به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۵

دی ماه ۱۴۰۰

*فهرست

شرح دادگان و کد

بخش ۱

مجموعه‌داده‌ی IMDB_Movie_Reviews.csv برای بخش ۱ تمرین در نظر گرفته شده است. این مجموعه‌داده شامل ۲۵۰۰۰ نظر درباره‌ی فیلم‌ها است که این نظرات بر اساس احساسات (مثبت/منفی) برچسب‌گذاری شده‌اند. هر نظر به یکی از دو رده‌ی 0 و 1 تعلق دارد که برچسب 0، متناظر با نظر منفی و برچسب 1، متناظر با نظر مثبت است.

بخش ۲

مجموعه‌داده‌ی dataset.txt شامل ۱۰۰ سند از Associated Press می‌باشد که برای انجام بخش ۲ تمرین در نظر گرفته شده است.

فایل PLSA.py شامل کد پیاده‌سازی PLSA به زبان پایتون برای مدل کردن موضوع می‌باشد. متد preprocessing جهت انجام پیش‌پردازش‌های لازم همچون tokenization، حذف stop word ها، ایجاد ماتریس document-word و ... می‌باشد. طبق کامنت‌ها این متد را کامل کنید. (می‌توانید از کتابخانه‌های NLTK و Scikit-learn استفاده نمایید.)

پیش‌نیازها

- برای انجام پیاده‌سازی‌ها حتما از زبان پایتون استفاده کنید.
- لطفاً کدهای خود را تا حد امکان مرتب و همراه با کامنت بنویسید.

بخش ۱ – Text Classification

هدف این تمرین، رده‌بندی احساسات (Sentiment Classification) است. برای این بخش، استفاده از کتاب‌خانه‌های Scikit-learn و NLTK پیشنهاد می‌شود.

۱. در ابتدا داده‌ها را با تکنیک‌هایی مانند حذف stop word ها و stemming پیش‌پردازش کنید. برای این منظور، استفاده از کتاب‌خانه‌ی NLTK پیشنهاد می‌شود. سپس، ۸۰ درصد دادگان را به صورت تصادفی انتخاب کرده و آن‌ها را به عنوان دادگان آموزشی در نظر بگیرید و مابقی دادگان را نیز به عنوان دادگان تست در نظر بگیرید.

۲. تعدادی از کلمات را به عنوان ویژگی انتخاب کنید و برای هر نظر، تعداد وقوع هر یک از کلمات انتخاب شده را به عنوان مقدار متناظر با آن ویژگی در نظر بگیرید. توضیح دهید که معیار شما برای انتخاب کلماتی که به عنوان ویژگی در نظر می‌گیرید، چیست.

۳. رده‌بندهای Naïve Bayes، Logistic regression و SVM را بر روی داده‌های آموزشی، آموزش دهید. سپس، عملکرد رده‌بندهای مختلف را بر روی داده‌های تست با استفاده از معیارهای accuracy، precision، recall و f1-score ارزیابی کنید. این معیارها را به ازای هر رده‌بند گزارش کنید. (نیازی به مقایسه‌ی عملکرد رده‌بندهای مختلف در این قسمت نیست.)

۴. به جای تعداد وقوع هر یک از کلمات انتخاب شده (tf)، ترکیبی از tf و idf را به عنوان مقدار ویژگی‌ها در نظر بگیرید و دلیل انتخاب خود را توضیح دهید. سپس، مرحله‌ی ۳ را با توجه به این مقادیر از ویژگی‌ها تکرار کنید. تاثیر انتخاب tf و تاثیر انتخاب ترکیبی از tf و idf بر عملکرد رده‌بندها را بررسی کنید.

۵. هر ویژگی دیگری را که به نظرتان به بهبود عملکرد رده‌بندها می‌تواند کمک کند، به مجموعه‌ی ویژگی‌ها اضافه کنید و دلایل خود را برای انتخاب این ویژگی‌ها شرح دهید. سپس، مرحله‌ی ۳ را انجام دهید. نهایتاً عملکرد رده‌بندهای مختلف را با هم مقایسه و تحلیل کنید.

بخش ۲ – Topic Modeling

در این تمرین هدف پیاده‌سازی مدلسازی موضوعی PLSA با به‌کارگیری موضوع زمینه[۱] می‌باشد. مجموعه اسناد $D = \{d_1, d_2, \ldots, d_N\}$ به صورت ترکیبی از K موضوع $\theta_1, \theta_2, \ldots, \theta_K$ و با استفاده از تابع احتمال لگاریتمی زیر مدل می‌شوند:

$$\log p(D \mid \Theta, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|d_i|} \log \left\{ \sum_{k=1}^{K} p(z_{i,j} = k \mid \pi_i) p(d_{i,j} = w \mid \theta_k) \right\}$$

با به‌کارگیری موضوع زمینه‌ی ثابت، تابع احتمال به صورت زیر تغییر می‌کند که در آن، کلمه با احتمال λ از موضوع زمینه P(w|D) و با احتمال $1 - \lambda$ از ترکیب K موضوع تولید می‌شود:

$$\log p(D \mid \Theta, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|d_i|} \log \left\{ \lambda p(d_{i,j} = w \mid D) + (1 - \lambda) \sum_{k=1}^{K} p(z_{i,j} = k \mid \pi_i) p(d_{i,j} = w \mid \theta_k) \right\}$$

الف) در فرمول دوم، P(w|D) مدل زبانی زمینه است که در طول یادگیری پارامترها ثابت است. تخمین احتمال بیشینه[۲] برای این مدل زبانی را بیان کنید.

ب) کد پیاده‌سازی الگوریتم PLSA فرمول اول (بدون در نظر گرفتن موضوع زمینه) به زبان پایتون در اختیار شما قرار گرفته است. کد را برای دادگان با پارامترهای پیش‌فرض اجرا کنید و مجموعه‌ی ۱۰ کلمه‌ی برتر هر ۱۰ موضوع را گزارش نمایید.

ج) کد را به گونه‌ای تغییر دهید که PLSA با استفاده از موضوع زمینه (فرمول دوم) مدل شود. به صورت مختصر در گزارش، قسمت‌های تغییریافته در کد را توضیح دهید.

کد تغییریافته را برای مجموعه دادگان و $\lambda = 0.9$ اجرا کنید. (مانند قسمت قبل، تعداد موضوع‌ها K=10 باشد و برای هر موضوع ۱۰ کلمه در نظر بگیرید). کلمات هر موضوع را گزارش نمایید.

د) قسمت ج را با $\lambda = 0.3$ اجرا کنید. کلمات هر موضوع را گزارش نمایید.

---

[۱] Background topic

[۲] Maximum Likelihood

چ) تاثیر اندازه λ روی موضوعاتی که توسط مدل پیدا می‌شوند، چگونه است؟ (λ بسیار بزرگ و یا بسیار کوچک باشد و یا مانند قسمت ب $\lambda = 0$ باشد) توضیح دهید که چرا این نتایج را مشاهده کردید و با توجه به نتایج چه مقداری (حدودی) برای λ پیشنهاد می‌دهید که موضوعات بهتری استخراج شود.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA5_StudendID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را به تفکیک هر بخش شامل شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل‌های زیر در ارتباط باشید:

(بخش ۱) مهسان اکبری mahsan.a.a@gmail.com

(بخش ۲) پریسا بزمی bazmi.parisa@gmail.com

مهلت تحویل بدون جریمه: جمعه ۱۷ دی ماه ۱۴۰۰

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: جمعه ۲۴ دی ماه ۱۴۰۰