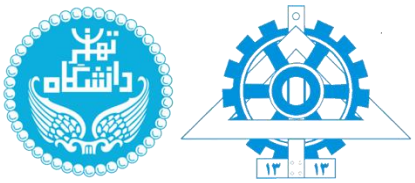
به نام خدا



سامانه‌های یادگیری ماشین توزیع‌شده (پاییز ۱۴۰۳)

تمرین کامپیوتری ۳

موعد تحویل: ۱۴۰۳/۱۰/۷

لطفا پیش از شروع کار بر روی تمارین، به نکات زیر توجه فرمایید.

- برای دسترسی به UI ماشین Spark Master، به آدرس http://raspberrypi-dml0:9000 رفته و از نام کاربری admin و گذرواژه dmlsAdmin استفاده کنید.
- برای دسترسی به UI مربوط به HDFS به آدرس http://raspberrypi-dml0:9870/explore.html بروید.
- لطفا فایلهای دانشجویان دیگر را تغییر ندهید.
- برای دسترسی به HDFS در کد خود، می‌توانید از آدرس http://raspberrypi-dml0:9000 استفاده کنید.
- برای سوال های ۱ و ۲ میتوانید از کولب یا کامپیوتر شخصیتان استفاده نمایید ولی سوال سوم باید روی کلاستر درس (بردهای رزبری پای) انجام شود.
- برای پیاده سازی از زبان پایتون و کتابخانه‌ی PySpark استفاده نمایید. برای استفاده از مدل‌ها یا توابع یادگیری ماشین از کتابخانه‌ی ml به جای mllib استفاده نمایید.
- قبل از شروع تمرین بهتر است ویدیو آپلود شده در سامانه درس را مشاهده نمایید.
- سوالات خود را در گروه تلگرام درس مطرح نمایید. به هیچ وجه کد یا پاسخ سوالات را در گروه به اشتراک نگذارید.
- برای تمامی سوالات در فایل گزارش کدها را نیز توضیح دهید.
- می‌توانید از طریق آدرس ایمیل alisalemmi.1379@gmail.com یا از طریق تلگرام با من در ارتباط باشید.

سوال اول (۳۰ نمره): در این تمرین متن شاهنامه فردوسی به شما داده شده است. به کمک Spark RDD به سوالات زیر پاسخ دهید.

الف) تعداد کل ابیات و کلمات (بدون در نظر گرفتن تکرار) را به دست آورید. (۱۰ نمره)

ب) ۱۰ قافیه پر تکرار را به همراه تعداد تکرار آن به دست آورید. (۵ نمره)

ج) یکی از ساده‌ترین روش‌های مدل‌سازی زبان n-gramها هستند. هر n-gram یک دنباله از n کلمه متوالی است. با محاسبه تعداد n-gramها در یک متن می‌توان یک مدل احتمالاتی از زبان ایجاد کرد. 3-gram های شاهنامه را محاسبه کنید. چند 3-gram در شاهنامه وجود دارد؟ ۱۰ 3-gram پر تکرار را نمایش دهید. (۱۵ نمره)

سوال دوم (۱۰+۴۰ نمره): word2vec روشی برای تبدیل کلمات به بردارهای معنایی است. word2vec به دو روش skip gram و CBOW قابل پیاده سازی است.

الف) الگوریتم word2vec را در حالت skip gram در Spark پیاده سازی کرده و با داده‌های شاهنامه آموزش دهید. در صورت استفاده از پیاده سازی‌های آماده word2vec نمره‌ای به این بخش تعلق نمی‌گیرد. (۲۰ نمره)

ب) embedding کلمات رستم، سهراب، اسفندیار، رخش و زال را به دست آورید. (۱۰ نمره)

ج) میزان شباهت دو به دو کلمات قسمت الف را به دست آورده و در قالب نمودار heat map نمایش دهید. (۱۰ نمره)

د) بردارهای embedding قسمت الف را با روش PCA به دو بعد کاهش داده و در نمودار دو بعدی نمایش دهید. (۱۰ نمره اضافه)

سوال سوم (۳۰ نمره): در این سوال اطلاعات ۲۰۰ مشتری یک فروشگاه در فایل customers.csv به شما داده شده است. هدف از این سوال آشنایی با کتابخانه یادگیری ماشین اسپارک است.

الف) در hdfs پوشه‌ای با نام شماره دانشجویی خود ایجاد کرده و مجموعه داده را در آن آپلود کنید (۵ نمره)

ب) اطلاعات آماری شامل مینیمم، ماکسیمم، میانگین و واریانس مربوط به درآمد سالانه مشتریان را محاسبه کنید. (۵ نمره)

ج) با روش k-means مشتریان را بر اساس درآمد سالانه و spending score طبقه بندی کنید. (۱۵ نمره)

د) نتیجه را در قالب یک نمودار رسم کرده و هر دسته را با یک رنگ نمایش دهید. (۵ نمره)

نحوه‌ی تحویل پروژه:

فایل‌ها را به صورت زیر نام گذاری کرده و در آخر همه را در یک فایل zip به نام شماره دانشجوی خود در سامانه ارسال کنید.

سوال	نام فایل
گزارش	report.pdf
۱	ngram.ipynb
۲	Word2vec.ipynb
۳	k-means.py