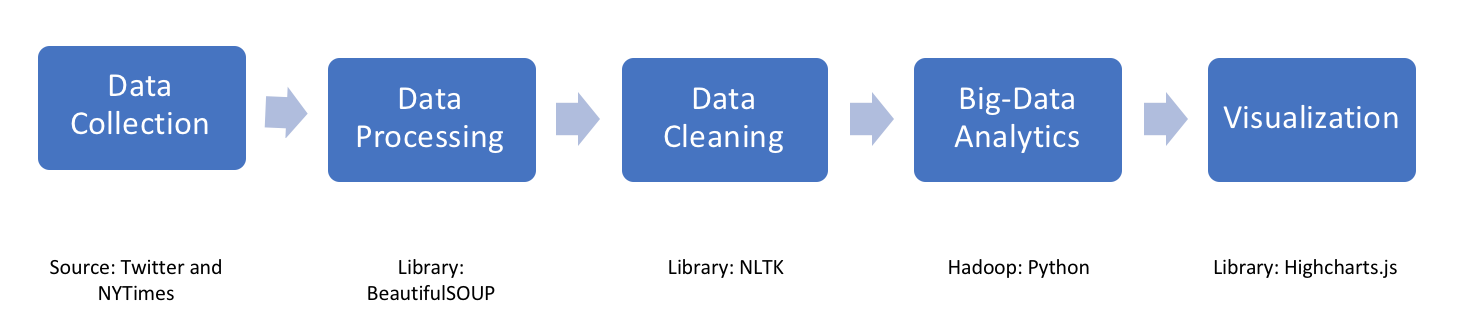
**BIG DATA ANALYSIS AND VISUALIZATION**

**Veerappan Saravanan**

**Aim:**

The goal of the project is to analyze, visualize and infer information on how the topic of interest (**blockchain**) is featured across various sources. Specific sources considered for this project are New York Times article and Twitter tweets.

**Process:**



**Data Collection and Processing: (Python and R)**

The source of data for this lab were New York times articles and tweets. A python script was used **(DataCollectionNYTimes.ipynb)** to fetch the article contents using the API provided by NYT article search and twitteR package was used to collect tweets in R language **(DataCollectionTwitter.ipynb)**. The data collected from both sources were organized into NewsData folder and TwitterData folder respectively and forms the input for the mapreduce.

**Data Cleaning (Inside Mapper):**

The raw data collected had many undesirable words/ characters that needed to be filtered out. NLTK library in python was used to assist with the filtering process.

1. Tokenized the input corpse
2. Removed non-alpha numeric characters
3. Removed punctuation
4. Removed stop words
5. Stemmed the words using Potter stemmer

**Big-Data Analytics:**

**Word count: (mapper.py and reducer.py)**

The filtering process happens in the mapper. The mapper picks the words of interest and emits its count.

The reducer then gets the output of the mapper, and performs the aggregation (sum of each keyword) and returns the output

Command to run:

After starting hadoop, and placing the data into input fil run the following command to start the map reduce

hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input input -output output

**Word co-occurrence: (mapper-twitter.py, mapper-news.py and reducer.py)**

The filtering process happens in the mapper. The mapper picks the words of interest and emits its count. The output of mapper is different from the previous case. Here it emits as **pairs and count** of the co-occurring words

The reducer then gets the output of the mapper and performs the aggregation similar to the word count.

Command to run:

After starting hadoop, and placing the data into input fil run the following command to start the map reduce

hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar -file mapper\_twitter.py -mapper mapper\_twitter.py -file reducer.py -reducer reducer.py -input input -output output

