بسم الله الرحمن الرحیم

فاز دوم پروژه بازیابی پیشرفته اطلاعات

دسته‌بندی و خوشه‌بندی داده‌های متنی

مدرس : مهدیه سلیمانی

نیم‌سال دوم سال تحصیلی ۹۸-۹۹

دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

مقدمه

این فاز پروژه از دو قسمت دسته‌بندی و خوشه‌بندی تشکیل شده است. در قسمت اول روش‌های دسته‌بندی kNN و Naive Bayes را از پایه پیاده‌سازی می‌کنید و با استفاده از کتابخانه‌های موجود روش‌های Random Forest و SVM اجرا می‌کنید. در قسمت دوم، الگوریتم K-Means را برای خوشه‌بندی پیاده‌سازی می‌کنید و با استفاده از روش t-SNE نتایج به دست آمده را تحلیل خواهید کرد.

مجموعه داده‌ها

مجموعه داده‌ای که برای این بخش انتخاب شده، قسمتی از مجموعه داده‌ی AG News است که شامل بیشتر از ۱ میلیون سند در دسته‌های مختلف است. در این پروژه تنها از قسمت کوچکی از مجموعه داده اصلی استفاده خواهیم کرد. هر سند شامل عنوان، متن و دسته‌ی خبر است. هر خبر در یکی از چهار دسته زیر قرار دارد:

World: 1
Sports: 2
Business: 3
Sci/Tech: 4

این مجموعه داده، در دو دسته‌ی آموزش (training) و اعتبارسنجی (validation) در اختیار شما قرار گرفته است.

معیارهای ارزیابی عملکرد پیاده‌سازی

۱- Precision و Recall به ازای هر دسته

۲- Accuracy

۳- Confusion matrix

۴- Macro averaged F_1 به ازای $\beta = 1$

بخش اول (۷۰ نمره + ۱۰ نمره امتیازی)

همان‌طور که اشاره شد، در این قسمت اخبار را دسته‌بندی خواهید کرد. ۱۰ نمره از این بخش به دقت الگوریتم پیاده‌سازی شما تعلق خواهد گرفت.

KNN (۳۰ نمره)

ابتدا اسناد را به فضای tf-idf و حالت ntn برده (۵ نمره) و سپس الگوریتم kNN را به ازای k های ۱، ۳ و ۵ و همچنین معیار فاصله‌های cosine similarity و Euclidean distance پیاده‌سازی کنید (۱۵ نمره). سپس با استفاده از معیارهای به دست آمده از داده‌های اعتبارسنجی، بهترین مقدار برای پارامتر k را گزارش کنید (۱۰ نمره).

Naive Bayes (۱۵ نمره)

الگوریتم Naive Bayes با smoothing را بر روی داده‌های آموزش، آموزش دهید. با استفاده از داده‌های ارزیابی، بهترین پارامتر (α) smoothing را پیدا کنید.

$$P(t,c) = \frac{T_{t,c}}{\Sigma_{t' \in V} T_{t',c}}$$

$$\hat{P}(t,c) = \frac{T_{t,c} + \alpha}{(\Sigma_{t' \in V} T_{t',c}) + \alpha|V|}$$

تاثیر روش‌های پردازش متن بر روی دسته‌بندی (۱۵ نمره)

در این بخش با استفاده از کتابخانه nltk اثر stemming، lemmatization و stopword removal را بر روی مدل های بخش قبل بررسی می‌کنیم (می‌توانید از لیست stopword های موجود در کتاب‌خانه‌ی nltk استفاده کنید). هر کدام از روش‌های معرفی شده را به تنهایی بر روی مدل‌های قسمت قبل اعمال کنید و معیارهای ارزیابی جدید را محاسبه کنید. نتایج به دست آمده را مقایسه کنید و در گزارش خود بنویسید. کدام روش‌ها بهترین و بدترین اثر را بر روی مدل دارند؟ (برای الگوریتم k-NN تنها یکی از حالت‌ها را به دلخواه خود انتخاب کنید.)

