Csci 581/543: Data Mining
Project 3
Deadline: May 4, 2018. 11:59 PM

Date: April 22, 2018
Points: 100

Requirements:

- The project must be handed in electronically through Blackboard, in **zip format only**.
- Your submission **must not contain** any information which may reveal your identity.
- Anonymous grading is enabled. If there is any identity-revealing information, **50 marks will be deducted**.

Overview

The objective of this project is to be familiar with unsupervised learning from data. We will use the famous 20-newsgroups dataset. You can download the data from here http://qwone.com/~jason/20Newsgroups/.

Task Description

In the following tasks, you will group the articles into clusters. After that you will use both internal measures (SSE) and external measures (Entropy) to evaluate the clustering. You will use bag-of-words model to represent the documents as vectors. You can use simple euclidean distance to measure the proximity among the documents. To achieve better results, you can use the tf-idf model, cosine similarity, etc.

1. Calculate the 10-biggest newsgroups from the 20-newsgroup dataset. In all further tasks, use this 10 groups only.
2. Implement K-means clustering algorithm. You cannot use external implementation. Experiment with K = 2, 4, 6, 8, 10, 12, 14, 16. Show a figure with number of cluster on X axis and SSE value on Y axis.
3. Repeat Task 2 with Bisecting K-means algorithm.
4. Repeat Task 2 with DBSCAN algorithm. You have to find the suitable parameter threshold values. **You can use external implementation of DBSCAN. But, if you implement by yourself, there is 20 bonus points**. Here, you will not tune the value of K but the values of the two parameters. You will report how the SSE changes with respect to eps and minpts. You can set eps to a fixed value and then tune minpts to see its effect on SSE and vice versa.
5. Repeat Task 2, 3 and 4 but in the figures, show Average Entropy in the Y axis instead of SSE. You can find how to calculate Entropy for each cluster on Page 104 of L13.

Deliverables [in a single zipped file]

1. All of your source codes.
2. Project Description: Write a document (**pdf file**) about this project, which should include the following information:
    - Describe the method you used to represent the documents.

- Describe your observation from cluster evaluation. Which algorithm appeared to be better- kmeans or bisecting k-means or DBSCAN?
- Describe your observation from internal and external evaluation.
- The document should be as detailed as possible, with figures and tables, but no more than 4 pages.

Grading

The full mark of this task is 100.

Basic Points (50 points)

All required materials are submitted as in required format. The code can compile, execute, and produce clusters. The clustering should come from the above specified clustering algorithms (certainly, it cannot be a random guess). Incomplete submission will result in points deduction.

Document Quality Points (50 points)

This part evaluates the quality of your comments in source code and the pdf file in submission. A well written document should be complete, readable and well formatted.

Bonus Point (20 points)

- If you implement the DBSCAN by yourself, then you will get 20 bonus points. Explain in the pdf document if you have put any bonus effort.
- If you use word2vec instead of bag-of-words model, then there is 20 bonus points. You can get guidelines for word2vec use in Python from this tutorial here http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/