ClusTop: A Clustering-based Topic Modelling Algorithm for Twitter using Word Networks

In social media many peoples will share their reviews and ideas on any matter and large amount of data will be gather and it’s become mandatory to identify topic from it. Many existing algorithms such as LDA will perform well but it require lots of predefine variable and needs to train with known topics. To overcome from this problem author of this paper has introduce ClusTop algorithm which is based on word UNIGRAMS and this technique does not require known topics or any predefine values. Author has introduce UNIGRAM algorithm in many variants such as ClusTop-Word-NA (based on coword), ClusTop-BiG-NA (based on bigram), ClusTop-TriG-NA (based on trigram) and many more. Implementing all this algorithms with different variants is bit difficult so we have implemented

1. ClusTop-Word-NA (based on coword)
2. ClusTop-BiG-NA (based on bigram)
3. LDA modelling

For all variants author has implemented 3 methods

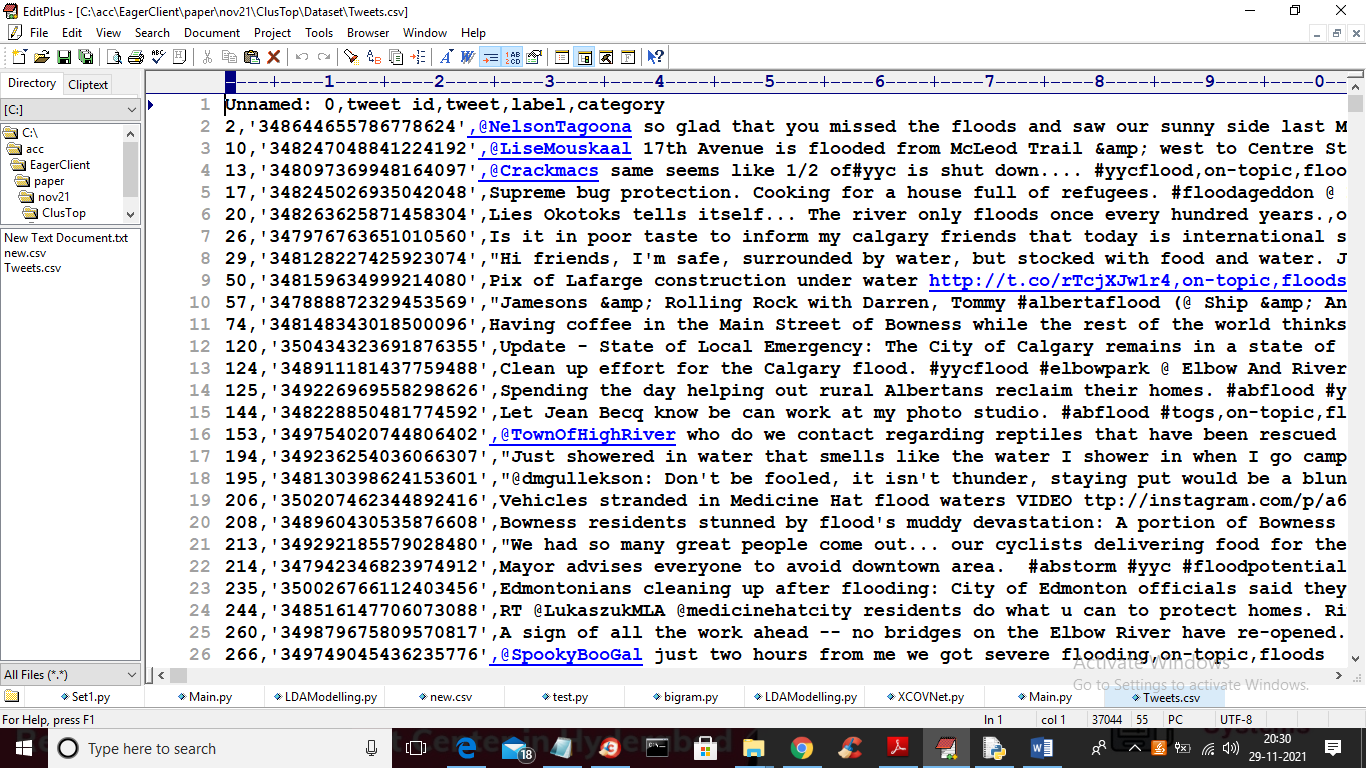
1. Network Construction: In this method we will extract all words from a tweet and then build a undirected graph where each word act like a VERTEX and then we calculate weight based on word occurrence and then if two words co-occur or having relation then both word will be connected with edges
2. Community Detection: word with high weight in the graph will be elected as topic
3. Topic Assignment: in this method detected community or topic will be assigned to tweets

