Comprehensive Exam – Tommy Jones

Implementation

To answer the following questions you may use python, or any other programming language of your choice. You are allowed to use existing packages, such as python scikit-learn, provided that you reference them properly.

Dataset

To answer the following questions, please download the 20 Newsgroups Dataset (http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups).

Preprocessing

You may need to do some preprocessing of the data (e.g., the removal of stop words etc.) in order to answer the questions that follow. In this case, please document all the preprocessing steps that you used.

Question 1. Exploratory visualization

What are the top 20 words? Please provide a plot showing their frequency distribution.

Question 2. Clustering

Calculate the TF IDF vectors for each document. Perform hierarchical clustering of the documents, using a distance metric based on cosine similarity. You may use packages such as python scikit-learn, or another programming language of your choice. Use the elbow method to determine the most suitable number of clusters and explain why this is the case. Please include the elbow plot in your answers. What do you observe from this plot? Is there a clear elbow? Please discuss.

Question 3. Topic Modeling

Perform Latent Dirichlet Allocation (LDA) using any package and programming language of your choice.

- Use the elbow method to determine the optimal number of topics. Does it agree with the number of clusters found in Question 2?
- Using the above number of topics, run the LDA algorithm:
  (a) for 1000 iterations;
  (b) for 200 iterations, stop and re-train for 800 iterations;
  (c) for 200 iterations, stop and re-train for another 200, etc. until a total of 1000 iterations.

Plot the Intertopic Distance Plot for the resulting topics of each of the above 3 cases. What do you observe in these results? Please discuss what your results mean.