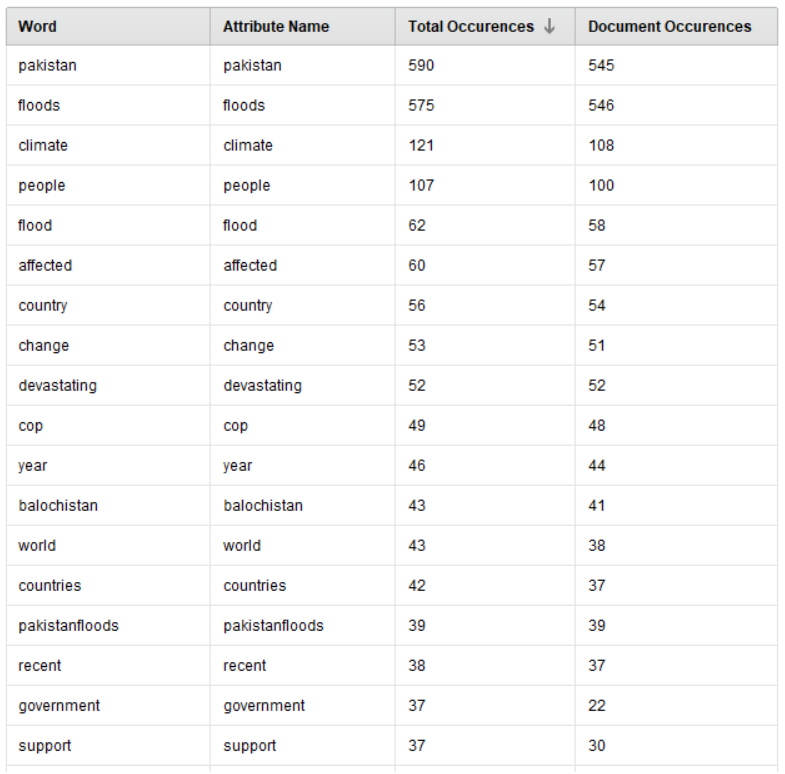
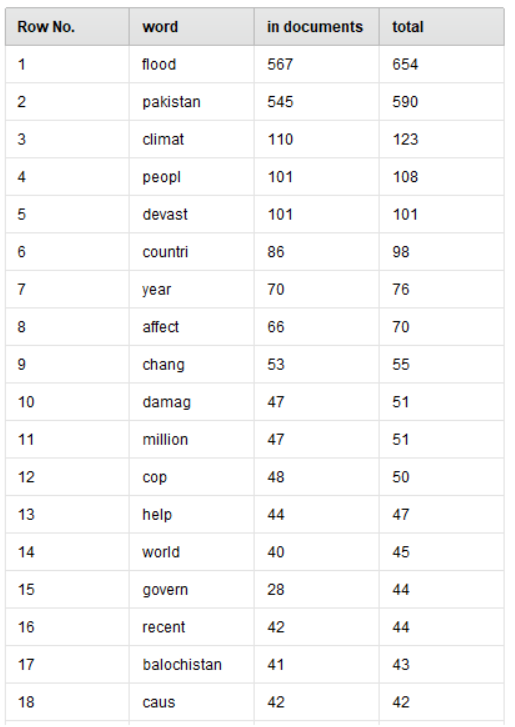
**Project 5810 documents**

This list of words is the output of process after tokenizing the tweets data. It shows the name of the words and its total occurrence in the all tweets and the number of documents in which this particular word occurred. After seeing this data, we got to know that we have redundancy of words such as floods and flood both are same words but here considered as different words.

