به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس

کلان داده و تحلیل داده‌های حجیم

دکتر اسدپور

نیمسال دوم سال تحصیلی ۱۴۰۰-۱۳۹۹

تمرین شماره ۳

Spark

طراح تمرین :

مجتبی بنائی

مهلت تحویل : ۲۰ خردادماه ۱۴۰۰

مقدمه

هدف از این تمرین آشنایی با **Spark** به عنوان یکی از اصلی ترین فریمورک‌های حال حاضر کار با کلان‌داده در جامعه جهانی است که در بسیاری از شرکتها و کاربردها به صورت روزانه مورد استفاده قرار می‌گیرد.

در این تمرین ابتدا با اصول اولیه اسپارک و اجرای دستورات پایه‌ای آن آشنا خواهید شد و سپس با دو کتابخانه جانبی و اصلی آن یعنی **Spark Graph** و **Spark SQL** کار خواهید کرد. کار با بخش پردازش جریان در اسپارک را در پروژه نهایی این درس انجام خواهید داد.

این تمرین را با محیط **Colab** گوگل انجام دهید. قبل از شروع کار، بهتر است آموزش سریع و کاربردی راه‌اندازی اسپارک در کولب که در آدرس زیر قرار گرفته است را انجام دهید تا برای تمرین اصلی، آمادگی لازم را کسب کنید :

https://jacobcelestine.com/knowledge_repo/colab_and_pyspark/

دیتاست‌های مورد نیاز هر تمرین هم همراه با تمرین آپلود شده است .

برای هر سوال، یک کتابچه پایتون (Python Notebooks) ایجاد کنید و در انتهای کار، کتابچه ها را دانلود کرده ، زیپ نموده و همراه گزارش توضیحات تمرین به صورت تک‌نفره ، آپلود نمایید.

گام اول - دستورات پایه

این تمرین را با دستورات پایه اسپارک یعنی دستورات کار با RDD ها انجام دهید و از دیتافریم و اسپارک‌اسکیوال استفاده نکنید. برای آشنایی با این دستورات پایه، می‌توانید از لینک زیر استفاده کنید :

http://yun.ir/jlilx9

بخش اول

تعداد لغات فایل **Input.txt** را شمارش کرده و نمایش دهید. همچنین گزارش کنید که هر کلمه چند بار تکرار شده است و خروجی را در یک فایل **txt**. ذخیره کنید. در این گام تنها علائم نقطه گذاری (علامت تعجب، سوال، نقطه و ...) را حذف کنید و پیش‌پردازش دیگری لازم نیست.

بخش دوم

تعداد تمامی کلماتی که با حرف (**M**) آغاز می‌شوند را بیابید. (مستقل از کوچک و بزرگ بودن **M**)

بخش سوم

در این بخش تعداد لغات 5 حرفی موجود درفایل **Input.txt** را یافته، لغاتی که با حروف صدادار شروع می‌شوند را از خروجی حذف کنید و نتیجه نهایی را به صورت مرتب نمایش دهید.

بخش چهارم

به کمک مراحل قبلی، ایست واژه‌ها(**stop words**) را بیابید. **کلمه‌ای را ایست واژه درنظر بگیرید که جزء ده درصد کلمات پرتکرار این فایل قرار بگیرد.** سپس تابعی بنویسید که یک خط را گرفته، تمام حروف غیر الفبایی و ایست‌واژه‌های آنرا حذف کند. این تابع را روی تمام خطوط اعمال کرده، نتیجه را در یک فایل، ذخیره کنید.

بخش پنجم

تعداد دو کلمه‌ای هایی که بیشتر از یک بار در فایل اصلی (**input.txt**) کنار هم آمده‌اند را به ترتیب فرکانس، یافته و نمایش دهید. منظور از دوکلمه‌ای(**bigram**)، دو لغتی هستند که پشت سر هم به کار رفته‌اند.

گام دوم - بررسی یک فایل لاگ وب سرور

فایل لاگ پیوست این تمرین با نام **Log** که مربوط به در خواست های **HTTP** است، برای این گام در نظر گرفته شده است. بااستفاده از این فایل به سوالات زیر پاسخ دهید (برای این بخش می‌توانید از دستورات پایه اسپارک، **Spark SQL** و یا Spark Dataframes استفاده کنید - هر کدام از این سه روش برای انجام این گام، مجاز است):

بخش اول

چند**Host** یکتا در این لاگ فایل وجود دارد؟

بخش دوم

متوسط تعداد درخواست های روزانه برای هر میزبان منحصر به فرد (آی‌پی یا نام دامنه) چقدر است؟ ابتدا متوسط تعداد درخواست‌های هر دامنه در هر روز را به دست آورید و سپس، متوسط نهایی را برای هر دامنه یا آی‌پی، تعیین کنید .

بخش سوم

تعداد فایل‌های گیف درخواست شده در این فایل لاگ چقدر است؟

بخش چهارم

دامنه‌های پرتقاضا (بیش از ۳ بار) را یافته، آنها را به صورت مرتب شده نمایش دهید. آی پی ها را جزء این دامنه‌ها در نظر نگیرید. سپس دامنه پرتقاضا به ازای هر روز را پیدا کنید (دامنه‌ای با بیشترین تعداد درخواست در یک روز).

