به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس کلان داده

تمرین شماره ۳

Spark

خرداد ماه ۱۳۹۹

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.

با توجه به مشکلاتی که در نصب و راه‌اندازی هدوپ در دو تمرین گذشته با آن مواجه بوده‌اید و با هدف کار با محیط‌های آنلاین پردازش داده، توصیه می‌شود این تمرین را به کمک محیط رایانش ابری شرکت Databricks[1] به آدرس https://community.cloud.databricks.com انجام دهید.

توصیه می‌کنیم قبل از شروع کار با این محیط، این آموزش ساده و کاربردی اسپارک را که با تمرکز بر این محیط توسعه آنلاین نوشته شده است را مطالعه کرده و دستورات آنرا به عنوان دست‌گرمی انجام دهید :

https://bit.ly/SparkUT

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 تک‌فره ، آپلود نمایید.

[1] https://databricks.com/

سوال اول - دستورات پایه

بخش اول

در این قسمت با استفاده از تابع نگاشت-کاهش(map_reduce) تعداد لغات فایلInput.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

حال تمام در این گام با استفاده از تابع نگاشت_کاهش(map_reduce) تعداد تمامی کلماتی که با حرف (M) آغاز می شوند را بیابید کنید. (مستقل از کوچک و بزرگ بودن M)

بخش سوم

در این بخش نیز همانند دو گام قبلی، با استفاده از تابع نگاشت_کاهش(map_reduce) تعداد لغات 5 حرفی موجود درفایل words.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.

سوال دوم - بررسی یک فایل لاگ وب سرور

فایل لاگ پیوست این تمرین با نام"Log "که مربوط به در خواست های HTTP است. بااستفاده از این فایل به سوال زیر پاسخ دهید (برای این بخش از دستورات پایه اسپارک استفاده کنید):

بخش اول

چندHost 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.

سوال سوم - کار با دیتافریم‌ها / Spark SQL

با توجه به دیتاست stock.csv به سوالات زیر پاسخ دهید . این دیتاست، داده های بورس مربوط به یکی از کمپانی های بزرگ از سال ۲۰۱۲ تا ۲۰۱۷ می باشد.

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 :

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

بخش اول:

یک ستون اطلاعات جدید با ستونی به نام HV ایجاد کنید که این نسبت بالاترین قیمت بر حجم سهام معامله شده برای یک روز است.

بخش دوم:

پیک بالاترین قیمت، برای چه روزی بوده است؟

بخش سوم:

میانگین ستون، Close، چه مقدار است؟

بخش چهارم:

مقدار ماکزیمم و مینیمم ستون Volum را مشخص کنید.

بخش پنجم:

چند روز ستون Close کمتر از 60 دلار بوده است؟

بخش ششم:

Pearson correlation بین ستون های High و Volum چقدر است؟

بخش هفتم:

ماکزیمم ستونHigh در هر سال چقدر است؟

سوال 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 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/