Homework 6- 690V- Visual Analytics

CS 690V, UMass Amherst, Fall 2017

Aditya Agrawal, Sanjay Reddy S
adityaagrawa@cs.umass.edu, ssatti@umass.edu

News Clustering Example

For document clustering, we clustered various news articles into 10 genres. The data is taken from kaggle. But it was over 600MB and thus we focused on only 5 publications instead of 15 (Just 500 records instead of 150,000! Note that the same program will work for the whole dataset. Just that it will take lot of time).

The processed dataset, used for this program can be downloaded from here (this one is about 200 mb. Place it in the same folder as news_cluster.py file). Requires nltk library. Runs on Python 3. (For including more records, just modify line 18. The program will run fine but will take more time)

To run: python3 news_cluster.py.

Will generate 3 .html files (rendered ones are already present in the folder).

It uses stemming on the contents of news articles (and other common techniques like removing stopwords) to reduce the vocabulary and increase the accuracy. It uses Tf-Idf vectorizer to create vectors from the words (used in clustering by K-Means). It uses SVD techniques to perform PCA dimensionality reduction (from over 20,000 dimensions to 100) and clusters them into 10 classes.

The idea was inspired from this blog post. His insights into the graphs are also good.