# Assignment:

The data files about news items and their respective social feedback on platforms are present in “data” folder.

The information about these files is provided in the file readme.txt

Using Apache Spark with any choice of language, complete the following tasks.

1. Apply following business rules on the data
   1. Rule 1 : If title contains less than 12 alphabets then remove those rows
   2. Rule 2 : If topic has any value other than following topics: ‘obama’, ‘economy’, ‘microsoft’ or ‘palestine’ then remove such rows
2. Save the IDLink values for the rows rejected by Rule 1 and Rule 2 in the files named ‘Rule1\_reject\_log.csv’ and ‘Rule2\_reject\_log.csv’ respectively.
3. Save the modified data in the file named “News\_cleansed.csv”
4. Generate a report in the text file format having following information
   1. Total number of records in the original file
   2. Total number of records after applying business rules
   3. Year-wise average Sentiment score of the text in the news items' headline for each topic
   4. Year-wise average Sentiment score of the text in the news items' title for each topic
   5. Top 10 most popular items by average Final value of the news items' popularity on Facebook, GooglePlus, LinkedIn

Refer to the file sample\_report.txt

1. Save the SOCIAL FEEDBACK DATA from GooglePlus as well as Facebook for the topic ‘economy’ for the news published in 2015 in the file named ‘economy\_social\_feedback\_2015’.

Use News.csv only for performing tasks 1,2 and 3.