COMPSCI546_132815_FA20-Applied Information Retrieval Fall 2020

Moodle home / My courses / COMPSCI546_132815_FA20 / Classification and Clustering / Programming Assignment: Clustering

Programming Assignment: Clustering

The purpose of this project is to explore clustering on a small collection of documents. You will use the collection and your index from the previous project.

Clustering

Implement clustering for the online clustering task (filtering). That is, implement agglomerative clustering, using your index from the previous projects. Beginning with the first document in the index, treat the contents as a stream of documents, clustering them in order of occurrence in the index. You may find it necessary to modify your Indexer to support the clustering activity.

Implement all four linking choices (single, complete, average, and average group). Use your implementation to cluster the shakespeare scenes collection across thresholds from 0.05 to 0.95 with mean linking.

Evaluation

Grading Rubric

(5%) Submission is in the correct format.

A ***single*** archive file (zip, tar.gz) was submitted to Moodle and it contains at least the following contents:

- ***report.pdf*** - your report (see below)
- ***src/**** - your source code
- ***README***
  - It has instructions for downloading dependencies the code.
  - It has instructions for building the code.
  - It has instructions for running the code.
- ***cluster-<thresh>.out***
  - Clustering Output Files: One file for each of the threshold settings. These should be in two column format:

<clusterID> <DocumentID>

(50%) Source code that implements online clustering.

Please be aware that you may lose points for problems in your code even if it appears to run correctly. We expect to see code used for all parts of the assignment in your submission.

(45%) The report discusses your work.

- (5%) Description of the system, design tradeoffs, questions you had and how you resolved them, etc. List the software libraries you used, and for what purpose.
- (20%) Explain what happens as the threshold value increases. Table the clustering results (number of clusters, size of each) for all of the threshold values.
- (20%) Explain the difference between the four linking strategies. What behavior would you expect from each of them. Compare to the results you obtained using mean.

Submission status

Submission status	No attempt
Grading status	Not graded
Due date	Friday, November 6, 2020, 11:59 PM
Time remaining	10 days 5 hours
Last modified	-
Submission comments	⊞ Comments (0)

Add submission

You have not made a submission yet

◀ Inference Network Solution Jump to... Quiz: Evaluation ▶