Homework 2

Due: November 10, 11:59pm CST

Instructions

Like the last homework, this homework consists of two parts: a coding part and a written part. Unlike the last homework, I will not be asking specific questions. Rather, the point of the written part of the homework is to report on your findings in the coding part.

As before, I will provide the skeleton of the code for you to fill in. However, differently from last time, this time you will have more freedom. The dataset you choose will be up to you, as are the methods you choose to discuss and any patterns you find.

The goal of this homework is to provide practice for one part of your upcoming project, namely exploratory data analysis. Hopefully by finishing this homework you will familiarize yourself with SKLearn's clustering classes, and methods for visualization.

Report Requirements

You can structure the report how you see fit, but I expect to see the following:

- Description of dataset, and how you obtained it. Are you trying to embed documents or words? Are you doing any preprocessing of the data? If so, describe it. Feel free to do a Google search for common datasets that other people use, or feel free to scrape a new dataset.
- Description of all embedding models. Choose at least two models and talk about the method the models use for embedding, discussing also any choices of hyperparameter (e.g. if doing any dimensionality reduction via SVD, what dimensionality did you choose?).
- Cluster your models. Use hierarchical clustering to determine roughly how many natural clusters your embeddings have and provide a visualization of this. Having done so, use at least two of SKLearn's many clustering algorithms, and discuss the results of your clusters. You may or may not

have gold-standard embeddings available.[1] If you do not, use the output of your hierarchical clustering for your choice of k (e.g. if you feel there are five natural clusters, you would choose $k = 5$) as your gold-standard embeddings. Use the Adjusted Rand Index to compare your different clustering methods.

- Summary. Write a summary of your findings. You will have compared at least two models on your dataset, and performed clusterings with at least two different methods. Did you find any interesting differences between the different models/methods?
- Bonus: In class, the only traditional clustering method we learnt was K-Means clustering. Above, I ask you to use at least two methods. Practically, this won't be a problem because SKLearn's interface is the same regardless of method. However under the hood, different clustering algorithms work differently. The bonus part of the assignment is to describe from a high-level (i.e. no mathematical details necessary) what your second clustering method is doing. There are many popular datascience websites which provide descriptions of different clustering methods. Feel free to summarize what one of these websites says.

[1] This will depend on your dataset. When we embedded the S&P 500 dataset, we had a natural gold-standard for clustering, namely the GICS sector. Since you will be choosing your own datasets, you may or may not have such a gold-standard against which to test your clusterings.