Algorithm details:

1. ClusTop-Word-NA (based on coword): in this algorithm we will scan tweet to find words which occur two or more number of times and then build a graph with weight values based on occurrence and then extract a word as topic with highest weight value
2. ClusTop-BiG-NA (based on bigram): this algorithm is similar to first algorithm but before calculating co-words algorithm will convert tweets into BIGRAM and then build a graph with BIGRAM words and then extract word with highest weight
3. LDA modelling: this is python built in algorithm which will take all tweets as input and then find topics from it

Based on predicted topics and labelled data we can calculate precision, recall and FSCORE and PMI and TC will be calculated based on word co-occurrence

We have used tweets events (earthquake, flood and bombing etc.) dataset to trained above 3 algorithms and below is the algorithm screen shots



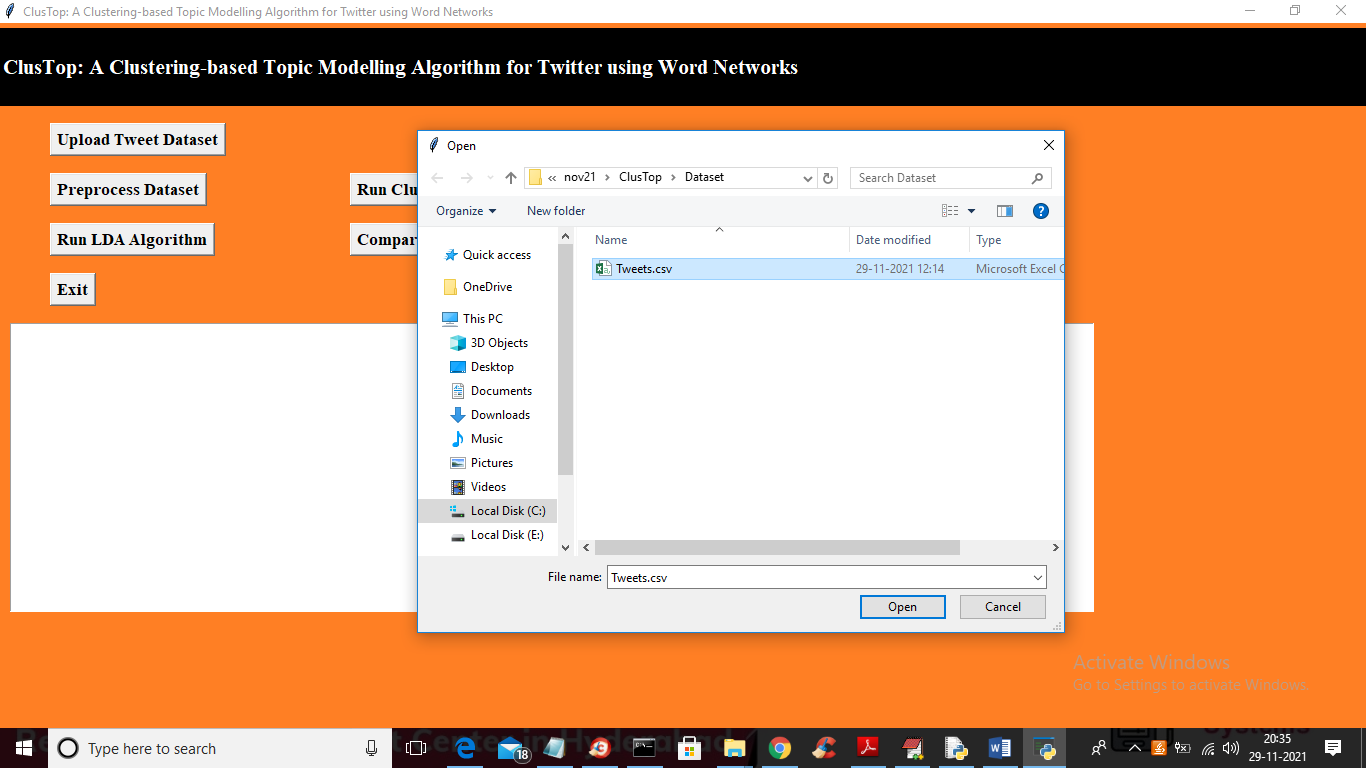
In above screen first row represents dataset column names and remaining rows represents dataset values and we are using above dataset to implement and test algorithm.

SCREEN SHOTS

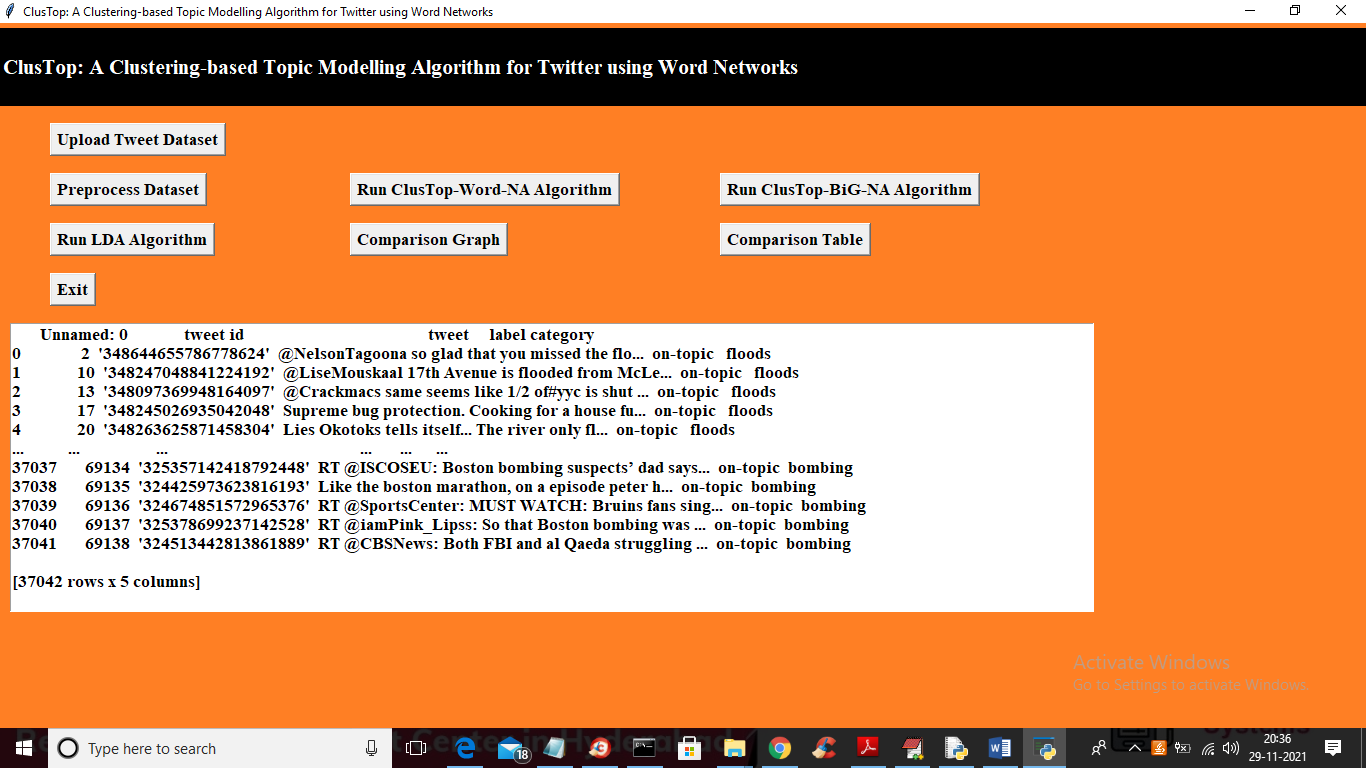
To run project double click on ‘run.bat’ file to get below screen