دقت دسته بندی (۱۰ نمره امتیازی)

معیار ارزیابی در این بخش دقت مدل شما بر روی داده‌های تست(که در دسترس شما نیست) خواهد بود. برای ارزیابی دو فایل judge.py و model.py در اختیار شما قرار گرفته است. از بین مدل‌هایی که در قسمت قبل پیاده سازی کرده‌اید بهترین مدل را انتخاب کرده و در فایل model.py قرار دهید. سپس با اجرای judge می توانید دقت مدل خود را بر روی داده‌های اعتبارسنجی مشاهده کنید.

دقت شود که برای بهبود دقت دسته‌بندی، تنها استفاده از الگوریتم‌های k-NN و Naive Bayes (که در بخش قبل پیاده‌سازی کرده‌اید)مجاز است و تنها می‌توانید از کتابخانه‌های numpy و nltk استفاده کنید.

استفاده از کتابخانه scikit-learn (۱۰ نمره)

هدف از این قسمت، پیاده‌سازی دو الگوریتم دسته‌بندی SVM و Random Forest با استفاده از کلاس‌های آماده کتابخانه scikit-learn و سپس گزارش دقت بر روی داده‌های اعتبار سنجی می‌باشد.

برای الگوریتم SVM، از مدل SVC با کرنل خطی استفاده کنید و سعی کنید بهترین مقدار پارامتر رگولارایزر C را با استفاده از معیارهای ارزیابی بیابید. در مورد Random Forest نیز از تعدادی درخت تصمیم استفاده کنید و با تغییر هایپرپارامترها(تعداد درخت‌ها و عمق آن‌ها) و استفاده از معیارهای ارزیابی، بهترین هایپرپارامترها را گزارش نمایید.

بخش دوم (۳۰ + ۱۰ نمره امتیازی)

در این بخش با استفاده از نمایش برداری داده‌ها که در قسمت قبل با استفاده از tf-idf به دست آوردید، الگوریتم k-means را با استفاده از ۴ خوشه پیاده‌سازی کنید. در نهایت با استفاده از t-SNE خوشه‌ها را در فضای ۲ بعدی نمایش دهید و با برچسب واقعی داده‌ها مقایسه کنید. نتایج به دست آمده را به صورت خلاصه تحلیل کنید.

- نیازی به نمایش همه‌ی داده‌ها در t-SNE نیست. قسمتی از داده‌ها را از هر دسته انتخاب کنید.

- برای t-SNE می‌توانید از کتاب‌خانه‌های موجود مانند sklearn استفاده کنید.

t-SNE روشی برای کاهش ابعاد داده‌ها از فضای بالا به فضایی با ابعاد پایین‌تر می‌باشد (که معمولا ابعاد فضای پایین‌تر ۲ یا ۳ بعدی می‌باشد که بتوان داده‌ها را در آن فضا نمایش داد). به همین منظور توزیعی بر روی جفت داده‌ها در فضای ابعادی بالاتر در نظر گرفته می‌شود که در آن احتمال انتخاب جفت‌هایی که شباهت زیادی بهم دارند، بالاتر است (می‌توان معیار شابهت را بر اساس فاصله اقلیدسی سنجید). سپس سعی می‌شود تا داده‌ها به نحوی به فضای ابعادی پایین‌تر منتقل شود که توزیع میان جفت داده‌ها تا حد ممکن حفظ شود.

بخش امتیازی

سعی کنید با استفاده از روش word2vec نمایش برداری مناسب برای هر سند به دست آورید و قسمت قبل (خوشه‌بندی و نمایش خوشه‌ها) را با استفاده از این نمایش جدید اجرا کنید. تأثیر پارامترهای طول پنجره و ابعاد embedding را بر روی نتایج بررسی کنید. به عنوان پیشنهاد می‌توانید از کتابخانه‌ی gensim استفاده کنید.