Csci 517: Natural Language Processing

Project: Text Clustering

Deadline: 11:59 PM, November 28

Overview

Through this project, you will get familiar with text clustering, cluster comparison, and evaluation. You will be using the NSF Abstract dataset in this project. You can download the data from here [Data Source and Description]. You will only be working with the "Part 1".

Goals

1. **K-means Clustering:** Each file contains the abstract and some metadata of an NSF award. You will use just the abstract, not the metadata. Note that some files may not have an abstract. Your goal is to implement a **K-means** clustering algorithm to cluster the abstracts into separate groups. You will use K = 5, 7, 10, 12, 15. You can use cosine similarity to measure distance between two abstracts. Note that you have to implement the algorithm by yourself; third-party based implementation can't be used.

2. **DBSCAN Clustering:** You will need to implement the DBSCAN clustering algorithm by yourself and apply that over the dataset. Note that DBSCAN doesn't require you to set the number of clusters beforehand. However, you have to set the value of MinPts and Epsilon either by trial and error or by following the guidelines given in Lecture_11 page 83. If you follow the guidelines, you will get **5 bonus points**. Your project description should clearly explain, using a graph, how you followed the guideline and what value you used.

3. **Comparison:** You will use SSE to measure the quality of your clusterings. You have to determine whether or not K-means perform better than DBSCAN in this dataset.

Presentation

You will need to present your project and findings in the class on November 29. Each student will have 5 minutes to present.

Deliverables

1. All source code: Your code must have plenty of comments so that it is easier for anyone to understand the flow and function of your code.
2. Project description: You have to submit a two page project report. The report should indicate how you implemented K-means and DBSCAN, what data structures you used, how you set the parameters, what results you got and so on.
3. Readme.txt: It should describe in details how to execute your source program.
4. Presentation slides: You will need to submit a **PDF version** of your presentation slides. Any other format won't be acceptable.

All deliverables should be submitted in a single zipped file through the Blackboard system.

Grading

- Program correctness [40]
- Documentation
  - Readme.txt [10]
  - Code comment [10]
  - Project description [20]
- Presentation [20]