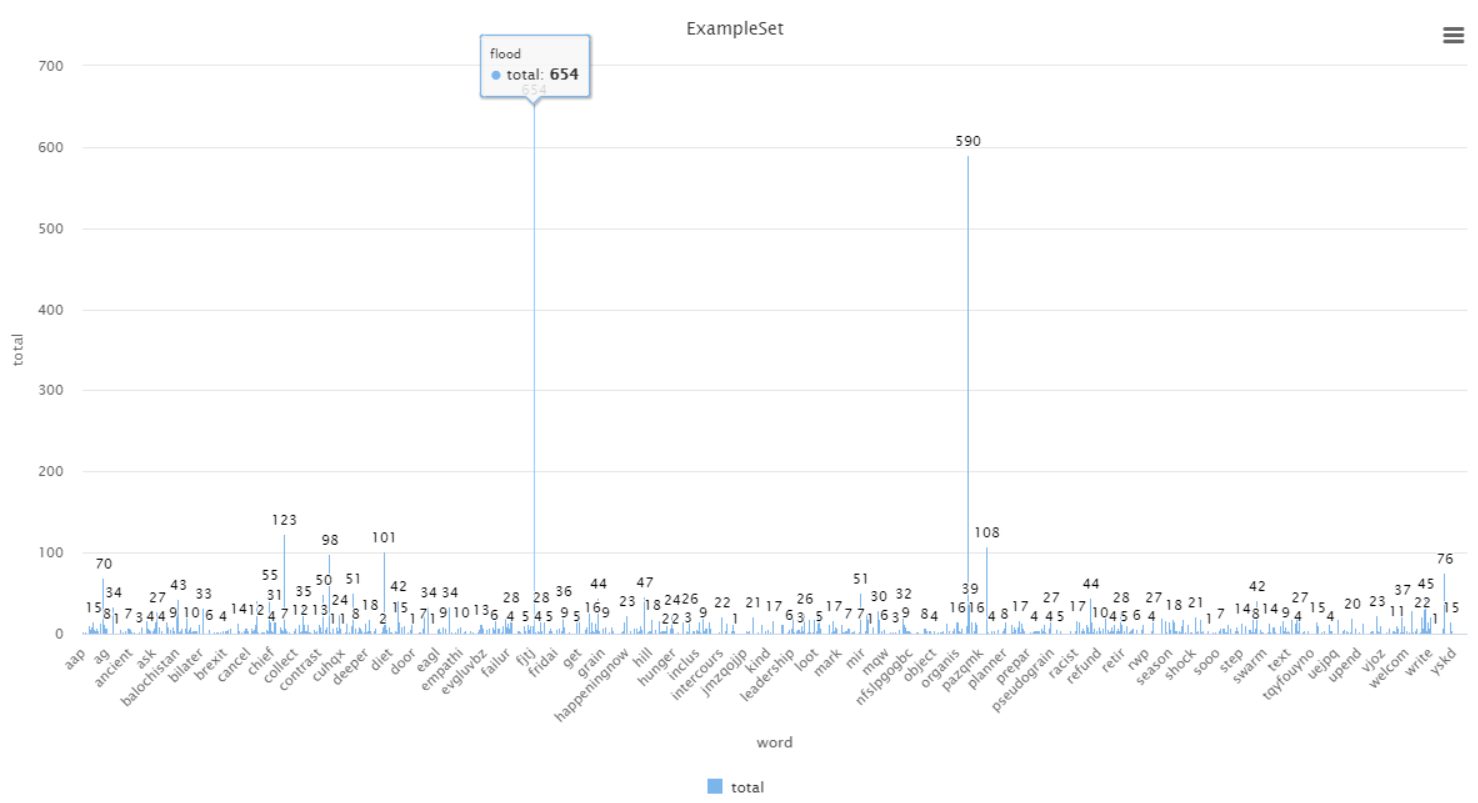
To tackle this issue, we use the “Stem” operator which will consider the same words that have same root.

