دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس: تحلیل داده های حجیم

تمرین اول و دوم

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 programming را خواهند گرفت. assignment
- پاسخ های قسمت های عملی می بایست حتماً در فرمت ipynb. باشند ، بنابراین میبایست تمامی بخش های عملی به صورت‌یک jupyter تحویل داده شوند. notebook
- 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 ۱۲ روز تاخیر داشته باشید و به ازای هر تمرین بیشتر از ۴ روز تاخیر، مشمول کسری نمره می باشد، بطوری که بعد از روز ۴ 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

تمرین اول

فایل question1.ipynb را باز کنید و به ترتیب بخش های مختلف آن را اجرا کنید.

بخش اول

دیتاست تهیه شده، مربوط به اخبار فارسی است. در ابتدا نیاز است که دیتاست را با توجه به لینک قرار داده شده دانلود کنید و آن را در دایرکتوری مناسب قرار دهید. در این بخش نیازی نیست کاری انجام دهید تنها باید سلول های jupyter را اجرا کنید تا دیتاست، در قالب RDD آماده گردد.

بخش دوم

در این بخش نیاز است که بدنه (متن) خبر را تمیز و نرمالایز کنید. در قدم اول باید تمام کاراکترهای بدون معنا مانند ویرگول، نقطه، علامت‌های ریاضی و ... را حذف کنید. پس از آن نیاز است stopwordها مانند است، هست، را و ... را پیدا کنید و آن‌ها را حذف کنید.

بخش سوم

در این بخش، به تمام موارد مورد پرسش پاسخ دهید و نتایج را با visualization مناسب ارائه نمایید.

توجه: در word cloud یا ابر واژگان نیاز است که از فیلد keywords استفاده کنید و نمایش تنها ۲۰ واژه ای که بیشترین امتیاز را دارد کافی است.

بخش چهارم

قصد داریم الگوریتمی را پیدا کنیم که با کمک آن بتوانیم مجموعه های سه کلمه ای را پیدا کنیم که بیشترین مشاهده حضور در کنار یکدیگر را دارند.

با توجه به مطالب درس نیاز است که روشی برای حل این مسئله ارائه دهید. در قدم بعدی نیاز است نتایج بدست آمده را صحت سنجی کنیم. یکی از روش‌ها برای صحت سنجی استفاده از الگوریتم tf-idf است. با مطالعه‌ی tf-idf روشی را پیشنهاد دهید تا بتوان این صحت سنجی را انجام داد.

تمرین دوم

در ادامه تمرین اول قصد داریم با کمک الگوریتم LSH، کدی را پیاده سازی کنیم تا اخبار مشابه را پیدا کنیم. آنچه که در این سوال از شما خواسته می‌شود این است که در ورودی unique id یک خبر داده شود و و تمام اخبار مشابه با آن‌را نمایش دهید.

- استفاده از روش‌های visualization مناسب مزیت محسوب می‌شود و امتیاز اضافه دارد.
- سعی کنید تعداد cluster های مناسب را با توجه به مطالبی که آموخته اید بیابید.
- روش انتخاب تعداد cluster های مناسب را توضیح دهید.