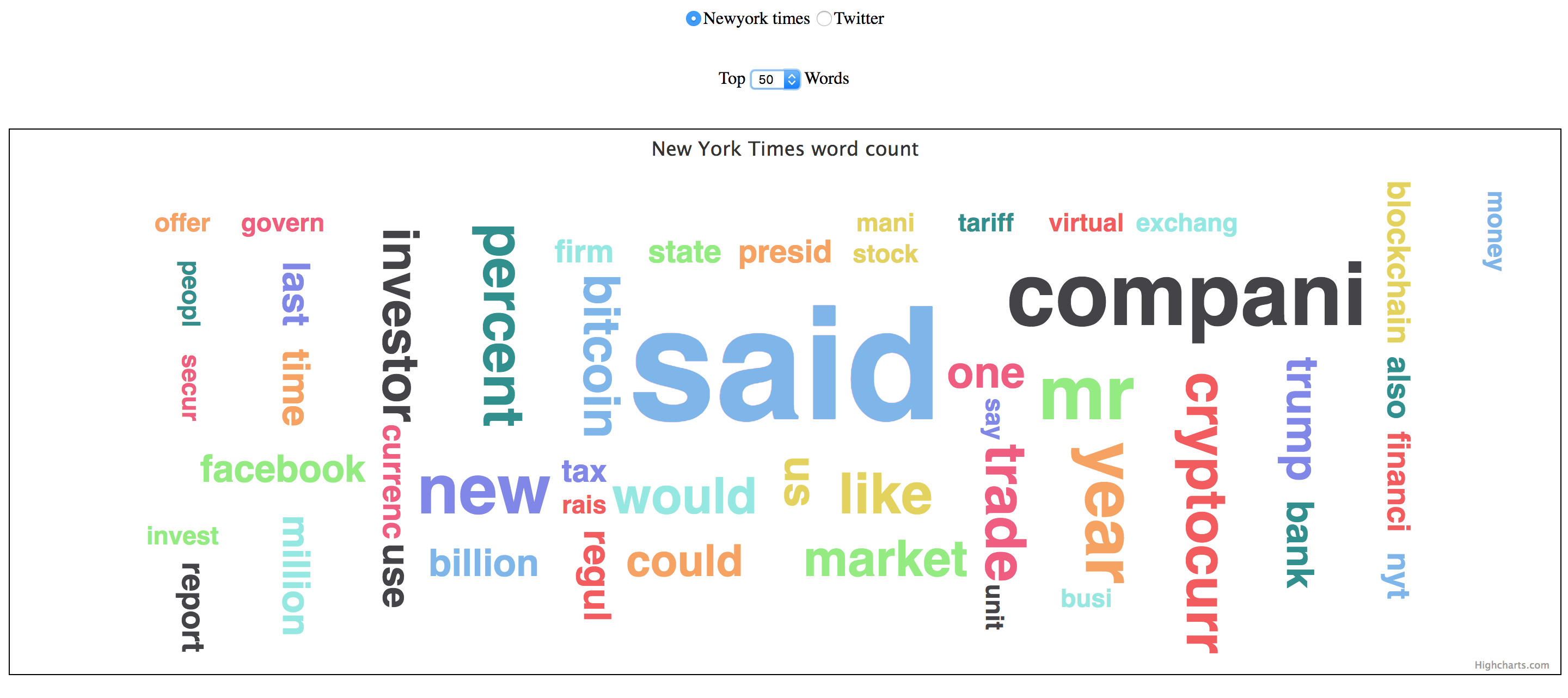
hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar -file mapper\_news.py -mapper mapper\_news.py -file reducer.py -reducer reducer.py -input input -output output

