چقدر است. جزئیات کامل این قسمت در بخش ۲.۴.۲ از کتاب مرجع درس قابل مشاهده است. برای پیاده‌سازی این قسمت

[1] Stop Words

می‌توانید به اختیار خود یک ساختمان داده‌ی مناسب را انتخاب کنید. (دقت کنید که ساختمان داده‌ی انتخابی به‌گونه‌ای نباشد که در زمان جستجو و دیگر عملیات، سرعت مدل را پایین آورد.)

۱-۳ پاسخ‌دهی به پرسمان کاربر

در این بخش با دریافت پرسمان کاربر باید بتوانید اسناد مرتبط با آن را به صورت دودویی[2] بازیابی نمایید. پرسمان کاربر به دو صورت زیر می تواند باشد:

تک کلمه: تنها کافی است که لیست اسناد مربوط به آن را از روی دیکشنری بازیابی نمایید.

چند کلمه: در این بخش لیست فایل‌ها باید بر اساس میزان ارتباط مرتب شده باشد. مرتبط‌ترین سند، سندی است که تمام کلمات را به همان ترتیب موجود در پرسمان داشته باشد. (به طور مثال اگر پرسمان شامل ۳ کلمه بود، سندی مرتبط‌تر است که هر سه کلمه را داشته باشد، بعد از آن سندی مرتبط است که دو کلمه از کلمات پرسمان را در خود دارد.)

۱-۴ مجموعه داده

مجموعه داده مورد استفاده در این پروژه مجموعه‌ای از خبرهای واکشی‌شده از چند وب‌سایت خبری فارسی است که در قالب یک فایل اکسل در اختیار شما قرار خواهدگرفت. لازم است تنها ستون "content" را بعنوان محتوای سند پردازش کنید. شماره‌ی هر خبر را به عنوان id آن سند (خبر) در نظر بگیرید و در زمان پاسخ به پرسمان، عنوان خبر مربوط به سند بازیابی‌شده را نمایش دهید تا امکان بررسی صحت عملکرد سیستم وجود داشته‌باشد.

۱-۵ گزارش

۱. با ذکر مثال شرح دهید که در گام پیش‌پردازش چه عملیاتی انجام داده‌اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

۲. صحت قانون Zipf را در دو حالت قبل از حذف کلمات پرتکرار از واژه‌نامه و بعد از حذف کلمات پرتکرار بررسی کنید. در صورت برقراری/ عدم برقراری این قانون در هر حالت، علت را شرح دهید.

۳. صحت قانون heaps را در دو حالت قبل و بعد از ریشه‌یابی بررسی کنید. برای بررسی این قانون لازم است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت مقایسه و تحلیل شود. آیا در هر دو حالت قانون برقرار است؟ چرا؟

۴. حداقل سه مورد از مواردی که در ریشه‌یابی با چالش روبرو بودید را ذکر کنید. (بطور مثال کلماتی که نیازی به ریشه‌یابی ندارند اما طبق روند ریشه‌یابی از دست می‌روند.)

[2] Boolean

۵. پاسخ به پرسمان در حالت‌های زیر:

الف) یک پرسمان از کلمات ساده و متداول تک کلمه‌ای (بین‌الملل)

ب) یک پرسمان از عبارات ساده و متداول دو کلمه‌ای (دانشگاه امیرکبیر)

پ) یک پرسمان از عبارات ساده و متداول چند کلمه‌ای (دانشگاه صنعتی امیرکبیر، سازمان ملل متحد، جمهوری اسلامی ایران)

ت) یک پرسمان دشوار و کم تکرار تک کلمه‌ای (ژیمناستیک)

ث) یک پرسمان دشوار و کم تکرار دو کلمه‌ای (واکسن آسترازنکا)

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- 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.

۲- فاز سوم

در این بخش از پروژه مقیاس موتور جستجویی که در دو مرحله‌ی گذشته طراحی و پیاده‌سازی شده، بزرگ‌تر می‌شود. با افزایش حجم اسناد ورودی، مقایسه پرسمان با تمام اسناد به صورت کارا و در زمان مناسب امکان‌پذیر نیست. در این فاز برای حل این مسئله می‌خواهیم از خوشه‌بندی استفاده کنیم و بردار ویژگی پرسمان را به جای مقایسه با تمام اسناد فقط با اسناد یک (یا چند) خوشه مقایسه کنیم. علاوه بر خوشه‌بندی، دسته‌بندی اخبار نیز در این مرحله از پروژه بایستی پیاده‌سازی شود. به این معنا که هر خبر به یکی ازدسته‌های ورزشی، اقتصادی، سیاسی، سلامت و فرهنگی نگاشت شود تا در هنگام جستجو بتوان مشخص کرد نتایج از کدام دسته‌های خبری باشند. در ادامه به توضیح بیشتر در این خصوص می پردازیم.