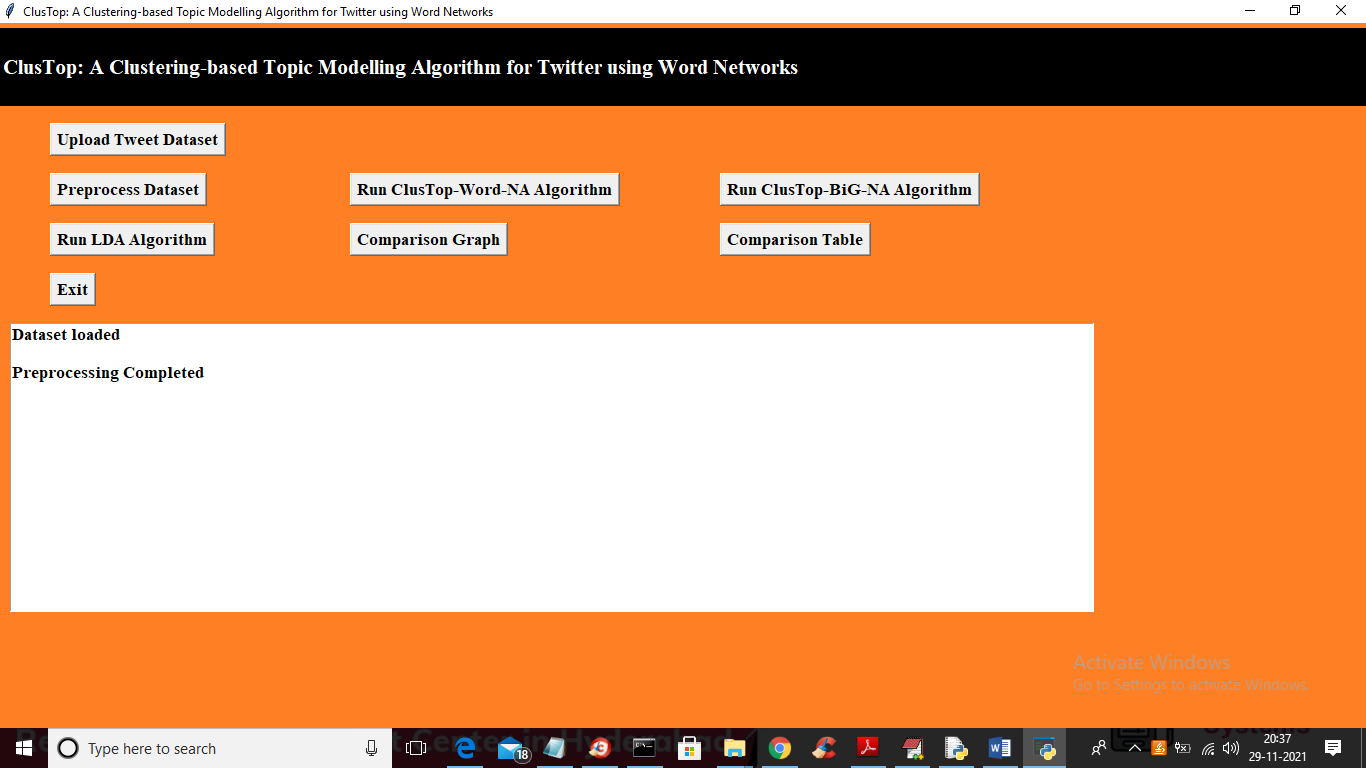
In above screen click on ‘Upload Tweet Dataset’ button to upload dataset and to get below screen and to upload dataset



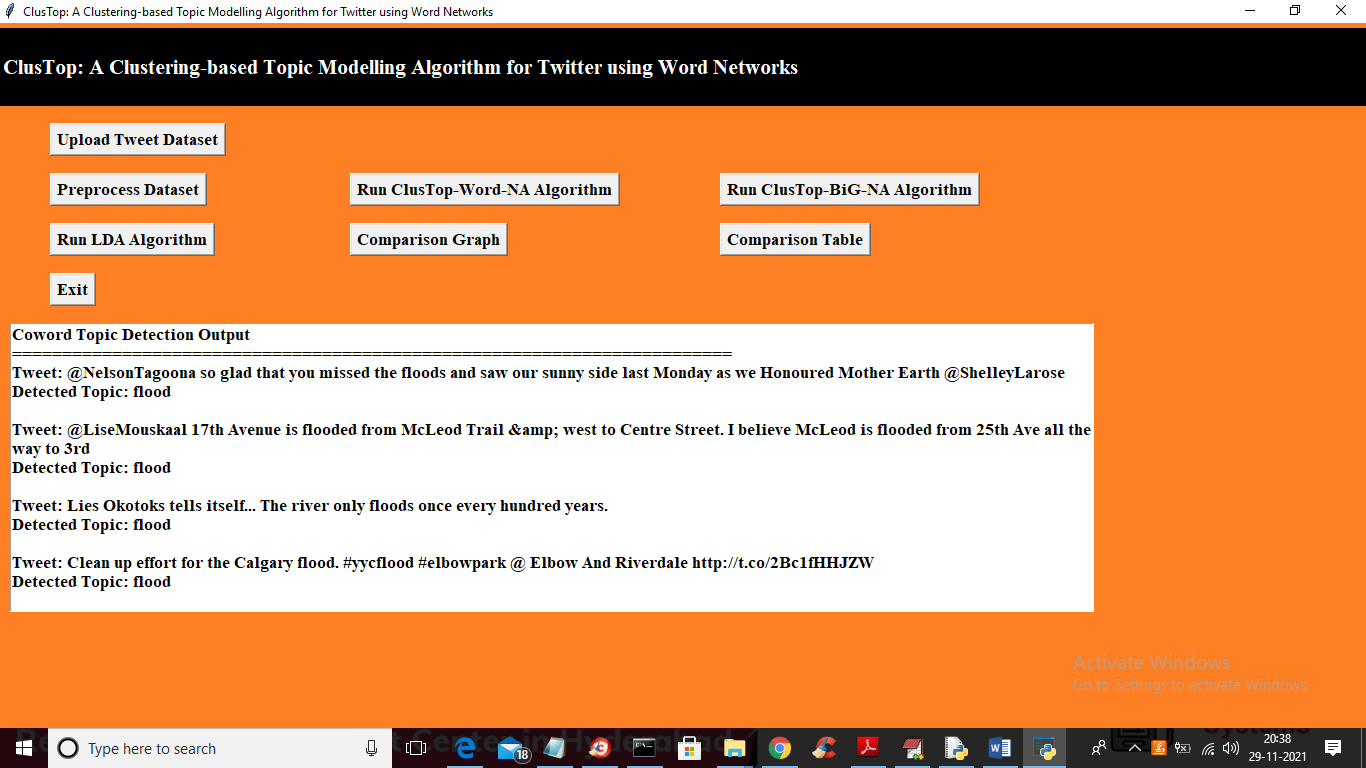
In above screen selecting and uploading ‘Tweets.csv’ file and then click on ‘Open’ button to load dataset and to get below screen



