تمرین دوم درس تحلیل داده های حجیم MDA2021

۱ - آشنایی با Pandas وnumpy

الف) فایل IMDB-Movie-Data.csv با استفاده از کتابخانه ی Pandas بارگذاری کنید و آن در یک دیتا فریم به نام df قرار دهید.

ب) مشخص کنید چه ژانرهای مختلفی در این دیتاست وجود دارد، در هر ژانر چند فیلم وجود دارد، و در هر ژانر میانگین درآمد چه قدر بوده است.

ج) بین فیلمهای سال ۲۰۰۶ تا ۲۰۱۶ فیلمهایی که حداقل دو بازیگر مشترک دارند را دریک دسته قرار دهید و این دسته ها را ذکر کنید. مثلا اگر سه فیلم در دو بازیگر مشترک بودند همگی در یک دسته قرار میگیرند.

د) ژانر همه ی فیلمهای بالای ۱۱۰ دقیقه که امتیازی بین ۷.۵ تا ۸.۵ دارند را به دست آورید.

ه) به ازای هر بازیگر، تعداد فیلمهایی را که در آن بازی کرده را به دست بیاورید. سپس بازیگران را ابتدا بر اساس تعداد فیلم به صورت نزولی مرتب کنید.

و) آیا نظر مردم نسبت به یک فیلم در میزان فروش آن موثر است با استفاده از کد و منطق استدلال کنید.

۲- تحلیل لغات ویکیپدیای فارسی

در این تمرین یک دامپ ۵ گیگابایتی از ویکیپدیای فارسی را تحلیل خواهید کرد و با مقدمات کار با RDD ها در اسپارک بیشتر آشنا خواهید شد.

در وبسایت databricks برای خود اکانت ساخته و وارد اکانتتان بشوید:

https://databricks.com/try tabricks#signup/communityda

- ایجاد کلاستر : از منوی سمت چپ گزینه Clusters را انتخاب کنید . روی دکمه Create Cluster کلیک کنید . برای کلاستر خود یک نام انتخاب کرده، نسخه پایتون ۳ را انتخاب کنید و دکمه Create Cluster را انتخاب کنید.
- واردکردن نوتبوک تمرین : از منوی workspace گزینه import را انتخاب کنید و نوتبوکی که در کنار این فایل تمرین قرار گرفته است را انتخاب کنید.
- از Colab به جای databricks می توانید استفاده کنید. روش راه اندازی spark در colab با یک مثال در کلاس مطرح شد و لینک مثال آموزشی آن نیز تقدیم شده است.
- نوت بوک مورد نظر را کامل کنید. تاحد امکان برای هر قسمت سوال توضیح مناسب به صورت mark down در نوتبوک قرار دهید. گزارشی شامل کدها و نتایج خود ارائه کنید. نتایج حاصل را در گزارش در یک باکس قراردهید.