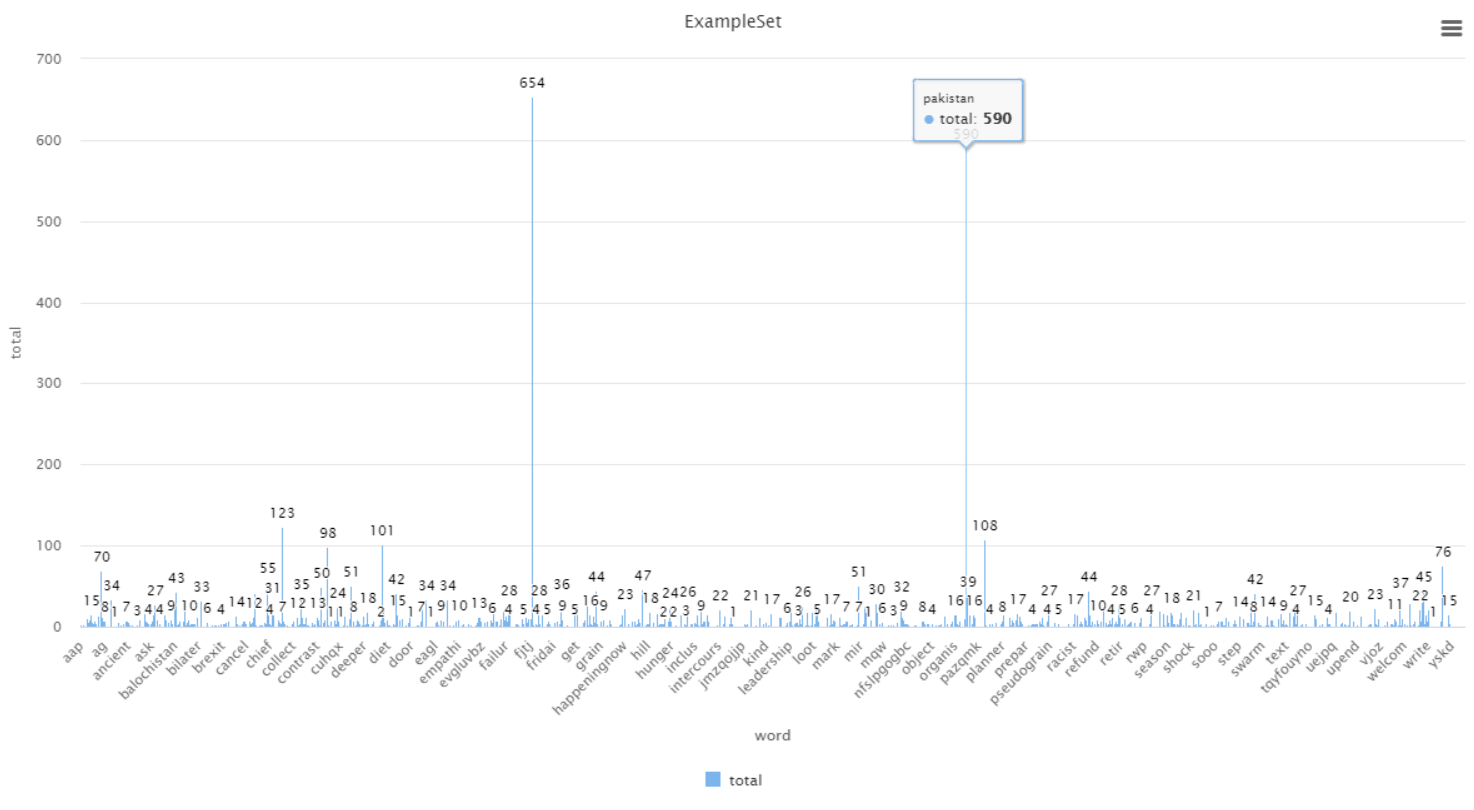


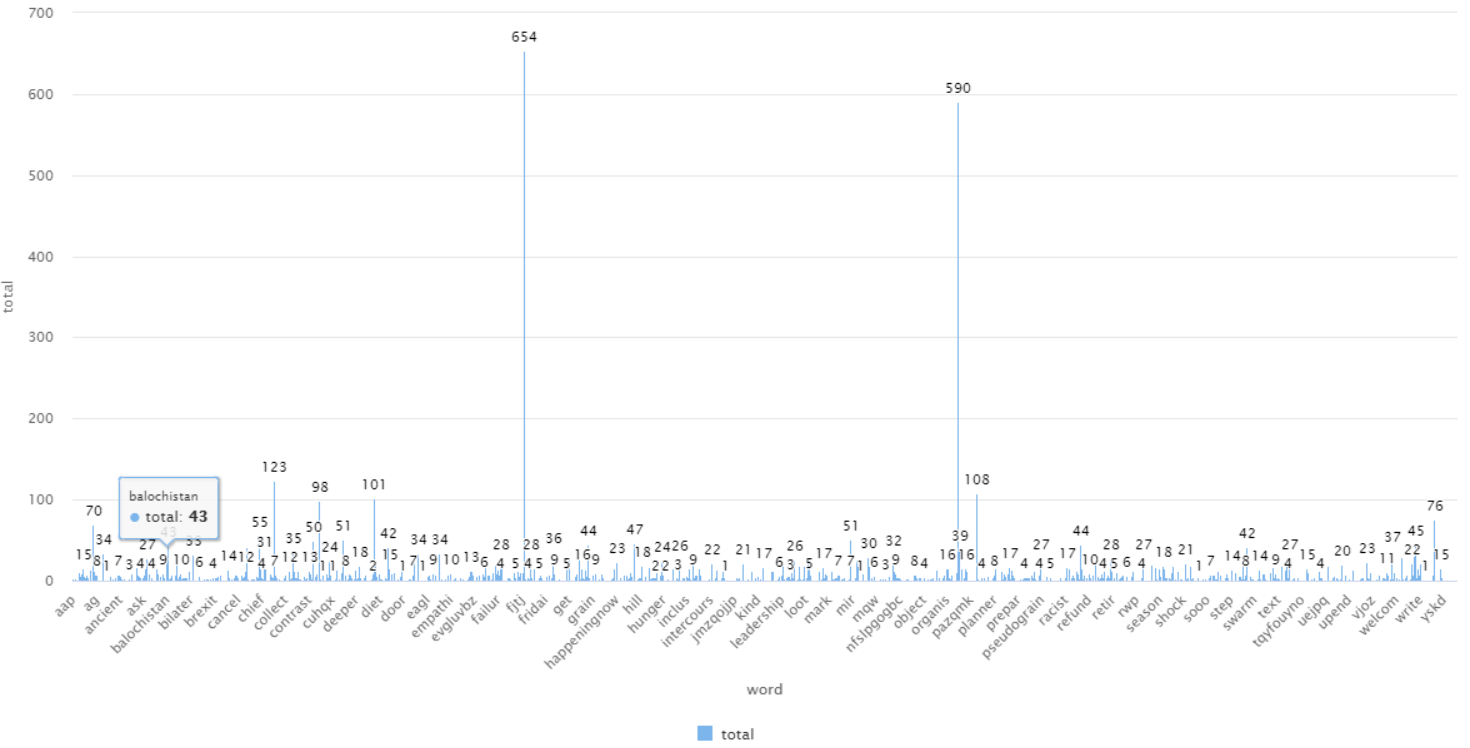
After doing the stemming means like different words with same root count as same words such as floods and flood count same.

