دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس آنالیز داده ها ی حجیم

تمرین سری اول

استاد: دکتر ایمان غلامپور

قوانین تحویل:

- پاسخ به تمرینات این درس می بایست حتماً تایپ شده باشند، لذا گزارش های دست نویس تصحیح نخواهند شد.
- بخش زیادی از نمره تمرینات به گزارش و نتیجه گیری های شما اختصاص دارد، لذا در نوشتن گزارش بخش های مختلف سوالات دقت کافی را داشته باشید و تمامی نتایج را تحلیل کرده و با حوصله آن ها را ذکر کنید، سعی کنید در تحلیل های خود از نمودارها و هر visualization ابتکاری دیگر استفاده کنید، گزارش هایی که صرفاً شامل کد باشند تنها نمره programming assignment را خواهند گرفت.
- پاسخ های قسمت های عملی می بایست حتماً در فرمت ipynb. باشند، بنابراین میبایست تمامی بخش های عملی به صورت یک jupyter notebook تحویل داده شوند.
- تمام فایل های خود را در قالب یک فایل زیپ به فرمت HWn_studentNumber_Family تحویل دهید، n شماره تمرین می باشد.

قوانین تاخیر:

در کل میتوانید برای تمامی تمرینات **حداکثر 12** روز تاخیر داشته باشید و به ازای هر تمرین بیشتر از 4 روز تاخیر، مشمول کسری نمره می باشد، بطوری که بعد از روز 4ام، به ازای هر روز اضافی، 20 درصد از نمره تمرین را از دست خواهید داد.

از آنجا که تمام سیاست به کار گرفته شده در این درس کار با دیتاهای واقعی و یادگیری عملی در دنیای واقعی در کنار مطالب تئوری ست، لذا وقت خود را با کپی کردن از یکدیگر هدر ندهید، در صورتی که در گزارش ها و کد ها، شباهت های غیرعادی دیده شود، بدون تذکر، 100 نمره منفی برای طرفین در نظر گرفته می شود، لذا می توانید صرفاً از یکدیگر مشورت بگیرید یا سوالات خود را به صورت حضوری (جنب مسجد، پژوهشکده الکترونیک، طبقه سوم، اتاق 306) یا به صورت ایمیل به آدرس زیر بپرسید.

alishojaei7697@gmail.com

سوال اول)

الف) فرض کنید که قصد داریم مسئله join روبرو را بررسی کنیم: $R(A,B) \bowtie S(B,C) \bowtie T(C,D) \bowtie U(D,E)$

که در آن R، S، T و U هر کدام یک جدول از دیتابیس با اندازه های r، s، t و u هستند. احتمال آنکه R و S در B، S و T در C و T و U در D در توافق باشند برابر با p است. در ابتدا برای حل مسئله بالا، با استفاده از مدل Map-Reduce الگوریتمی را طراحی کنید. این الگوریتم را از نظر هزینه محاسبات، کران های replication rate و reducer size و تعداد nodeهای Map و Reduce بررسی کنید.

ب) برروی طراحی الگوریتم Single Step در حل مسئله بالا فکر کنید و آنرا ارائه دهید و مزیت آن را با استفاده از تحلیل پارامتری مشابهی که در قسمت قبل انجام دادید اثبات کنید.

سوال دوم)

الف) فایل MOVIE_IMDB.csv را با استفاده از کتابخانه Pandas بارگذاری کنید و آنرا در یک دیتافریم با نام df بریزید.

ب) تعداد مقادیر Null را در تمامی سطرها و ستون ها مشخص کنید، سپس درصد مقادیر Null در هر ستون را بدست بیاورید (تا دو رقم اعشار round شود).

پ) ما به تعدادی از ستون های این دیتافریم نیازی نداریم، لذا ستون های زیر را از دیتافریم حذف کنید:

- color
- director_facebook_likes
- actor1_facebook_likes
- actor2_facebook_likes
- actor3_facebook_likes
- actor2_name
- cast_total_facebook_likes
- actor3_name
- duration
- facenumber_in_poster
- content_rating
- country
- movie_imdb_link
- aspect_ratio
- plot_keywords

حال قسمت ب را دوباره انجام دهید، متوجه خواهید شد که بعضی از ستون ها درصد بیشتری از مقادیر Null را دارا هستند(بیشتر از 5 درصد)، برای این ستون ها، سطرهایی که مقادیر Null دارند را حذف کنید و بار دیگر قسمت ب را انجام دهید، در گام بعدی از ستون language، تمامی سطرهایی که مقدار NaN دارند را با English پر کنید، در آخر درصد مقادیر Null در هر ستون را بار دیگر گزارش کنید. (علت وجود مقادیر NaN در این دیتافریم چه میتواند باشد؟)

ت) مقادیر duplicate را از دیتافریم حذف کنید، سپس ستون جدیدی به اسم profit بسازید که از اختلاف دو ستون gross و budget بدست می آید، نمودار profit را برحسب budget رسم کنید و آن را تحلیل کنید، سپس 10 فیلم پرسود این دیتافریم را گزارش کنید.

ث) چه ژانرهایی در این دیتافریم وجود دارند؟ در هر ژانر چند فیلم وجود دارد؟ میانگین درآمد در هر ژانر چقدر بوده است؟

ج) ژانر تمام فیلم هایی که بین 2007 تا 2015 تولید شده اند و امتیاز آنها بین 7.5 تا 8.5 هست را بدست بیاورید.

چ) آیا نظر مردم نسبت به یک فیلم در میزان فروش آن موثر است؟ با استفاد از کد و نمودار استدلال کنید.

سوال سوم)

هدف از این سوال، آشنایی کار با مقدمات RDDها در spark می باشد. در این بخش شما یک دامپ 1 گیگابایتی از ویکی پدیای فارسی را بررسی و تحلیل خواهید کرد. Note book ای که در ضمیمه در اختیار شما قرار داده شده است را به دقت کامل کنید و گزارش هر بخش را به صورت کامل ارائه دهید (تا حد ممکن از آوردن کدها در گزارش خودداری کنید مگر جایی که واقعا نیاز باشد).

نکته: برای حل این بخش درصورتی که ماشین مناسب در اختیار شما نیست، می توانید کدها را در Colab و همچنین سایت databricks ران کنید. (با توجه به اینکه در colab به دلیل مواردی مانند قطعی نت، بستن مرورگر و ... احتمالاً زیاد به مشکل colab session timeout برمیخورید، توصیه می شود از databricks برای حل تمارین این درس استفاده کنید).

موفق باشید