HW4

Deadline: Apr. 18th 2:59 P.M. (before class)

Data

Reuters-21578 is a collection of documents that appeared on Reuters newswire in 1987. The collection was released by Reuters and CGI in 1990 and now is available in the link below.
https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Download and unzip reuters21578.tar.gz from the link above. Besides some descriptive files, you will find the collection contains 22 data files (reut2-XXX.sgm).

TODO

Implement K-Means with Spark MLlib and classify the 22 data files into 4 clusters. Specifically, you will need to

1. Convert the 22 files lowercase
   (other pre-processing such as stop word removal is not required for simplicity.)
2. Transform the lowercased file to TF-IDF, so you have a matrix representation for the text files. The value in row i column j represents the TF-IDF value of token j in file i. (more details is available at https://en.wikipedia.org/wiki/Tf%E2%80%93idf)
3. Cluster the 22 files into 4 groups using K-Means algorithm implemented on the TF-IDF matrix.
4. Report cluster ID for each file. Example output format is

   reut2-000.sgm,0
   reut2-001.sgm,0
   reut2-002.sgm,1
   …
   reut2-021.sgm,3

Corresponding documentations can be found at

1. TF-IDF
   https://spark.apache.org/docs/1.6.0/mllib-feature-extraction.html#tf-idf

2. K-Means
   https://spark.apache.org/docs/1.6.0/mllib-clustering.html#k-means
   https://spark.apache.org/docs/1.6.0/api/python/pyspark.mllib.html#pyspark.mllib.clustering.KMeansModel

Submission

1. Spark program
2. Output file