دانشکده مهندسی برق

تحلیل داده‌های حجیم

تمرین سری اول

استاد: دکتر ایمان غلامپور

دستیار آموزشی: بهنام رئوفی

بخش تئوری و تحقیق:

سوال اول) فرض کنید می‌خواهیم مسئله join مقابل را بررسی کنیم: $R(A, B) \bowtie S(B, C) \bowtie T(C, D) \bowtie U(D, E)$

در این مسئله S، R، U و T هر کدام، یک جدول از دیتابیس با اندازه‌های s، r، u و t هستند. احتمال آنکه S و T در C، U و T در D، S و R در B در توافق باشند برابر با p است.

- با استفاده از مدل Map-Reduce و رویکرد Single Step، الگوریتمی برای حل این مسئله join طراحی کنید.

- حال این الگوریتم را از نظر Communication cost، Replication rate، Reducer size و تعداد node های Map و Reduce بررسی کنید.

سوال دوم) توضیح دهید که RDD و Dataframe چه تفاوت‌هایی با هم دارند و کدامیک برای کار با داده‌های ساختار یافته مناسب‌تر هستند.

سوال سوم) با چه روش‌هایی می‌توان کارایی Pyspark را افزایش داد؟ چند نمونه از تکنیک‌های بهینه‌سازی را ذکر کنید.

بخش عملی:

هدف این بخش از تمرین، آشنایی با مقدمات کار با Rdd ها در Spark می‌باشد. در این بخش شما یک دیتاست که شامل مقالات Arxiv می‌باشد را از لینک زیر دانلود کرده و بخش‌های مشخص شده در فایل notebook همراه با تمرین را تکمیل خواهید کرد.

لینک دانلود دیتاست

بخش اول

در این بخش پس از دانلود دیتاست و قرار دادن فایل json موجود در دایرکتوری مناسب، کافیست کدهای ابتدای این بخش را اجرا کرده تا دیتاست در قالب Rdd آماده گردد.

بخش دوم

در این بخش نیاز است تا متون موجود در قسمت title و abstract را در صورت نیاز تمیز کنید. برای این منظور می‌توانید اقدام به حذف علائم ریاضی، کاراکترهای بی معنا در متن، حذف stopwords ها و... نمایید.

بخش سوم

در قسمت اول این بخش، باید تعداد مقالات موجود در هر دسته بندی (برای مثال hep-ph یا math.co) را محاسبه کنید. در قسمت دوم، دسته‌ای را که بیشترین مقالات را دارد، شناسایی کنید. سپس در قسمت سوم، توزیع تعداد نویسندگان در هر مقاله را تحلیل کنید. قسمت چهارم از شما می‌خواهد که مقالاتی را که بیش از سه نویسنده دارند فیلتر کرده و عنوان و نویسندگان آن‌ها را فهرست کنید. در قسمت پنجم، تعداد مقالات ثبت شده در هر سال را رسم کنید. سپس در قسمت ششم، ۲۰ کلمه‌ی پرتکرار در بخش abstract مقالات را استخراج و نمایش دهید. در قسمت هفتم، کلمات پرتکرار شناسایی‌شده در بخش قبلی را با استفاده از Word Cloud به‌صورت تصویری نمایش داده تا تحلیل بهتری از این کلمات داشته باشید.

بخش چهارم

در این بخش ابتدا مقالاتی که در abstract آن‌ها کلمه algorithm آمده است را پیدا کنید. سپس تعداد کلمات موجود در abstract این مقالات را شمرده و در نهایت، آن‌ها را بر اساس تعداد کلمات به صورت نزولی مرتب کنید. پنج مقاله با بیشترین تعداد کلمات در abstract را به عنوان نتیجه نهایی نمایش دهید.