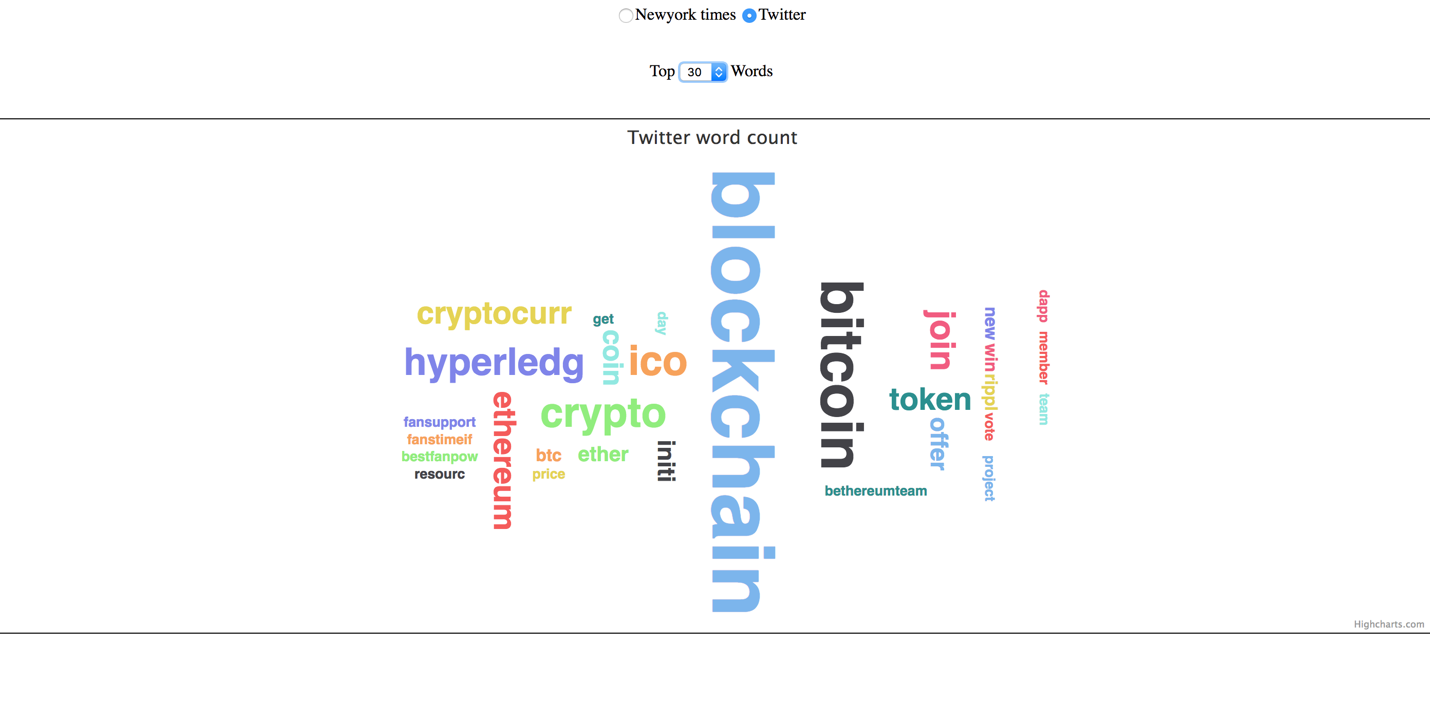
**Processing the output: (Python)**

The output of the reducers is processed to get the top counts and fed as input for visualization. This is done using python script **(DataVisualization.ipynb)**

