CS6200 Information Retrieval
Fall 2019
Instructor: Omar Alonso

Extra Homework. Clustering

Objective: Cluster the document titles from the AP collection using your implementation of k-means. This is for titles only (not the document body).

This assignment involves:

1. Run k-means with the following values for k = 50, 100, 200, 300. For each k value you need to run 5 iterations of the algorithm.
2. Design and implement a technique for summarizing the content of each cluster.
3. For each k in item #1, rank clusters by their intra-class similarity. You need to specify the intra-class similarity.
4. Plot the results of item #3 for the top 30 clusters and analyze the results.

What you need to submit

1. Your k-means implementation in Java source code
2. A document that covers item #3 in detail.
3. Documentation on how to compile and run the code.