K Means Mahout Clustering

1. hadoop fs -ls /user/tempuser

2. mahout seqdirectory -i /user/tempuser/news_articles_v2 -o news_articles_seq_files

3. hadoop fs -ls /user/tempuser

4. hadoop fs -ls /user/tempuser/news_articles_v2

5. hadoop fs -ls /user/input

6. hadoop fs -ls /user/tempuser/news_articles_v2

7. mahout seqdirectory -i /user/tempuser/news_articles_v2 -o /user/tempuser/news_articles_seq_files

8. hadoop fs -ls /user/tempuser

9. mahout seq2sparse -nv -i /user/tempuser/news_articles_seq_files -o /user/tempuser/news_articles_vectors

10. hadoop fs -ls /user/tempuser/news_articles_vectors

11. mahout canopy -i /user/tempuser/news_articles_vectors/tf-vectors -o /user/tempuser/news_articles_vectors/news_articles_canopy_centroids org.apache.mahout.common.distance.CosineDistanceMeasure -t1 1500 -t2 2000

12. mahout canopy -i /user/tempuser/news_articles_vectors/tf-vectors -o /user/tempuser/news_articles_vectors/news_articles_canopy_centroids -t1 1500 -t2 2000

13. hadoop fs -ls /user/tempuser/news_articles_vectors

14. mahout kmeans -i /user/tempuser/news_articles_vectors/tfidf-vectors -c /user/ tempuser/news_articles_canopy_centroids -o /user/tempuser/news_articles_kmeans_clusters –clustering -cl -cd 0.1 -ow -x 20 -k 3

15. mahout kmeans -i /user/tempuser/news_articles_vectors/tfidf-vectors -c /user/ tempuser/news_articles_canopy_centroids -o /user/tempuser/news_articles_kmeans_clusters -cl -cd 0.1 -ow -x 20 -k 3

16. hadoop fs -ls /user/tempuser/news_articles_kmeans_clusters

17. hadoop fs -ls /user/tempuser/news_articles_vectors

18. mahout clusterdump -dt sequencefile -d

/user/tempuser/news_articles_vectors/dictionary.file-* -i /user/tempuser/news_articles_kmeans_clusters/clusters-6-final -o clusters2.txt -b 100 -p /user/tempuser/news_articles_kmeans_clusters/clusteredPoints -n 20

19. cat clusters2.txt