**Visualization: (highchart.js)**

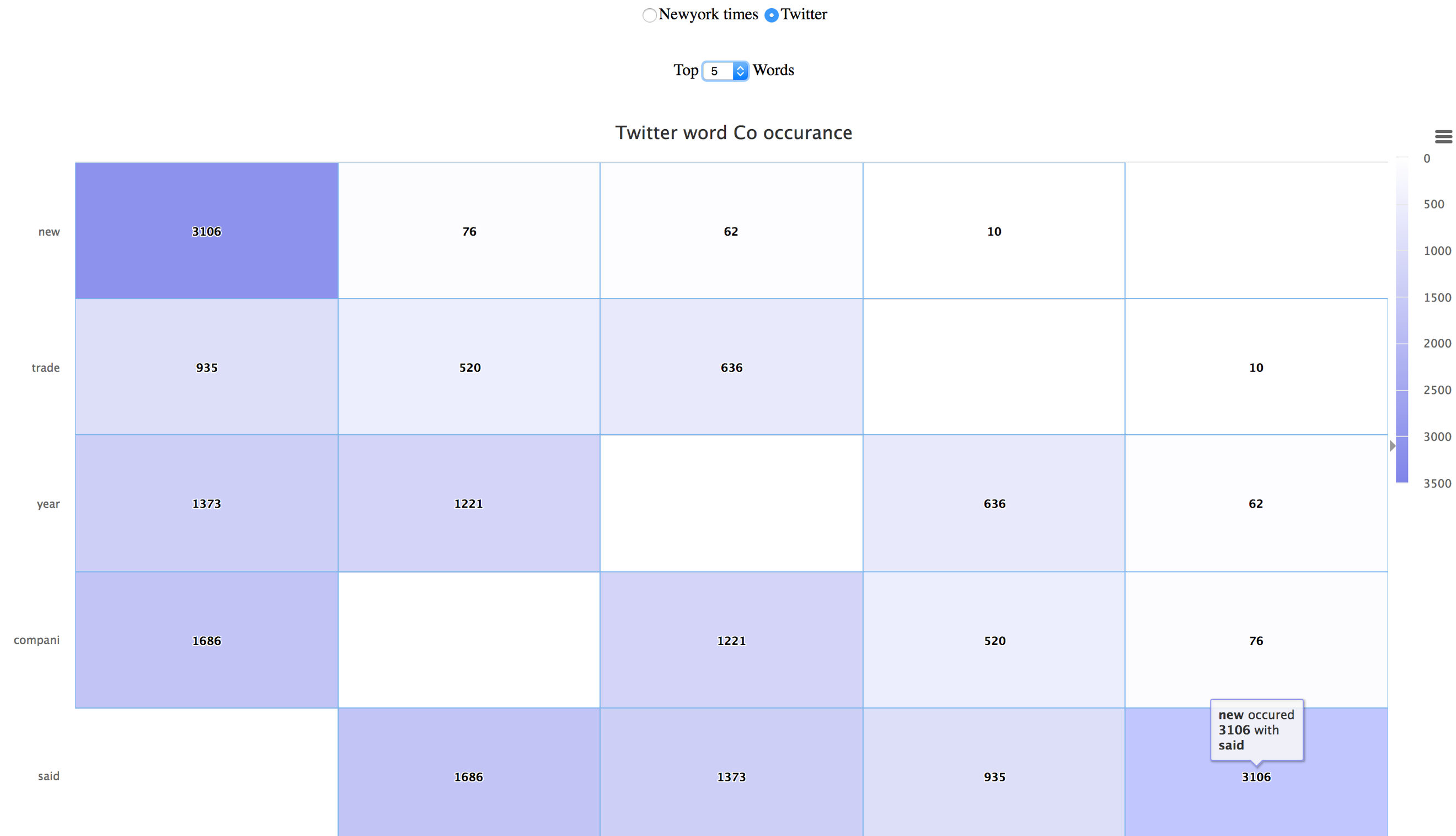
1. Word cloud for Word Count: An interactive website to visualize the data processed. The user can select the source of data (NY times and twitter) and can also specify the top of N words of interest. This will show a word cloud visualization of the selected parameters. The visualization speaks for itself.





1. Heat Map for Co occurrence: The data for co-occurrence is visualized as a heat map. Similar to the word count, the user can select the source and top N of words.





**Keywords (NYT and Twitter):**

1. Blockchain
2. Crypto
3. Bitcoin
4. Ethereum
5. Ether
6. dApp
7. Initial Coin Offering
8. R3 corda
9. Hyperledger
10. Litecoin

**Inference:**

**The data collected from NY Times:**

Prominent words occurring are **said, Bitcoin, Blockchain, investor, company, new, market, trade, cryptocurrency**. New York times articles are professional and well edited when published, that’s why one can see words like company, investors, new, market etc. These things are not prominent in tweets

In NYT articles, blockchain and its associated components are seen as **new technology** that is grabbing the attention of the **investors** and **market**. There has been a lot **said** about the technology components. We can see that words such as **would**, **could** and **like** which symbolizes the **unpredictability** of blockchain technology

**The data collected from Twitter:**

The data collected from twitter whereas has more technical terms associated with it like **bitcoin**, **ethereum**, **hyperledger**, **ico**. The trending components such as bitcoin and hyperledger have been more prominent, People have tweeted more about the technology and its components, especially about trading by means of **btc**, **coin**, **ether**, **xrp**.

To sum up, the tweets have all been about the technology and its various components and the new article is more about what effect the technology has in the real-world market.