CS 429 Information Retrieval

Bonus (optional/extra-credit) Assignment: k-means Clustering Algorithm

Points: 85

Submission deadline: Tuesday, 04/30/19, 11:59 PM

Late submission deadline (5% penalty): Thursday, 05/02/19, 11:59 PM

Note: 1) This assignment is not mandatory. It is extra credit. 2) This is individual work.

In this assignment, you will implement the k-means clustering algorithm as discussed in class (see lecture notes)

Part A: K-means algorithm [40pts]

- You will implement the k-means algorithm that uses a vector space representation for the documents with *tf-idf* weighting. You may reuse code from the previous assignments as needed.
- Your implementation should read the documents and generate its document vectors.
- Your algorithm should take *k*, the number of clusters as an input to the program.
- Your algorithm should print out:
    - For each cluster, its residual sum of squares (RSS) values and the document ID of the document closest to its centroid. The document IDs are integers, starting at 1.
    - The average RSS value
    - Time taken for computation.

Part B: Experimental study [35pts]

- You will conduct an experimental study, with the TIME dataset, to understand the relationship between RSS and the number of clusters (k). See fig 16.8 in text book. The goals is to measure the RSS values for various cluster sizes ranging from k=2 to k=30. Determine the value of ‘k’ that provides a good tradeoff with RSS values.
- Your report will have a plot comparing the RSS values with *k.*
- Also include in the report, the following details:
    - What is the procedure for selecting the initial set of centroids in your implementation?
    - What is the stopping condition in your implementation?
    - From the plot, what is the value of ‘*k*’ that provides a good tradeoff with change in RSS?

Other instructions:

- Implement in Python 3.0.
- Comment your code appropriately.
- You may reuse the code from earlier assignments.

Attachments:

- TIME.rar - collection of documents
- Skeleton code – implement the functions in the code. Use additional functions as needed.

Submission:

Submit the following files on blackboard as a .zip file.

1. Report.pdf: Report with experimental study.

2. Output.txt: containing the output generated by your code for three values of k. This is for testing your code.