پروژه دوم

درس مبانی داده کاوی

نیمسال دوم تحصیلی ۱۴۰۲-۱۴۰۱

در این پروژه دانشجویان باید از میان دیتاست هایی که در اختیارشان قرار گرفته یک دیتاست را به دلخواه انتخاب کرده و موارد خواسته شده که در قالب فاز های مختلف پروژه تعریف میشوند را پیاده سازی کنند.

فاز اول(Preprocessing):

در این بخش شما باید داده هایی که در اختیار دارید را به فرمی ساختار یافته و تمیز تبدیل نمایید.

- missing value ها هندل شوند.
- برای داده های عددی outlier ها شناسایی و حذف شوند.
- در صورت نیاز در این فاز data reduction انجام شود.
- در صورت نیاز داده های عددی به داده های categorical تبدیل شوند.
- برای داده های متنی در صورت نیاز عملیات stemming,lemmatizing و حذف stopword ها انجام شود.(برای این کار میتوانید از کتابخانه هایی مانند nltk استفاده کنید.)

هم چنین در این فاز انجام یک سری مقایسه های آماری با توجه به دیتاست تعیین خواهد شد که توسط هر گروه باید انجام شود.

فاز دوم(Extracting Frequent Pattern):

در این فاز هر گروه باید با توجه به الگوی مشخص شده توسط تیم حل تمرین برای هردیتاست،الگوهای مکرر موجود در دیتاست clean حاصل از فاز قبلی خود را استخراج نمایند.استفاده از کتابخانه های مربوط به الگوهای مکرر مجاز بوده و میتوان از کتابخانه هایی مانند mlxtend استفاده کرد.

فاز سوم(Classification & Clustering):

در این فاز اگر خوشه بندی یا طبقه بندی دیتاست بر اساس داده های متنی باشد،لازم است داده های clean فاز قبل را با استفاده از BERT تبدیل به vector نمایید تا برای ماشین قابل فهم بوده و بتوانید عملیات خوشه بندی و طبقه بندی را انجام دهید.برای استفاده ازBERTدر این فاز از پروژه میتوانید از مدل های پیش آموزش داده شده آن که در کتابخانه های مختلف مانندHugging Faceیاsentence-transformers موجود است استفاده نمایید.

در ادامه باید از خروجی مرحله قبل و به کمک الگوریتم های موجود عملیات خوشه بندی روی دیتاست انجام شود.

همچنین با توجه به دیتاست هر گروه یک مساله طبقه بندی تعریف شده که باید با توجه به دیتاست گروه خود طبقه بندی خواسته شده را پیاده سازی کنید.(دقت کنید که بخشی از دیتاست برای تست طبقه بندی پیاده سازی شده باید در نظر گرفته شود.)

نکات تحویل:

- لازم است با استفاده از کتابخانهmatplotlibیا دیگر کتابخانه های مشابه،بصری سازی های لازم جهت درک خروجی هر مرحله از کد خود را انجام نمایید.
- فایل ها باید در قالب studentOneName-studentTwoName-phaseNum.zip ارسال شود.
- ارسال فایل تنها از طریق سامانهVUمورد قبول بوده و فایل های ارسال شده در تلگرام و ... تصحیح نخواهند شد.