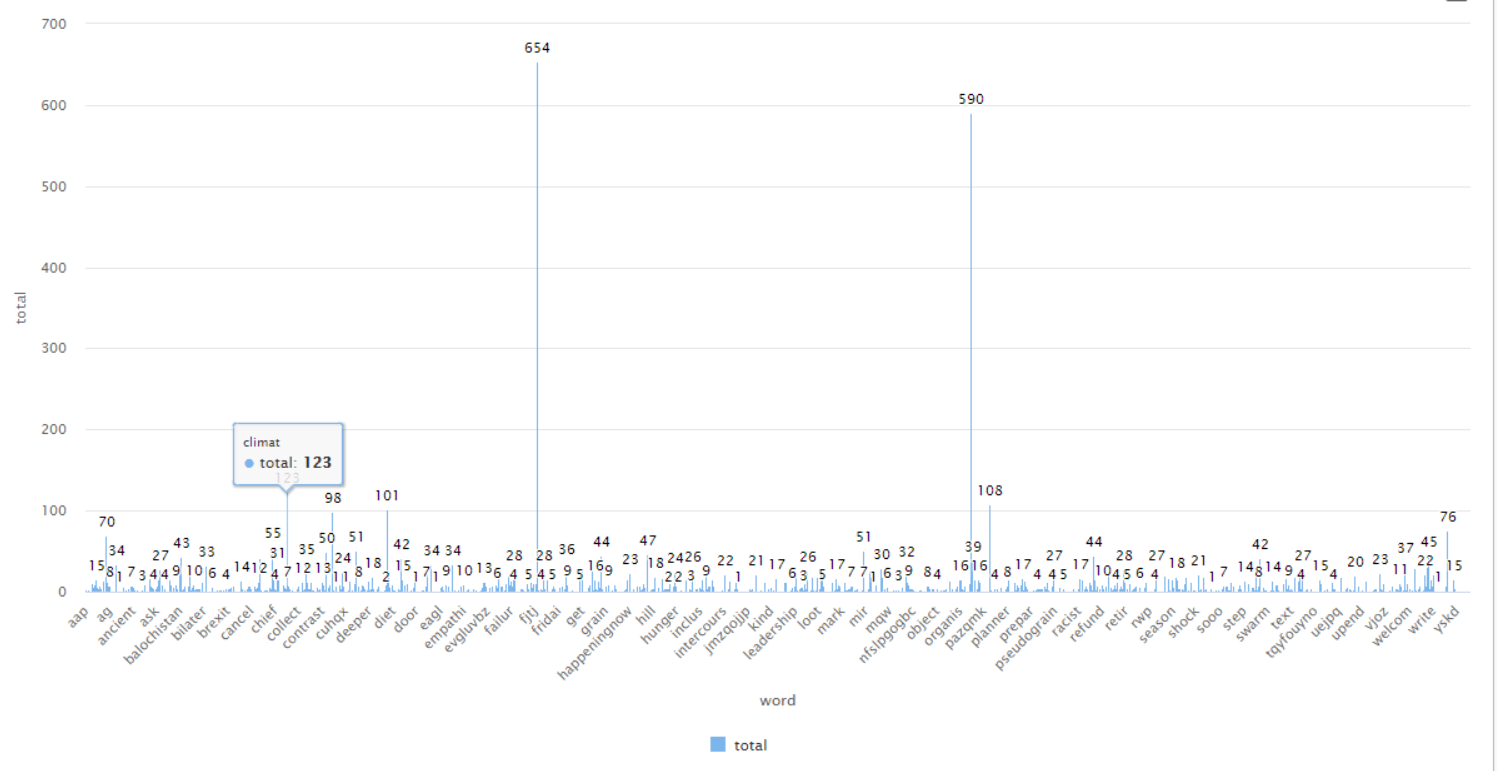


Showing the Bar graph between the words and the total occurrence of the word in the all tweets. From the graph, we also can see that flood word has the most occurrence in the whole tweets data and after the flood word, Pakistan word is used on the second place.

