Mark Torres
CPSC 453: Machine Learning in Biology
Professor Krishnaswamy
14 November 2019

Project Proposal

Project Idea: Comparing performance of clustering techniques against LDA when analyzing Twitter text data.

Summary: LDA is a common way to group text data into groups, via analyzing words and assigning topics to these texts. LDA doesn't require a distance metric, but rather uses probability distributions and bag-of-words models to create its groupings. In contrast, clustering techniques (e.g., k-means, Louvain) utilize and minimize some sort of distance metric in order to create groupings. I am interested in how typical clustering algorithms perform in contrast to LDA when performed on a corpus of Twitter data. I plan on using LDA to classify tweets into certain topics and examining what topics arise as a result. To compare with LDA, I plan to use document embedding methods in conjunction with a neural network in order to assign vector projects and distances to the corpus of the text data, which would allow me to perform clustering on the text. I then intend to compare the groups that arise from each of the algorithm and study how the differences in the groups arise as a result of the particular methodology that each algorithm implements in its classification. Moreover, I would like to examine the differences in the groupings that each algorithm creates and how each algorithm approaches certain test cases (e.g., how does the algorithm approach tweets that could fit in multiple topics?).

Overview of computational methods that I plan to implement:

Latent Dirichlet Allocation (LDA): a generative topic model that groups documents into topics.

Clustering Algorithms

- K-means clustering
- Spectral clustering
- Louvain

Some papers/resources I plan to cite and study:

- https://stats.stackexchange.com/questions/292281/clustering-with-latent-dirichlet-allocation-lda-distance-measure
- https://www.researchgate.net/publication/326163504_Revisiting_K-Means_and_Topic_Modeling_a_Comparison_Study_to_Cluster_Arabic_Documents
- https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d
- http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf