دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه (فاز اول، دوم و سوم)

پاییز ۱۴۰۰

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.)