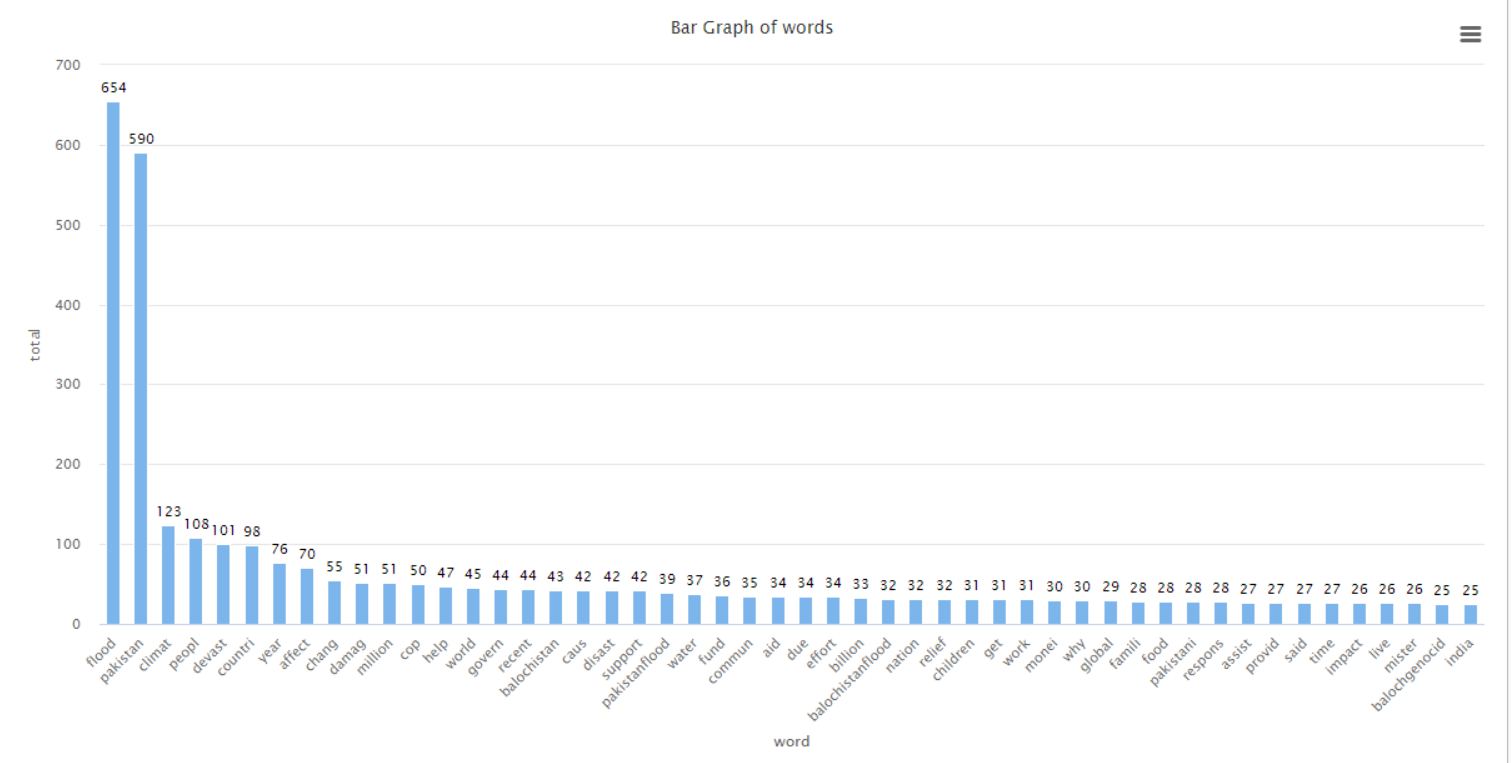








Top 50 words.

