Conclusion

Classification of Twitter Tweets using NLTK Library

Twitter tweets were taken from the internet. Based on studies done on the tweets we extracted major keywords which are related to the charity. Some of them are *charity*, *donation*, *fund* etc.

Tools and Libraries used

1. Python NLTK Library: To extract the text related to the keywords. NLTK(Natural Language Toolkit) is one of the most popular libraries used in text mining.
2. Python Pandas: One of the powerful libraries to play with the data frames. This library was used to load the CSV (Comma Separated Values) to the python object.
3. Matplotlib: Matplotlib is used to plot the charts.
4. Jupyter Notebook / Google Colab: As an interactive editor.

Steps we followed

1. Install the NLTK package on Colab (or any python interpreter)
2. Load the required libraries onto the notebook – Pandas, Matplotlib and NLTK
3. Mounting the drive on which the data set exist (In Google Colab we mounted Google Drive). Load the data sets – Three Data set containing the tweets and One Data series containing keywords.
4. We have to tokenize the paragraph to sentence then to words. We can use the NLTK's built in function to tokenize. We use NLTK Wordnet to tokenize the sentence to words.
5. Then in the next step we remove the common stop words from the tokenized words. The stop words. This helps us to play with much smaller data set by removing unwanted words. Hence reducing the computational cost
6. We have to remove the multiple forms words by using stemming. Using the PoterStemmer to stem the keywords and the text column in the tweet data frame.
7. Then using the apply function we construct a custom function to check whether any keyword exist in the tokenized tweet. We will store the tweets returned by the function as true in a new pandas object.
8. For better understanding we count the frequency and plot the bar chart of keyword versus count in the given corpora.
9. The categorized result is now in pandas object. As a next step we will write to a file as CSV or Excel.

Detailed source code and data set is available in the following link.
https://github.com/dkowsikpai/TwitterTweets