بخش پنجم

خطاهای **HTTP** (غیر ازکد ۲۰۰، بقیه را همه خطا در نظر بگیرید.) را یافته، تعداد تکرار آنها در یک نمودار ستونی نمایش دهید.

گام سوم - کار با دیتافریم‌ها / Spark SQL

در این تمرین، از داده‌های بورس دانلود شده در تمرین هدوپ استفاده خواهیم کرد. داده‌های روزانه بورس برای یک بازه دوماهه که بتوان حداقل سی روز متمایز را دانلود کرده و در یک پوشه در گوگل درایو خود آپلود کنید. در محیط کولب، می‌توانید به راحتی به این پوشه دسترسی داشته باشید. ابتدا به کمک اسپارک، فایلها را باز کرده و ستون روز و ماه و سال را به آنها اضافه کنید . (یا یک ستون تاریخ - شمسی یا میلادی)

تمام سوالات بخش هدوپ را در این قسمت با اسپارک و با دو رهیافت مختلف (**Spark SQL** / **Spark Dataframe**) انجام خواهیم داد.

- گرانترین و ارزانترین نمادهای بورسی کدام ها هستند (ده نماد) ؟ روز آخر را ملاک بگیرید یعنی جدیدترین فایل موجود در پوشه فایلهای بورس. برای راحتی کار، پردازش را به صورت خاص بر روی همین یک فایل انجام دهید.
- چه نمادی بیشترین حجم خرید را در شش ماه گذشته داشته است ؟
- در هر ماه، چه نمادهایی بیشترین میزان افزایش قیمت را تجربه کرده‌اند(۱۰ نماد) ؟ یعنی اگر نمادها را بر اساس قیمت در هر ماه مرتب کنیم و سپس اختلاف بین ابتدا و انتهای لیست را به دست آوریم، ده نمادی که بالاترین عدد را به خود اختصاص داده‌اند در هر ماه، چه نمادهایی هستند .
- چه نمادهایی بیشترین ریزش قیمت را در شش ماه اخیر داشته اند ؟ (البته برخی نمادها بعد از افزایش سرمایه و افزایش تعداد سهام، افت قیمت پیدا می کنند که به آن ریزش قیمت نمیگوییم و فعلا مد نظر ما نیست)
- چه نمادی بیشترین میزان بسته بودن را داشته است؟ به دلایل مختلف مانند برگزاری مجمع عمومی یا عادی و یا افشای اطلاعات با اهمیت، ممکن است یک سهم چندین روز بسته باشد . احتمالا نمادهایی که بسته بوده‌اند در فایل اکسل، حجم معامله صفر دارند و یا اصلا درج نشده اند. باید بررسی کنید . مثلا وهور در دهه دوم اسفند بسته بوده است و در فایل‌های اکسل این دهه این موضوع را چک کنید .

برای انجام این تمرین از دو روش استفاده کنید یعنی برای هر بخش، خروجی مورد نظر را با هر کدام از دو روش زیر به صورت جداگانه به دست آورید :

- **Spark DataFrames** – با توابع دیتافریم (**DataFrame Operations such as min,avg,...**)
- **Spark SQL** - با دستورات **SQL** (**spark.sql**)

برای آشنایی دقیق‌تر با دیتافریم‌ها، علاوه بر مستندات رسمی خود بنیاد آپاچی، از لینک زیر هم می توانید به عنوان یک آموزش سریع و کاربردی استفاده کنید :

https://towardsdatascience.com/the-most-complete-guide-to-pyspark-dataframes-2702c343b2e8

گام چهارم - Spark GraphX

فایل پیوست **edgs.txt** یال ها و فایل پیوست **vertex.txt** درجه های یک گراف هستند. گراف مورد نظر ما از مقالات ویکی پدیا استخراج شده اند. هر گره یک مقاله ویکی پدیا و یال از مقاله A به مقاله B نشان دهنده این است که مقاله A به مقاله B ارجاع داده است.

- ***نکته*** : می‌توانید برای کار با گراف در اسپارک می‌توانید از **GraphFrames**[1] استفاده کنید.

بخش اول:

با استفاده از فایل یال ها و گره ها، این گراف را ایجاد کنید.

بخش دوم:

بیشترین درجه ورودی در این گراف چقدر است؟ بیشترین درجه خروجی (مقاله‌ای که احتمالا Survey بوده و شامل لینک زیادی به سایر مقالات است.) چند است ؟

بخش سوم:

- سایز هرکدام از **ConnectedComponent** ها چقدر است؟

بخش چهارم:

ده تا از مقالات برتر را بیابید (مقالاتی که بیشترین درجه ورودی را داشته‌اند).

بخش پنجم (نمره اضافی)

آیا می‌توانید گراف فوق را به صورت بصری نمایش دهید ؟

1 graphframes.github.io/graphframes/