PART – II

a) The current topic of interest that we chose is, FacebookDataBreach.

We extracted the data from both twitter and NYTimes using TwitterR and NYTimes API respectively.
We also extracted the data both for a longer duration and shorter duration.

b) The data that is extracted has been loaded into HDFS using commands, -put, -copyFromLocal etc.

c) Using -mkdir command in the command line different directories TwitterData, NewsData are created to store respective data.

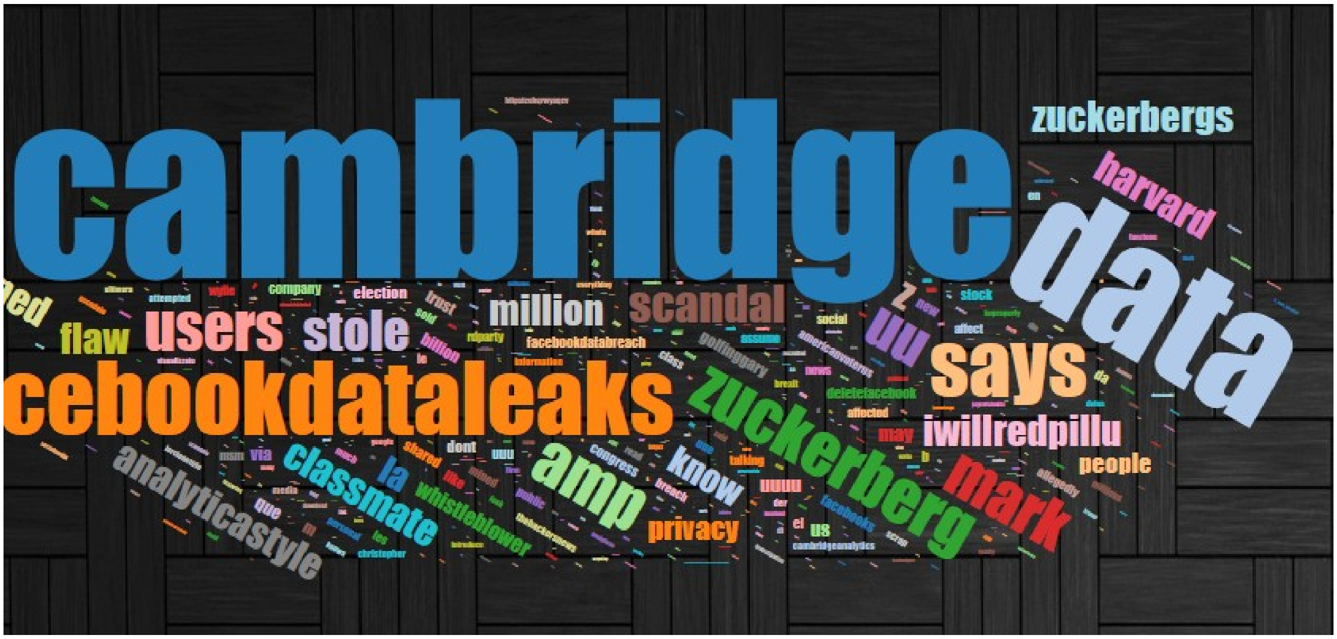
d) Using python as the language, files mapper and reducer are created. We are removing the punctuation, standard stop words as per the package nltk.

f) The same steps as above have been performed for the data that has been collected for a longer duration of times.