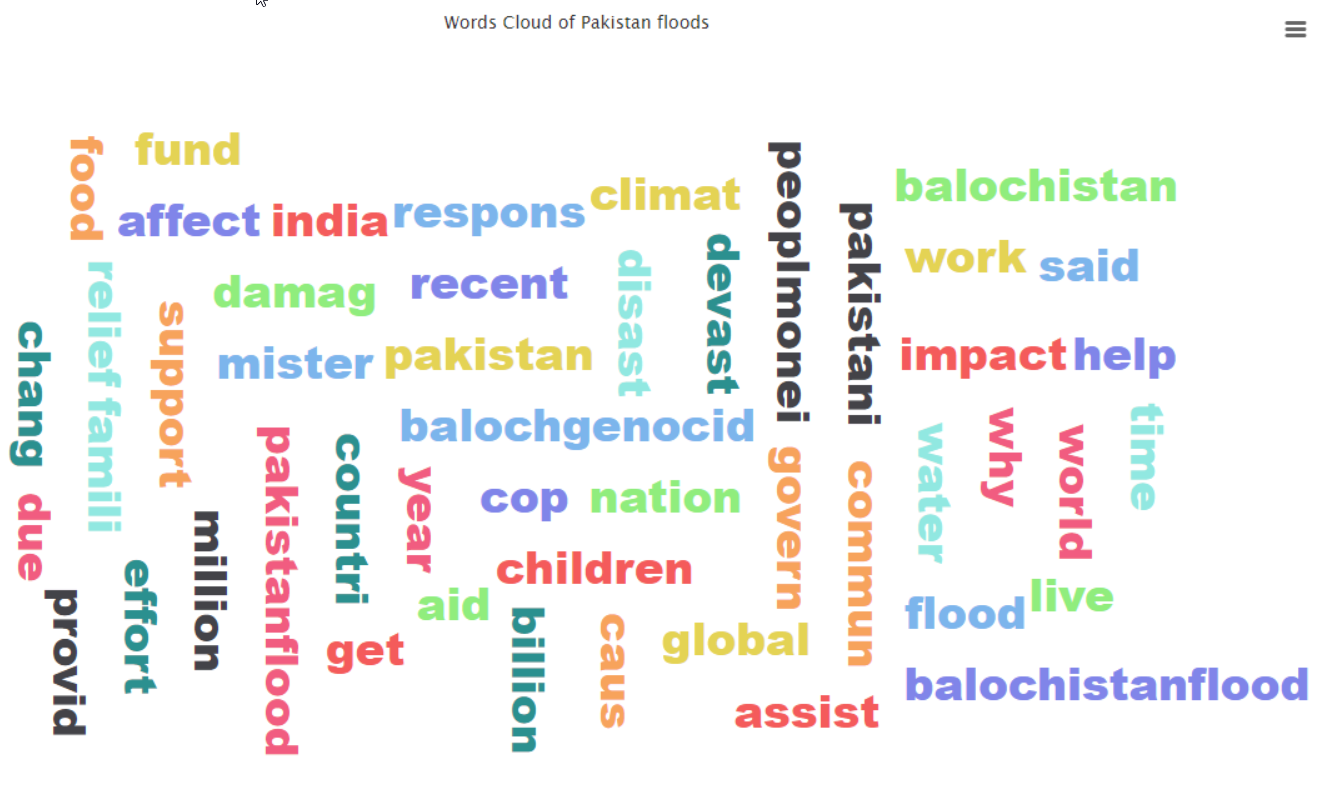


From the top 50 words, Balochistan words comes at 17 place from the top. Balochistan is the state of Pakistan and as per multi-time occurrence of this word, it means this state is affected more by floods as compare to the other states.

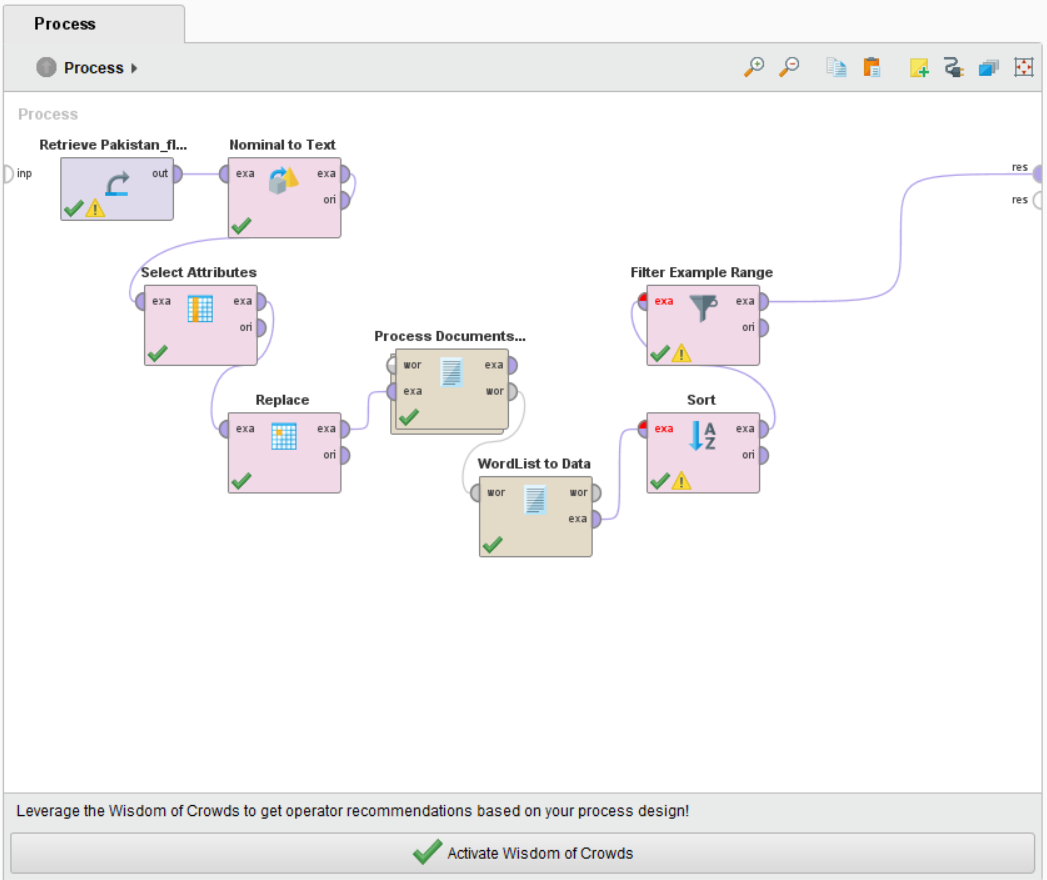
Climate word comes at 3rd position from the top, so climate change is the big cause of the floods and it is also a big concern for the people of the country.

