کاوش داده‌گان انبوه

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده علوم و فنون
ترم اول سال تحصیلی – پاییز ۱۳۹۹

دستیاران آموزشی
مهدی صادقی (mahdisadeghi@ut.ac.ir)
شیوا پارسا راد (shiva.parsarad@ut.ac.ir)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

مجموعه تمرین اول : MapReduceیی فکر کنید!

آخرین مهلت تحویل این تمرین روز چهارشنبه، ۵م آذر ساعت ۲۳:۵۹ می‌باشد.

در این مجموعه تمرین، یک الگوریتم MapReduce برای اجرای یک عملیات جبر رابطه‌ای طراحی می‌کنید. همچنین از سیستم Hadoop که یک سیستم گردش کاری مبتنی بر MapReduce است برای پیاده سازی یک مثال ساده استفاده خواهید کرد. در قسمت پژوهش این تمرین با چند چارچوب دیگر برای کاربردهای خاص داده انبوه آشنا خواهید شد.

۱- (طراحی و پیاده سازی الگوریتم) فرض کنید دو رابطه $R(A,B)$ و $S(C,D)$ را داریم. عملیات پیوند زیر تمام تاپل‌های (a,b,c,d) را برمی‌گرداند بطوریکه (a,b) در R و (c,d) در S قرار داشته باشد با این شرط که $b < c$ است:

$$R(A,B) \bowtie_{B<C} S(C,D)$$

الف) با استفاده از ساختار MapRedeuce، الگوریتمی طراحی کنید که این عملیات را انجام دهد. حتما از شکل‌های گرافیکی برای توصیف الگوریتم خود استفاده کنید.

ب) (امتیاز اضافی) با استفاده از یک دیتاست کوچک از اعداد صحیح، این الگوریتم را به صورت MapReduce پیاده سازی کنید.

۲- (طراحی و پیاده سازی الگوریتم)

می‌خواهیم برای پیاده سازی یک سیستم توصیه‌گر[1] آماده شویم. فعلا اما یک سیستم خیلی ساده پیاده‌سازی خواهیم کرد.

داده مورد نیاز این تمرین در یک دیتاست عمومی به نام **MovieLens 100K** قرار دارد. این دیتاست شامل ۱۰۰هزار امتیاز (۱ تا ۵) از ۹۴۳ کاربر به ۱۶۸۲ فیلم است. این دیتاست شامل کاربرانی است که حداقل به ۲۰ فیلم امتیاز داده‌اند.

الف) می‌خواهیم شباهت کاربران **(شناسه ۱ تا ۵۰)** را بر اساس ژانر[2] فیلمهایی که به آن امتیاز داده‌اند پیدا کنیم. برای اینکار ابتدا امتیاز کاربران به فیلم‌ها و ژانر آنها را بررسی می کنیم. اگر متوسط امتیاز کاربر در هر ژانر برابر یا بیشتر از ۳ بود کاربر به آن ژانر علاقمند است و بهتر است از این ژانر فیلمی به او پیشنهاد نشود.

ب) حال باید کابران مشابه را بر اساس این ژانرها پیدا کنیم **(برای کاهش تعداد مقایسات هر کاربر با فقط یک کاربر دیگر بصورت تصادفی مقایسه می‌شود)**. برای اینکار از شباهت جاکارد[3] استفاده می‌کنیم. کاربرانی را مشابه فرض می‌کنیم که شباهت آنها برابر یا بزرگتر از 0.5 باشد.

برای پیاده سازی هر دو بخش فوق از MapReduce و کتابخانه MrJob استفاده کنید..

۳- (پژوهش) در یک یا دو صفحه Spark و Hadoop را با هم مقایسه کنید. همچنین بطور مختصر Apache Storm و Apache Kafka را معرفی کنید. با درج رفرنس در کنار مطالب خود، آنها را به ما نیز معرفی کنید.

دیدن ویدیوی زیر خالی از لطف نیست:

Going from Hadoop to Spark: A Case Study (San Francisco Bay ACM)

[1] Recommender System
[2] Genre
[3] Jaccard

نکات تکمیلی

- از زبان برنامه سازی پایتون استفاده کنید.
- کدها با نامگذاری مناسب در پوشه SOURCES قرار گیرد.
- گزارش کامل تمرین را با ساختار مناسب، بدون ابهام و به زبان ساده نوشته و با فرمت PDF در پوشه دیگری قرار دهید. برای نگارش گزارش‌ها نیز امتیاز مثبت و منفی در نظر گرفته شده است، پس برای نگارش گزارش خود وقت مناسبی اختصاص دهید.
- پاسخ تمرین خود را به ایمیل barazandeh.iman@ut.ac.ir ارسال نمایید. هر دو پوشه را در یک فایل ZIP با نام MMDS_STID_FullName_HW# (مثل MMDS_830123451_ImanBarazandeh_HW1) قرار داده و در ایمیلی با همین عنوان نیز ارسال کنید.
- در صورت مشخص شدن هر نوع تقلب انفرادی یا مشترک نمره تمرین صفر در نظر گرفته می‌شود.
- به ازای هر روز تاخیر ۲۵٪ از امتیاز تمرین از دست خواهید داد.
- هر گونه ابهام را با دستیاران آموزشی مطرح کنید.