EAST TENNESSEE STATE
UNIVERSITY

Project 4: Natural Language Processing

Overview:

This project is somewhat open-ended. I am not providing pre-written code. Instead, I am providing a corpus of all inaugural addresses of each presidential inauguration since Washington in 1789. The document corpus contains 58 documents that you will use in the course of this project. For the dates associated with all 58 inaugurations of the President of the United States, see https://en.wikipedia.org/wiki/United_States_presidential_inauguration#List_of_inaugural_ceremonies

Assignment Requirements and Constraints

Given the provided ZIP folder named "speeches," you must perform a natural language processing methodology to determine two things: 1. What major themes have existed in the United States over time? To do this, you will want to do a Bag-of-Words technique (see below), and 2. Determine what presidential speeches have shared similar terminology using Clustering Analysis? Note that the purpose of this assignment is not to make political comments or determinations, but rather to track important topics. To do this, you will need to determine themes within and across documents.

1. You **must use a bag of words (n-grams) technique** with frequency counts.
   a. Stop words, stemming, lemmatization are up to you, but you must clearly justify why you did what you did in the writeup.
   b. You may use n-grams larger than words if you feel this will provide a better result, but again, you must justify your process in the writeup.
2. You **must construct a vector model** (specifically, a document-term matrix or a term-document matrix).
   a. You must save this vector model to a CSV file, named **project4_dtm.csv**
3. You **must implement a clustering process** to determine what major themes were important to what presidents, regardless of time.
   a. You should do this using a k-means clustering algorithm on your vector model. Note that the value for k is not well-defined for this data, so part of your job is going to be finding a suitable value for k, and then conducting the clustering using that value for k. You may want to use the **calinski_harabaz_score** metric from scikit-learn.metrics to determine the value for k. Start at k=2 and continue up from there. The value of k that produces the highest calinski_harabaz_score would be considered the best clustering. See https://scikit-learn.org/stable/modules/clustering.html#calinski-harabaz-index.
   <u>**Hint**</u>: Also see Python library sklearn.cluster (http://scikit-learn.org/stable/modules/clustering.html#clustering)
4. You **must report your findings in a well-defined and well-written document** (named **project4_writeup.docx**) that details:
   a. What you did, exactly. Detail your process and why you made the choices you made.
   b. What conclusions you reached. Word clouds, charts, graphs, etc., are encouraged.
      i. Report on the major themes found over time. These should emerge from your Bag-of-Words technique. Many of these will be the same over time, so you may have to dig a little deeper into your word frequencies.
      ii. Report the important topics for groups of presidents (based on the k-means clustering). For example, those who were presidents during wartime might share common language and cluster together.

Submission and Due Date

Place all files used for this project into a folder named lastname_project4 (of course, you should use your own last name). ZIP the folder and submit it to the dropbox **at or before Monday, April 8, 2018 at 11:59pm**. The rubric follows.

Rubric

	Points Possible	Points Gained	Comments
Bag of Words/n-gram Technique			
Word Frequencies, Stemming, Lemmatization	10		
Vector Model (CSV)	10		
Clustering Process	20		
Results	20		
Writeup			
1. What you did (the process) and justifications	10		
2. What problems you had	10		
3. What conclusions you reached	20		