فاز یک – پیش پردازش و استخراج ویژگی

هدف از این فاز آشنایی بیشتر با فرآیندهای متن کاوی و همچنین روش های پیش پردازش بر روی دیتاستی که در اختیار دارید و استخراج ویژگی برای انجام عملیات های خوشه بندی و طبقه بندی در فاز بعدی می باشد.

Preprocessing

در این بخش شما باید داده هایی را که از صفحات مختلف وب جمع آوری نموده اید به فرمی ساختار یافته و تمیز تبدیل نمایید.

لازم است شما اطلاعات مدنظر را با توجه به ساختار اطلاعاتی که از دانشگاه موردنظر گرفته اید پیدا نموده و عملیات را بر روی انها انجام نمایید(این اطلاعات احتمالا در ویژگی های outcome یا objective یا description هر درس موجود می باشند اما شما می بایست با بررسی دقیق دیتاست خود ویژگی های مطلوب را انتخاب نمایید.). برای انجام فاز بعد لازم است شما این ویژگی ها را با یکدیگر ترکیب نموده و عملیات Embedding را بر روی ترکیب آنها انجام نمایید.

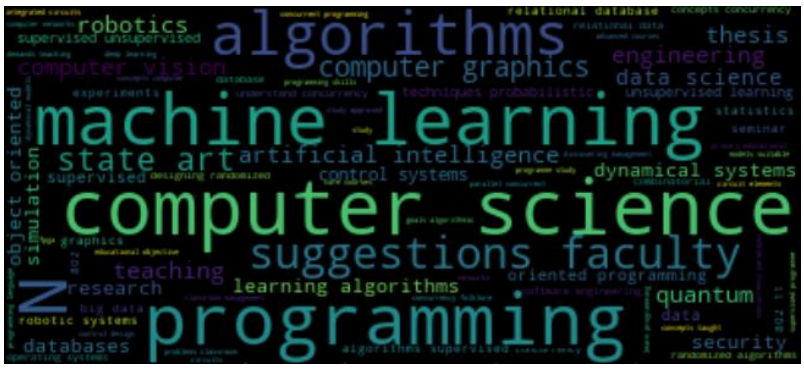
پس از ترکیب ویژگی های فوق شما می بایست عملیات پیش پردازش مربوط به داده های متنی همانند stemming و lemmatizing و حذف stopword ها را انجام داده و متن clean شده ای را برای بخش بعدی آماده نمایید.

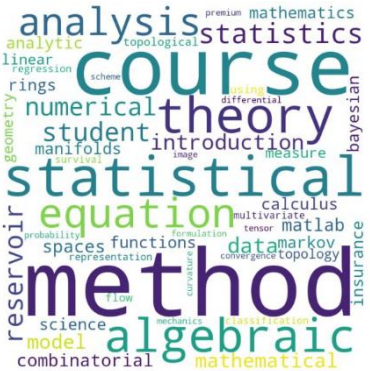
همچنین لازم است مقایسه ای بین تعداد دروس ارائه شده توسط هر دپارتمان انجام و اطلاعات آماری آنها مانند میانگین دروس ارائه شده هر دپارتمان و ... را به دست آورید.

Keyword Extraction

برای این بخش شما می بایست کلمات کلیدی موجود در متن های به دست آمده پس از انجام پیش پردازش های فاز قبل را استخراج نمایید.برای استخراج کلمات کلیدی با استفاده از مدل آموزش داده شده های BERT به روشی ساده تر می توانید از KeyBERT استفاده نموده و یا از دیگر کتابخانه های موجود در این حوزه استفاده نمایید.

نمونه هایی از کلمات کلیدی استخراج شده از دروس رشته های مختلف:





Frequent Pattern Extraction

در این بخش لازم است شما سبدهای خود را بر اساس کلمات کلیدی استخراج شده در بخش قبلی ایجاد و سپس آیتم های پرتکرار را استخراج نمایید. شما می‌بایست روابط بین کلمات کلیدی را نیز با استفاده از استخراج قوانین انجمنی به دست اورید. استفاده از کتابخانه های مربوط به الگوهای مکرر مجاز بوده و شما میتوانید از کتابخانه هایی مانند mlxtend و ... استفاده نمایید.

نکات تحویل

- لازم است با استفاده از کتابخانه matplotlib یا دیگر کتابخانه های مشابه، بصری سازی های لازم جهت درک خروجی هر مرحله از کد خود را انجام نمایید.
- این فاز در قالب گروه های ۲ نفره و یا به صورت انفرادی قابل انجام میباشد.
- در این فاز شما می بایست فایل notebook خود(بدون حذف خروجی های حاصل از اجرا) و دیتاستی که پس از انجام عملیات پیش پردازش در اختیار دارید را ارسال نمایید.
- فایل ها باید در قالب Student1Name-Student2Name-CourseCode-phase1.zip ارسال شود.
  - CourseCode برای گروه ساعت ۴ عدد ۱ و برای گروه ساعت ۶ عدد ۲ می باشد.
- ارسال فایل تنها از طریق سامانه VU مورد قبول بوده و فایل های ارسال شده در تلگرام و ... تصحیح نخواهند شد.