دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس: تحلیل داده های حجیم

تمرین سوم و چهارم

مدرس: 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 تحلیل‌های خود از نمودارها و هر visualization ابتکاری دیگر استفاده کنید، گزارش هایی که صرفاً شامل کد باشند تنها نمره assignment programming را خواهند گرفت.
- پاسخ های قسمت های عملی می بایست حتماً در فرمت ipynb. باشند، بنابراین میبایست تمامی بخش های عملی به صورت یک notebook jupyter تحویل داده شوند.
- فایل تحویلی را بصورت فشرده و با نام گذاری مناسب تحویل دهید:

MDA2023-HWn-StudentNumber.zip

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 4 ام، به ازای هر روز اضافی، ۲۰ درصد از نمره تمرین را از دست خواهید داد.

ازآنجا که هدف این درس تحلیل داده های 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، ۱۰۰ نمره منفی برای طرفین در نظر گرفته می شود، تحویل تمرین فقط بصورت آنلاین و از طریق CW در زمان تعیین شده مورد پذیرش خواهد بود.

a.r.zargaran7@gmail.com

داده مورد استفاده در این تمرین داده های توییتر در یک ماه اخیر است. در ابتدا داده‌ها را دانلود کنید و به سوالات زیر پاسخ دهید.

فیلدهای مهمی که در این تمرین ممکن است به آن نیاز داشته باشید به شرح زیر است: (فرمت زیر دارای شکل سلسله مراتبی است.)

- id: unique id for identifying tweet
- user
    - id: unique id for identifying user
    - screen_name: user name
    - name: the name of the user
    - description: biography of the user
- tweet_type: show the type of tweet (quoted, replied, retweeted, generated.)
- in_reply_to_status_id_str: if tweet_type is “replied”, shows what tweet is this tweet in reply to.
- in_reply_to_user_id_str: if tweet_type is “replied”, shows what user is this tweet in reply to.
- quoted_status: if tweet_type is “quoted” shows some information of the tweet quoted from.
    - id: unique id for identifying tweet
    - created_at: the time which tweet is created
        - user:
            - id: unique id for identifying user
            - screen_name: user name
            - name: the name of the user
            - description: biography of the user
- retweeted_status: if tweet_type is “retweeted” shows some information of the tweet retweeted from.
    - id: unique id for identifying tweet
    - created_at: the time which tweet is created
        - user:
            - id: unique id for identifying user
            - screen_name: user name
            - name: the name of the user
            - description: biography of the user

تمرین سوم Twitter User Engagement Recommender System

توییتر، یک پلتفرم محبوب میکروبلاگینگ، به دنبال بهبود تعامل و رضایت کاربران است. این هدف با پیاده‌سازی یک سامانه پیشنهاد دهی قوی تر به کاربران ارائه شده است. هدف نهایی ارائه پیشنهادات شخصی‌سازی شده برای توییت‌ها است که با سلیقه کاربران هماهنگ باشد و در نهایت تجربه کلی کاربران را در این پلتفرم بهبود بخشد.

یک مجموعه داده در اختیار شما قرار دارد که شامل اطلاعاتی در مورد تعاملات کاربران با توییت‌ها می‌شود. این مجموعه داده شامل اطلاعات زیر است:

1. User IDs
2. Tweet IDs
3. Retweets
4. Replies
5. Quoted Tweets

- با جست و جو در اینترنت، معیارهای دیگری را که می‌توان برای تعامل کاربران با توییت‌ها در نظر گرفت پیشنهاد دهید.
- سیستمی را پیاده سازی کنید که به کاربران بر اساس تعاملات و ترجیحات آن‌ها، توییت‌های شخصی‌سازی شده ارائه دهد. سیستم باید به ریتوییت‌ها، پاسخ‌های کاربران و نقل قول‌ها توجه کند و پیشنهادات شخصی برای هر کاربر ارائه دهد.
- چگونه با کمک ماتریس ساخته شده می‌توان کاربرانی که تعاملات مشابه‌ای دارند را به یکدیگر پیشنهاد دهیم؟
- الگوریتم پیشنهاد شده برای کاربر جدیدی که به این ماتریس اضافه می‌شود نیاز دارد که از ابتدا مدل را بدست آورد؟ (online یا offline است)