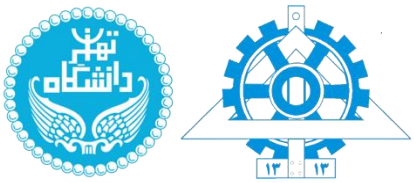
به نام خدا



سامانه‌های یادگیری ماشین توزیع‌شده (پاییز ۱۴۰۲)

تمرین کامپیوتری ۳

موعد تحویل: ۱۴۰۲/۱۰/۸

لطفا پیش از شروع کار بر روی تمرین، به نکات زیر توجه نمایید:

- برای دسترسی به UI ماشین Master Spark ، میتوانید به آدرس http://respberrypi-dml0:8080 رفته و از نام کاربری admin و گذرواژه dmlsAdmin استفاده کنید.
- برای دسترسی به UI مربوط به HDFS میتوانید 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 hdfs://raspberrypi-dml0:9000 استفاده کنید. برای مثال اگر بخواهید خروجی را در فایل tmp واقع در دایرکتوری rostami بنویسید، باید به کمک تابع مربوطه درpyspark ، آن را در آدرس hdfs://raspberrypi-dml0:9000/rostami/tmp بنویسید.
- برای سوال‌های ۱ و ۲ می‌توانید از کولب یا کامپیوتر شخصیتان استفاده نمایید ولی سوال سوم باید روی کلاستر درس (بردهای رزبری پای) انجام شود.
- برای پیاده‌سازی 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 و نوتبوک هندزآن 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.

۱. (۳۰ نمره) هدف از این تمرین آشنایی با Spark RDD است. در این تمرین ۱۲ خبر از اخبار مجموعه داده‌ی رویترز در فایل news.txt آورده شده است. مراحل زیر را به ترتیب بر روی داده‌ها ایجاد کنید:

الف) (۵ نمره) داده‌ها را خوانده و تعداد کل اخبار را بنویسید. سپس تعداد کل کلمات (بدون توجه به تکرار آن‌ها) را به دست آورید و ۱۰ کلمه‌ی اول را چاپ کنید.

ب) (۵ نمره) حروف هر کلمه را به حروف کوچک تبدیل کرده و تعداد تکرار هر کلمه را محاسبه نمایید. سپس آن‌ها بر اساس تعداد تکرارشان از بزرگ به کوچک مرتب نمایید و ۱۰ کلمه‌ی پر تکرار را چاپ کنید.

ج) (۵ نمره) به کمک ماژول punctuation از کتابخانه‌ی string می‌توانید به علائم نشانه‌گذاری انگلیسی دسترسی پیدا کنید. تعدادی از کلماتی که در RDD بخش (ب) وجود دارند، علائم نشانه‌گذاری هستند. این کلمات را از RDD بخش (ب) حذف کرده و دوباره ۱۰ کلمه‌ی پرتکرار را چاپ نمایید.

د) (۵ نمره) تعداد کلمات با حرف اول یکسان را به دست آورده و ۵ حرفی که بیشترین کلمات با آن‌ها شروع شده است را چاپ نمایید.

(۱۰ نمره) کدهای نوشته شده برای هر بخش را در گزارش توضیح دهید.

۲. (۳۵ نمره) الگوریتم tf-idf یکی از الگوریتم‌های حوزه‌ی بازیابی اطلاعات است. در این الگوریتم عبارت tf (term frequency) نشان‌دهنده‌ی تعداد تکرار هر کلمه در هر سند و عبارت idf (inverse document frequency) نشان‌دهنده‌ی تعداد اسنادی است که آن کلمه را دارند. به کمک ضرب این دو مقدار می‌توان بردارهای tf-idf را برای هر کلمه محاسبه کرد. در این تمرین قصد پیاده‌سازی این الگوریتم به کمک Spark RDD را داریم.

در فایل news.txt هر خط را یک سند در نظر بگیرید. جزئیات پیاده‌سازی به همراه مثال آورده شده در ویکی‌پدیا را مطالعه کرده و این الگوریتم را به کمک توابع Spark RDD و بدون استفاده از توابع آماده‌ی PySpark پیاده‌سازی نمایید. بردارهای tf-idf می‌توانند در پیدا کردن اسناد مرتبط با هر کلمه مورد استفاده قرار گیرند. به این صورت که در هر بردار، اندیس اعداد بزرگتر نشان‌دهنده‌ی شماره‌ی سند مرتبط‌تر است. برای کلمات japan ، gas و market ۳ سند مرتبط (در صورت وجود) را مشخص نمایید.

(۳۰ + ۵ نمره) جزئیات کد نوشته برای این بخش را توضیح داده و نتایج اسناد به دست آمده برای کلمات را بررسی نمایید.

۳. هدف از این سوال آشنایی بیشتر با کتابخانه‌ی یادگیری ماشین Spark است. در این سوال از مجموعه داده‌ی heart.csv استفاده می‌شود. این مجموعه داده شامل اطلاعات مختلف ۳۰۳ بیمار است و برای پیشبینی حمله‌ی قلبی بیماران بر اساس اطلاعات داده شده مورد استفاده قرار می‌گیرد. برای اطلاعات بیشتر در مورد این مجموعه داده، می‌توانید این لینک را مطالعه نمایید.

الف) (۵ نمره) در hdfs پوشه‌ای با نام شماره دانشجوییتان ایجاد کرده و داده‌ی heart.csv را در آن آپلود نمایید.

ب) (۵ نمره) داده‌ی heart.csv را از hdfs خوانده و مینیمم، ماکزیمم، میانگین و واریانس ستون‌های age ، trtbps و chol را بدست آورید.

ج) (۵ نمره) داده‌ها را به دو بخش آموزش و تست به نسبت ۸۵ به ۱۵ به صورت رندوم تقسیم نمایید.

د) (۱۰ نمره) با استفاده از Pipeline مدل‌های Logistic Regression و Random Forest را بر روی مجموعه داده‌ی آموزش، آموزش دهید (دقت کنید که از کتابخانه‌ی pyspark.ml استفاده نمایید).

ه) (۵ نمره) معیارهای Accuracy، Precision، Recall و F1-score برای هر مدل را بر روی مجموعه داده‌ی تست محاسبه نمایید.

(۵ نمره) کدهای نوشته شده در این بخش به همراه نتایج را گزارش نمایید. همچنین مدت زمان اجرای آموزش هر مدل را گزارش دهید.

نحوه تحویل پروژه

فایل‌ها را به صورت زیر نام گذاری کرده و در آخر همه را در یک فایل zip در سامانه ارسال کنید:

۱- گزارش report.pdf

۲- نام گذاری کدها را به صورت زیر انجام دهید:

سوال	بخش	نام فایل
۱	تمام بخش‌ها	Spark_rdd.ipynb (کدهای هر بخش باید مجزا و خروجی هر سلول باید مشخص باشد)
۲	-	Tfidf.ipynb (خروجی سلول‌ها باید مشخص باشد)
۳	تمام بخش‌ها	Logistic_regression.py
	تمام بخش‌ها	Random_forrest.py