توجه: در این مرحله می‌توانید برای بازنمایی اسناد از روش «بازنمایی با استفاده از تعبیه‌گذاری کلمه» نیز استفاده نمایید.

۳-۱ خوشه‌بندی

در این مرحله میخواهیم با استفاده از الگوریتم K-means خوشه بندی اسناد را انجام دهید. به منظور بهبود عملکرد الگوریتم خوشه‌بندی می‌توانید چندین بار آن را اجرا و سپس بر مبنای معیار RSS بهترین خوشه‌بندی را انتخاب کنید. بعد از انتخاب یک خوشه‌بندی مناسب، در زمان پاسخگویی به یک پرسمان، ابتدا بردار بازنمایی آن را مطابق با روش موردنظر استخراج کنید. سپس شباهت کسینوسی آن را با تمام مراکز خوشه‌ها محاسبه کرده و خوشه با بیشترین شباهت را انتخاب کنید. در نهایت شباهت کسینوسی بردار پرسمان با تمام سندهای آن خوشه را محاسبه کرده و از میان آنها شبیه‌ترین سندها به پرسمان را انتخاب و به عنوان نتیجه جستجو برگردانید.

توجه کنید لزومی بر اینکه فقط یک خوشه را برای جستجو انتخاب کنید وجود ندارد. به این معنی که بعد از محاسبه‌ی شباهت بردار پرسمان با مراکز خوشه ها، می‌توانید b مرکز خوشه با بیشترین شباهت را انتخاب کرده و جستجو را در تمام اسناد خوشه‌های مربوط به آنها انجام دهید. این کار خصوصا زمانی موثر است که تعداد خوشه ها زیاد باشد و در نتیجه تعداد اسناد در یک خوشه کم شده باشد. انتخاب مقدار b و تعداد خوشه‌ها با هم مرتبط هستند و بهترین مقادیر آنها مقادیری است که یک تعادل بین سرعت پاسخگویی و

کیفیت نتایج ایجاد کند. ارزیابی دقیق این موضوع مستلزم اندازه‌گیری دقیق زمان پاسخ به پرسمان‌های کاربر و دقت نتایج بازگردانده شده بر روی مجموعه‌ای از پرسمان‌های از قبل آماده شده است. در این پروژه می‌توانید این کار را به صورت شهودی انجام دهید و تنظیم دقیق مقدار b الزم نیست.

توجه: در این قسمت استفاده از کتابخانه مجاز نیست.

۳-۲ دسته‌بندی

موتور جستجوی طراحی شده در این حالت می بایست قابلیت تعیین دسته خبر را در زمان وارد کردن پرسمان به کاربر بدهد. این قابلیت با استفاده از کلمه کلیدی cat ارائه می‌گردد. به عنوان مثال زمانی که کاربر عبارت «استقلال cat:sport» را وارد می‌کند می‌بایست بازیابی در بین اخبار دسته‌ی ورزشی و زمانی که عبارت «استقلال cat:economy» را وارد می‌کند می‌بایست بازیابی در بین اخبار دسته‌ی اقتصادی انجام شود. بدین منظور با استفاده از روش‌های دسته‌بندی اسناد متنی ارائه شده در درس، دسته هر خبر را تعیین و ذخیره کنید تا در زمان جستجو بتوان از آن استفاده کرد. دسته‌های خبری مد نظر عبارتند از:

ورزشی، اقتصادی، سیاسی، فرهنگی، سلامت.

برای دسته بندی اسناد از الگوریتم k-نزدیکترین همسایه با مقادیر مختلف k استفاده کنید. در ابتدا باید الگوریتم دسته بند را پیاده‌سازی کنید و سپس با استفاده از مجموعه اسنادی که برچسب دارند (فایل ۵۰ هزار خبری)، اسنادی که برچسب ندارند (فایل ۷ هزار خبری) را برچسب بزنید. سعی کنید یک مقدار مناسب برای k پیدا کنید. برای پیدا کردن k مناسب و ارزیابی عملکرد دسته‌بند خود می‌توانید از روش ارزیابی 10-fold-cross-validation استفاده کنید.

توجه: در این قسمت مجاز به استفاده از کتابخانه نیستید ولی برای ارزیابی 10-fold-cross-validation می‌توانید از کتابخانه استفاده کنید.

۳-۲ گزارش

۱. سه پرسمان مناسب را انتخاب کرده، نتایج را از نظر عملکرد و سرعت موتور جستجو برای این سه پرسمان در دو حالت بدون خوشه‌بندی و با خوشه‌بندی مقایسه و تحلیل نمایید.

۲. به ازای هر دسته یک پرسمان مناسب انتخاب کنید و نتایج جستجوی این پرسمان را در دو حالت با دسته‌بندی و بدون دسته‌بندی مقایسه و تحلیل کنید.

(ذکر جزئیات در پرسمان‌ها و نتایج بازیابی‌شده در گزارش الزامی است.)