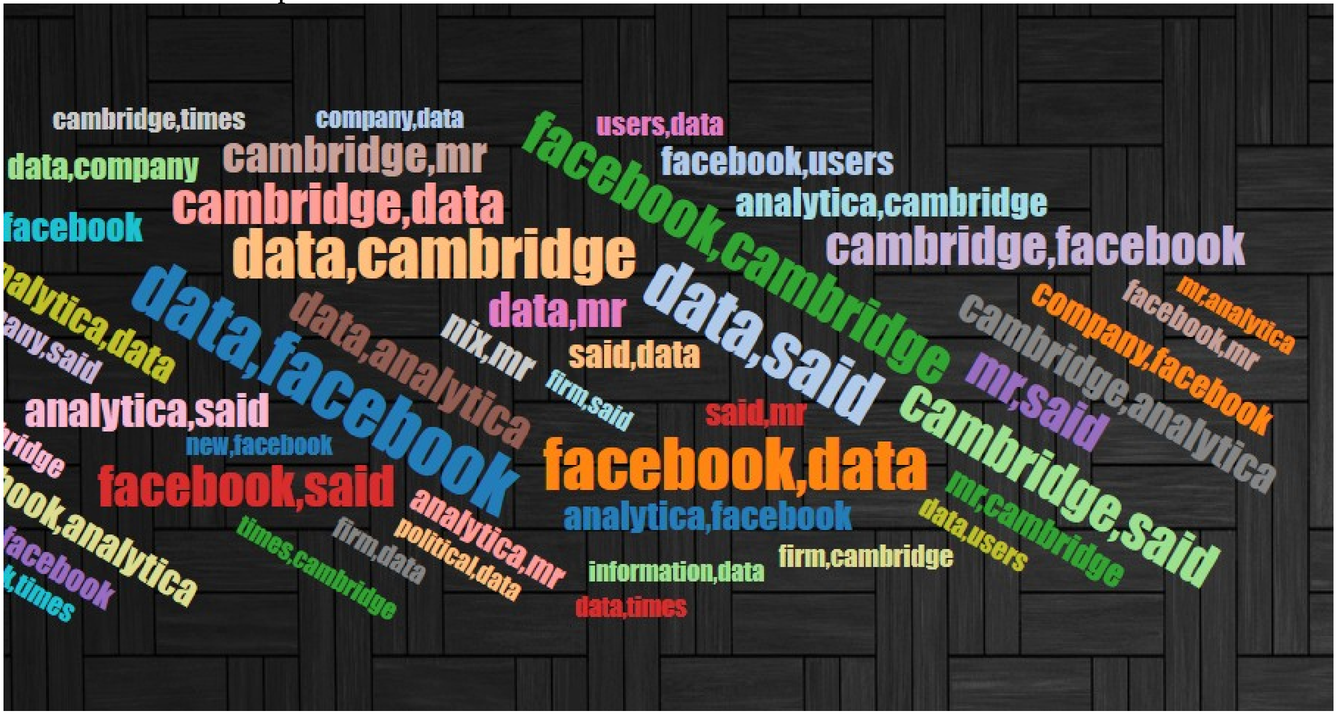
h) To analyse the co-occurance between the words from NYTimes data and Twitter data different mapper and reducer are created.
With mapper dealing with stop words and punctuation, reducer emitting the co-occurance values.

Following are the screenshots of visualizations obtained.

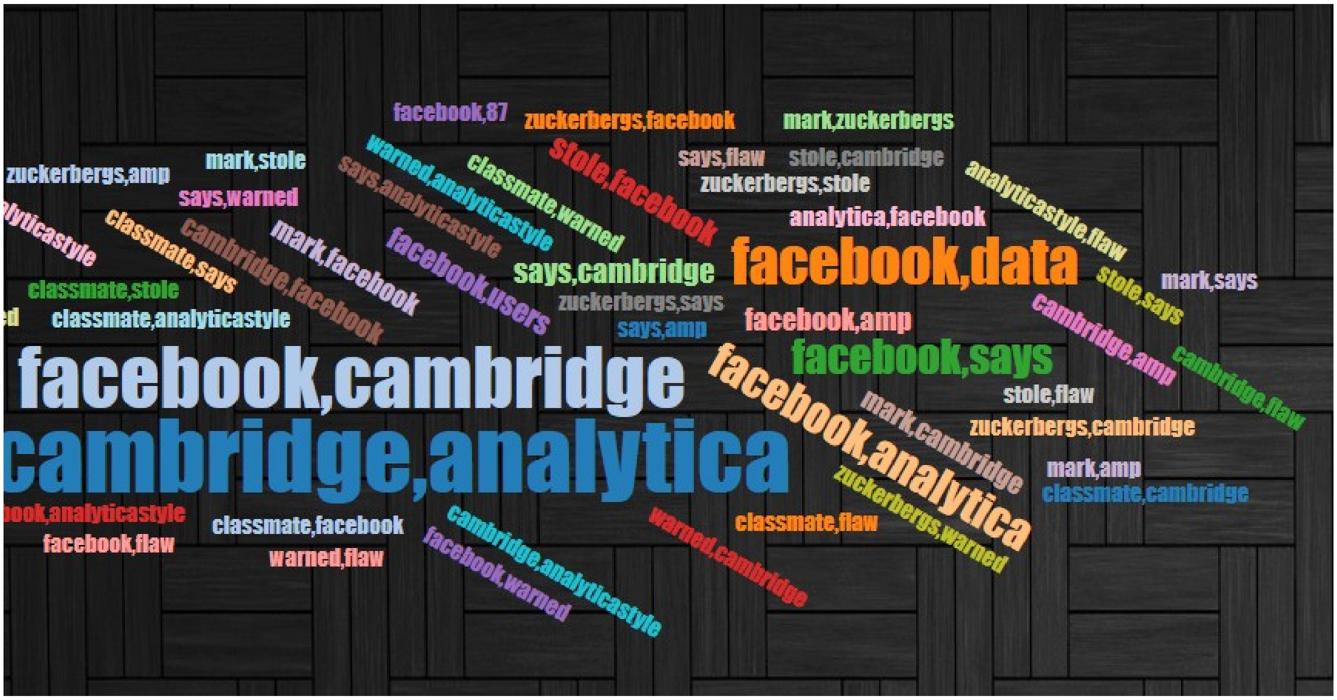
WordCloud:



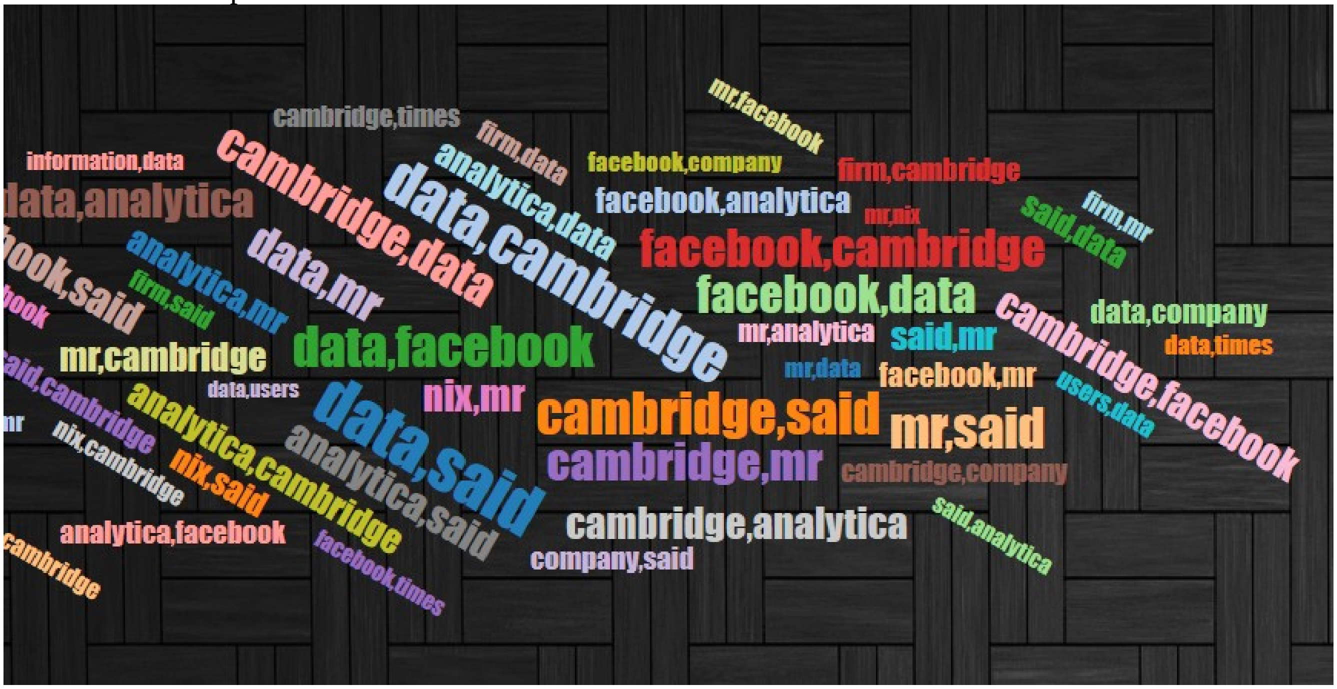
co-occurance on complete data of NYTimes:



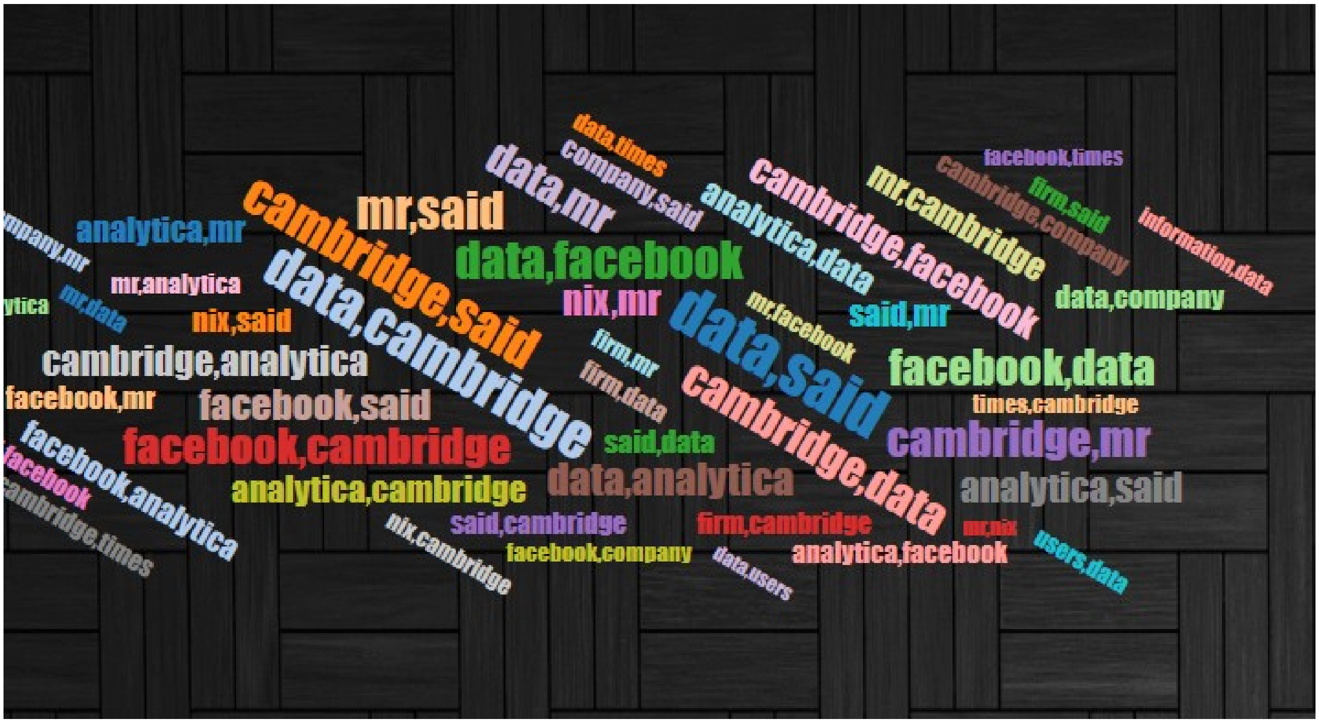
co-occurance on complete data of TwitterData:



Co-occurance on partial data from Twitter:



Co-occurance on partial data from NYTimes:


data,times
company,said
facebook,times
data,mr
cambridge,company
mr,said
analytica,data
cambridge,facebook
mr,cambridge
firm,said
information,data
mpany,mr
analytica,mr
cambridge,said
data,facebook
mr,analytica
ytica
mr,data
nix,said
nix,mr
data,cambridge
mr,facebook
said,mr
data,company
data,said
firm,mr
cambridge,analytica
firm,data
facebook,data
facebook,mr
facebook,said
cambridge,data
times,cambridge
cambridge,mr
facebook,cambridge
said,data
facebook,analytica
facebook
ambridge,times
analytica,cambridge
data,analytica
analytica,said
said,cambridge
firm,cambridge
mr,nix
nix,cambridge
facebook,company
data,users
analytica,facebook
users,data