UNIVERSITY OF ILLINOIS AT CHICAGO, INFORMATION AND DECISION SCIENCES

Clustering Using K-Means under Mahout

SOFTWARE REQUIREMENTS

Hadoop 1.0.3
Mahout 0.7.0

DATASET

Applying K-Means under Mahout and Hadoop environment to find clusters for the Yelp Review data, which can be found here:

https://www.yelp.com/dataset_challenge/dataset.

INSTRUCTIONS

(1). You first download the compressed data and unzip it into a folder. Then you extract all review texts into multiple files under the same folder or into one file. Each individual review here becomes a data point and K-means algorithm can be applied to find clusters. As we don't know the number of clusters (topics) among these reviews, therefore K can be set to be any numbers before you run the algorithm. Each document will be represented as a vector using TF-IDF of each unique word. You can remove all stop words, which can be found here:

http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

(2). Two strategies you can use to remove all stop words for all documents. (a) Write your own simple Java code to remove all stop words within the original documents to generate new documents. (b) Extends Analyzer class in the Lucence package to obtain your customized Analyzer. Then add -a <your customized analyzer class name> into command seq2sparse when you generate vector files.

(3). Once you get the clusters, you need to write another Job to get top 10 frequent words for each cluster.

Submission

You need to submit a zip file containing all source codes (*.java files) and all commands you use to run Mahout. Please submit it through the Blackboard. In addition, you should also include a README file to explain each file and command. The output you generate should be files containing cluster id and its corresponding top 10 frequent words.

$ClusterID_1 : Word_1, Word_2, \cdots, Word_{10}$
$ClusterID_2 : Word_1, Word_2, \cdots, Word_{10}$
$ClusterID_3 : Word_1, Word_2, \cdots, Word_{10}$
$\ldots$
$ClusterID_i : Word_1, Word_2, \cdots, Word_{10}$
$\ldots$
$ClusterID_k : Word_1, Word_2, \cdots, Word_{10}$