There is one more word “balochgenocid” which has more occurrence in the tweets, from the word, we can say that there are lots of casualty (deaths) found in the state of Balochistan in the Pakistan.

Word cloud of the top 50 words from the Pakistan floods data from the twitter.

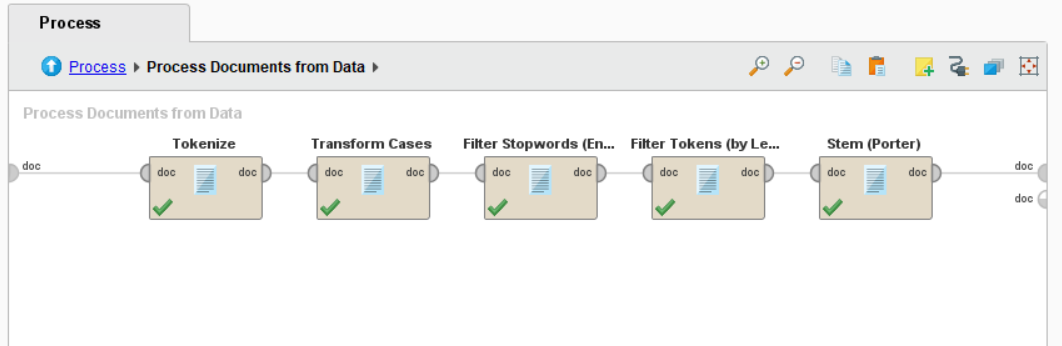


Process Diagram for Data analysis.



Here, we are using the cleaned data and we need to convert the data, for converting the data, we use nominal to text operator. Now, we ned to select the attributes from our data set, we have three attributes in our data set. First attribute is “att1”, second attribute is “CLEAN\_tweets”, and the third attribute is “likes”. So by selecting the attribute filter type “subset” then select only “CLEAN\_tweets”, and “likes”. Now, we used the replace operator and replace the “https” words with white space.

Now, we need to process the tweets data, for processing the tweets, we use the process documents from data operator but this operator is not available in rapider miner by default. For this operator, we go to the extension and click on the market place and search for operator tool box. After searching, click and install it and restart the laptop.



After updating or installing the operator toolbox, we select the process documents from data operator and drag into the process pan. Now, double click on this operator, a new empty process pan will open. For processing the documents from the data, we select the tokenize operator. This operator tokenize the words means it separates the words on the basis of space. After tokenizing the documents, we found that it counts the capitalized and small alphabets words differently. So, to remove this difference and consider the same word, use operator “Transform cases” from the operator.

Now, we need to remove the stop words such as the, a, in, for etc. from the words list. To remove these stop words, we use the “Filter Stopwords (English)” from the operator.

After processing the process documents from data, we found some words and the do not have useful meaning such as aa, aaa, aaaaa, t etc. To handle this issue or remove these meaning less words from the word list, we use the “Filter Tokens (by Length)” and we select the minimum length 3 and the maximum length 25.

We have words list in our hand, now, we need to check the occurrence of the words. For checking the total occurrence of a single word, we use the “WordList to Data” operator. This operator created a column in which it shown the total count of a words and the number of documents in which that word occurred.

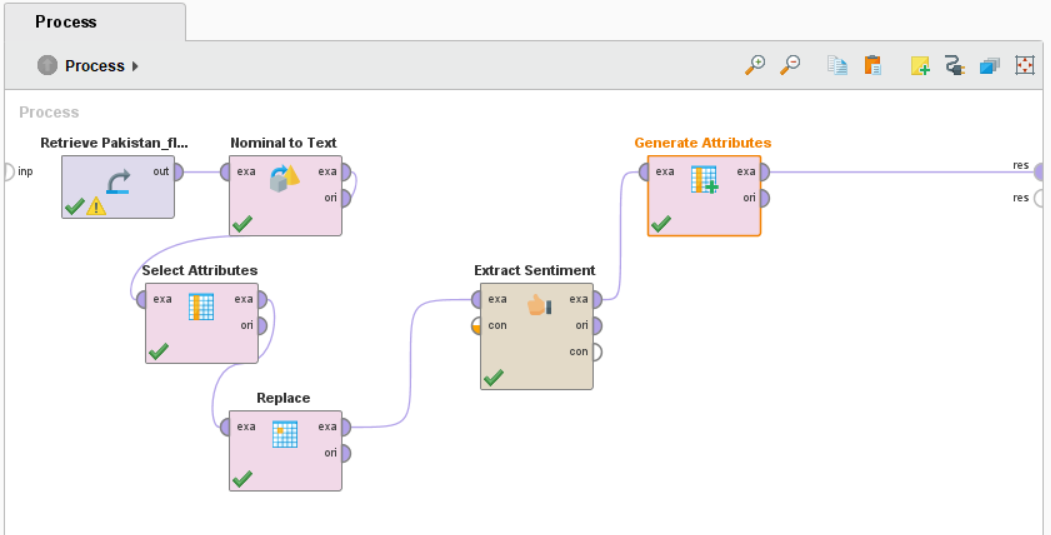
If we want to check the top 50 words that have high occurrence in the tweets, we need to use the “Sort” operator to select the attribute name “total” because total attribute has the total occurrence of a single word and sorting direction is “decreasing” that show the highest occurrence word first and second highest word at second position and so on.

