بازیابی پیشرفته اطلاعات

نیم‌سال اوّل ۹۹-۹۸

دانشکده‌ی مهندسی کامپیوتر

مدرس: حمید بیگی

فاز دوم پروژه (۱۰۰ نمره)

زمان تحویل: ۲۹ آذر ماه

هدف از فاز دوم پروژه پیاده‌سازی الگوریتم‌های دسته‌بندی و در نهایت مقایسه با معیارهای خواسته شده و گزارش نتایج است. الگوریتم‌های k-NN و Naive Bayes را از پایه و روش‌های SVM و Random Forest را با استفاده از کتاب‌خانه‌های آماده پیاده‌سازی کنید. مجموعه داده‌ای که در اختیارتان قرار دارد به زبان انگلیسی بوده و توسط موتور جستجوی آکادمیک ComeToMyHead جمع‌آوری شده است. توجه نمایید که این مجموعه‌ی داده نسخه‌ی کامل‌تری از آنچه در فاز اول به شما داده شد است. تعداد ۹۰۰۰ مستند برای تمرین (train) و ۱۰۰۰ مستند برای آزمون (test) در اختیار شما قرار گرفته است. اینکه چه مقدار از داده‌های تمرین را استفاده کنید بستگی به خودتان دارد. توجه کنید که مجموعه‌ی داده تمرین جدید شامل سه ستون است. ستون اول نشان‌دهنده‌ی دسته‌ی مستند است که عددی از ۱ تا ۴ است. ستون دوم و سوم نیز به ترتیب عنوان و متن خبر می‌باشند. شماره‌ی دسته‌ها به صورت مقابل است:

۱. World
۲. Sports
۳. Business
۴. Sci/Tech

بخش ۱. پیاده‌سازی دسته‌بندها (۶۰ نمره)

در این بخش از پروژه ابتدا باید مجموعه داده‌هایی که در اختیارتان قرار گرفته است را خوانده و مراحل پیش‌پردازش لازم را بر آن‌ها اعمال کنید. همانند فاز قبل می‌توانید از کتاب‌خانه‌های آماده برای پیش‌پردازش استفاده کنید. در ادامه باید اسناد را به فضای برداری tf-idf به روش ntn (رابطه‌ی $tf \times \log(\frac{N}{df})$) برده و در نهایت الگوریتم‌های دسته‌بندی را پیاده‌سازی نمایید.

۱. **Naive Bayes**: الگوریتم Naive Bayes را از پایه پیاده‌سازی کرده و بر روی داده‌های آموزش اجرا کنید.

۲. **k-NN**: الگوریتم k-NN را از پایه پیاده‌سازی کنید و بر روی داده‌های آموزش اجرا کنید. توجه کنید که پارامتر k ورودی برنامه‌ی شما باشد. این الگوریتم را با مقادیر ۱، ۵ و ۹ برای k اجرا کرده و نتیجه را گزارش دهید.

۳. **SVM**: الگوریتم SVM را با توجه به کتاب‌خانه‌های موجود برای حالت Soft Margin پیاده‌سازی کنید. توجه کنید که پارامتر C ورودی برنامه‌ی شما باشد. این الگوریتم را با مقادیر $\frac{1}{2}$، ۱، $\frac{3}{2}$ و ۲ برای C اجرا کرده و نتیجه را گزارش دهید.

۴. **Random Forest**: الگوریتم Random Forest را با کمک کتاب‌خانه‌های موجود بر روی داده‌های آموزش پیاده‌سازی کنید.

توجه: برای مقایسه‌ی پارامتر در الگوریتم‌های k-NN و SVM از بخشی از داده‌ی آموزش (مثلا ۱۰ درصد) به عنوان داده‌ی Validation استفاده کنید. در نهایت با اجرای الگوریتم به ازای هر پارامتر بر روی باقیمانده‌ی داده‌ی آموزش (توجه کنید که همچنان لازم نیست از تمام داده استفاده کنید) و مقایسه نتایج حاصل از داده‌ی Validation بهترین پارامتر را به دست آورده و گزارش دهید.

بارم‌بندی

۱. الگوریتم Naive Bayes (۲۰ نمره)

۲. الگوریتم k-NN (۲۰ نمره)

۳. الگوریتم SVM (۱۰ نمره)

۴. الگوریتم Random Forest (۱۰ نمره)

بخش ۲. بهبود سیستم بازیابی اطلاعات فاز اول پروژه (۳۰ نمره)

در این بخش قصد داریم تا با استفاده از الگوریتم‌های دسته‌بندی که در بخش قبل پیاده‌سازی کرده‌اید، سیستم بازیابی اطلاعاتی که در فاز اول پروژه پیاده‌سازی کردید را بهبود دهیم. برای این منظور می‌خواهیم که سیستم بازیابی اطلاعات فاز اول پروژه علاوه بر ورودی‌های پیشین، شماره یا نام دسته‌ای که قصد داریم جستجو در داخل آن انجام گیرد را نیز از کاربر دریافت نماید و ضمن دسته‌بندی مستندهای ارائه شده در فاز اول پروژه، اقدام به جستجو با توجه به ورودی کاربر نماید. توجه کنید که برای ورودی دادن دسته می‌توان شماره (از ۱ تا ۴) و یا نام دسته را از طریق رابط کاربری به سامانه ارائه کرد. انتخاب نحوه‌ی خواندن ورودی بر عهده‌ی خودتان است.

بارم‌بندی

۱. دسته‌بندی داده‌های فاز اول و افزودن امکان جستجو بر اساس دسته (۳۰ نمره)

بخش ۳. ارزیابی نهایی (۱۰ نمره)

در این بخش باید برای تمامی الگوریتم‌های پیاده‌سازی شده در بخش ۱ (با بهترین پارامتر به دست آمده) معیارهای خواسته شده در زیر را بر روی داده‌های آموزش و آزمون گزارش دهید.

۱. Accuracy

۲. F_1 با $\beta = 1$ و $\alpha = \frac{1}{2}$

۳. Precision و Recall برای هر کلاس

بارم‌بندی

۱. ارزیابی الگوریتم‌های دسته‌بندی (۱۰ نمره)

بخش ۴. نکات

۱. امکان تغییر بارم‌بندی وجود دارد.

۲. نوشتن گزارش فراموش نشود. به قوانین کلاس و پروژه که در پیاتزا قرار گرفته است رجوع کنید.