README

To run the TM_spark.py on the EMR Cluster follow the steps below -

Step 1:
upload the program file (TM_spark.py) and the data file (News_Category_Dataset_v3.json) to the aws S3 bucket and note the paths

Step 2:
On the EMR cluster, install pandas and sci-kit learn using the commands.
$ pip install pandas
$ pip install scikit-learn

Step 3:
Upload the program file and the data file to the EMR cluster

Step 4:
Run the code with the command
$ spark-submit TM_spark.py News_Category_Dataset_v3.json

Step 5:
Observe the results

The number of topics, topic words, and iterations can be changed in the code on line 250 in the function call to LDA