For creating the words cloud, we need to select some words like top 50 high occurrence word from the total attribute, for this purpose, we need to use the “Filter Example Range” with this operator we can select top 50 words and make the words cloud using these words. When we created the words clouds from the high occurrence words, we found that there are some words that have same meaning or we can say plural or singular forms of a word. To overcome this issue, we use the “Stem (Porter)” operator and this operator consider the same word for singular or plural form of the word.

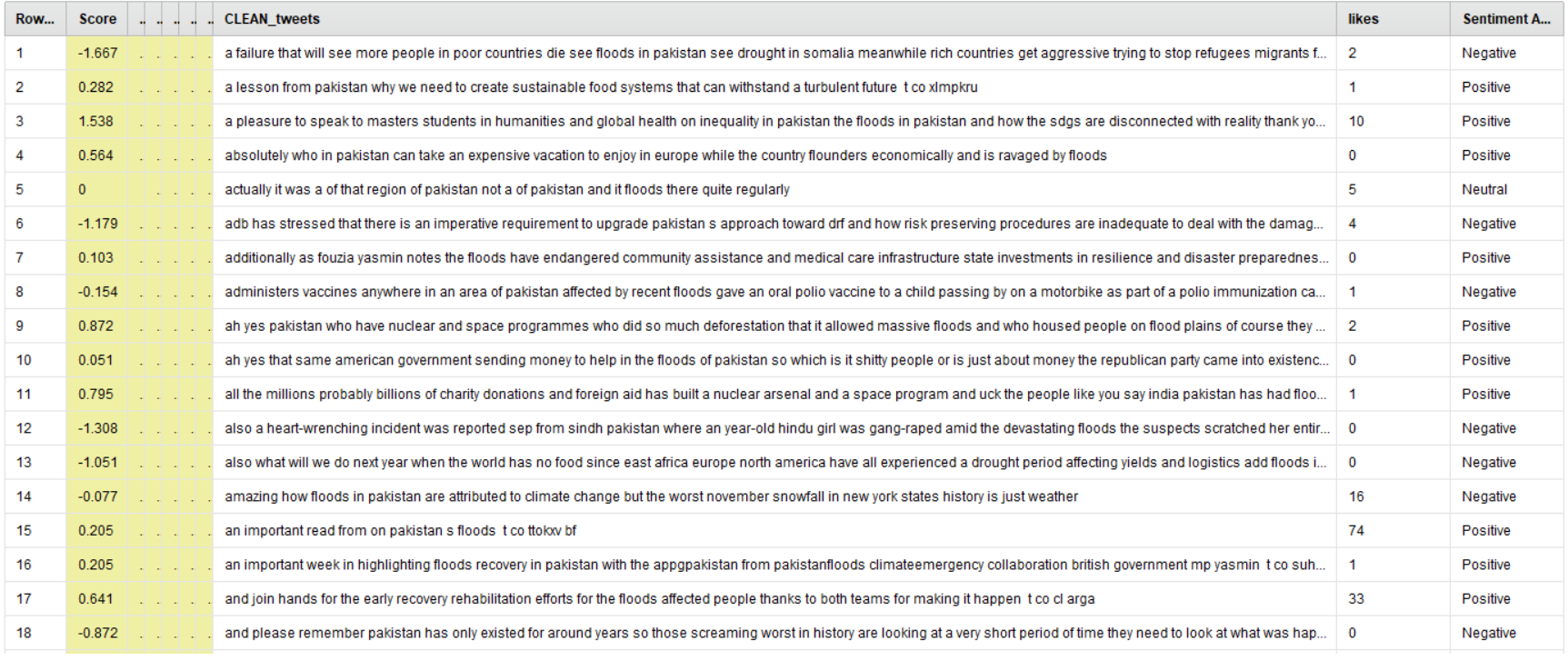
Process under the Process Documents from Data for data analysis.

Sentiment Analysis:

Process for Sentiment Analysis of the data.



Output of the sentiment Analysis process.



Saving excel (xlsx) file from the rapid mainer process into our local machine.