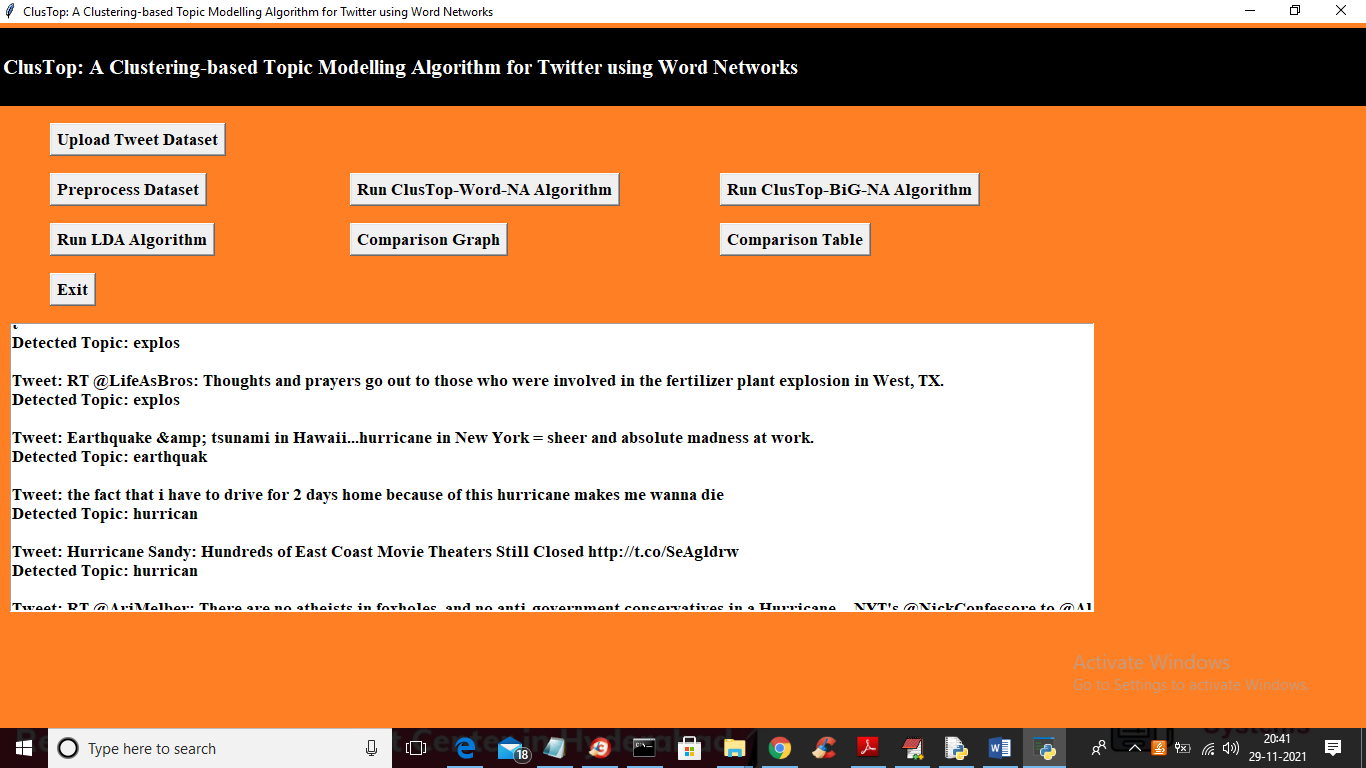
In above screen dataset uploaded and I am showing few tweets from dataset and in above dataset values we can see it contains lots of special symbols and stop words such as “and the in where” and we need to remove such words to clean tweets so we can perform accurate topic detection so click on ‘Preprocess Dataset’ button to clean tweets



