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 چهارم Random Walk algorithms

در این سوال می‌خواهیم مسئله‌ی بالا را با کمک الگوریتم‌های Random Walks حل کنیم.

- گره‌ها و یال‌های گراف، چه اطلاعاتی را نشان می‌دهند؟
- با پیاده سازی الگوریتم پیشنهادی، کاربرانی که تعاملاتی مشابه‌ای به یکدیگر دارند را پیدا کنید و به یکدیگر پیشنهاد دهید.
- یک کاربر را انتخاب کنید و کاربرانی را با کمک الگوریتم تمرین سوم و چهارم پیشنهاد دهید. آیا کاربران پیشنهادی در هر دو الگوریتم یکسان هستند؟
- این الگوریتم از نظر offline و یا online بودن چگونه رفتار می‌کند؟
- مزایا و معایب هر یک از الگوریتم چیست؟