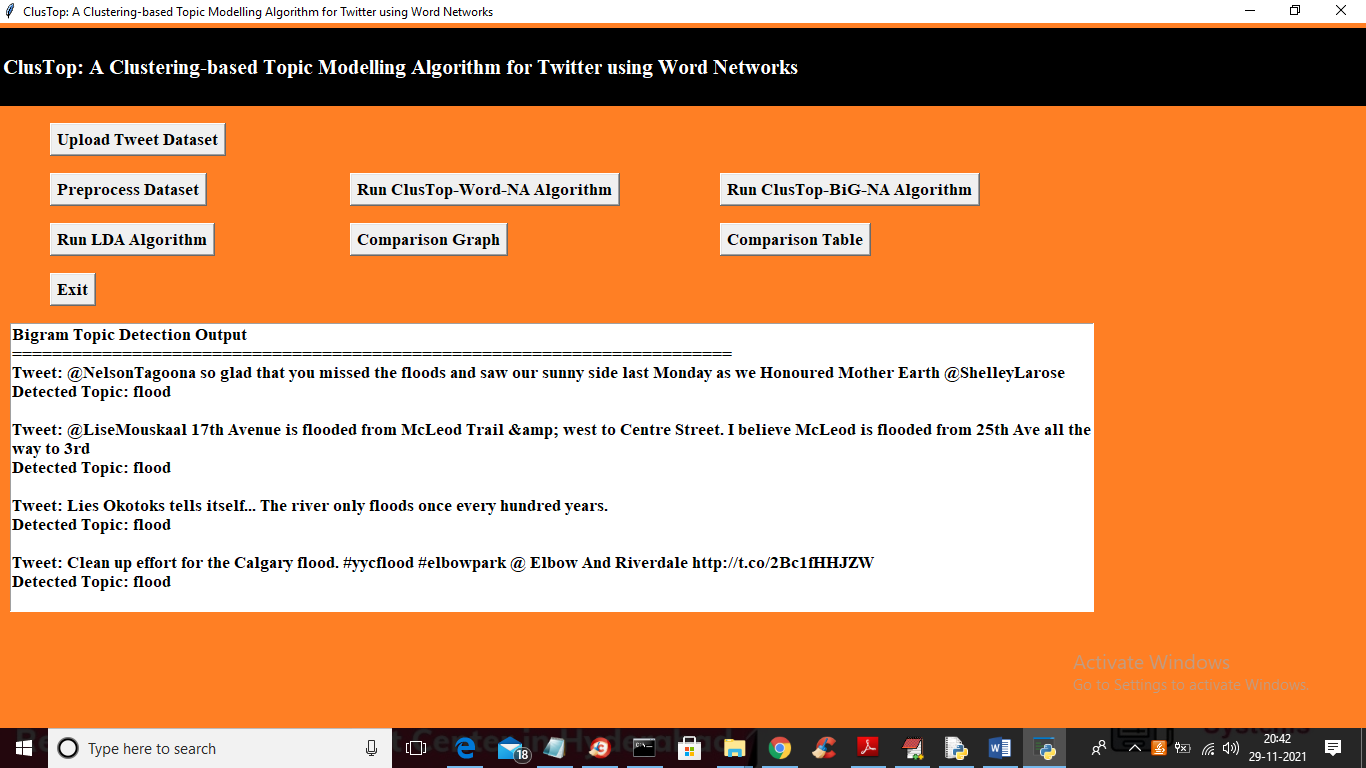
In above screen preprocessing and now click on ‘Run ClusTop-Word-NA Algorithm’ button to run algorithm with CO-WORDS technique and then detect topic



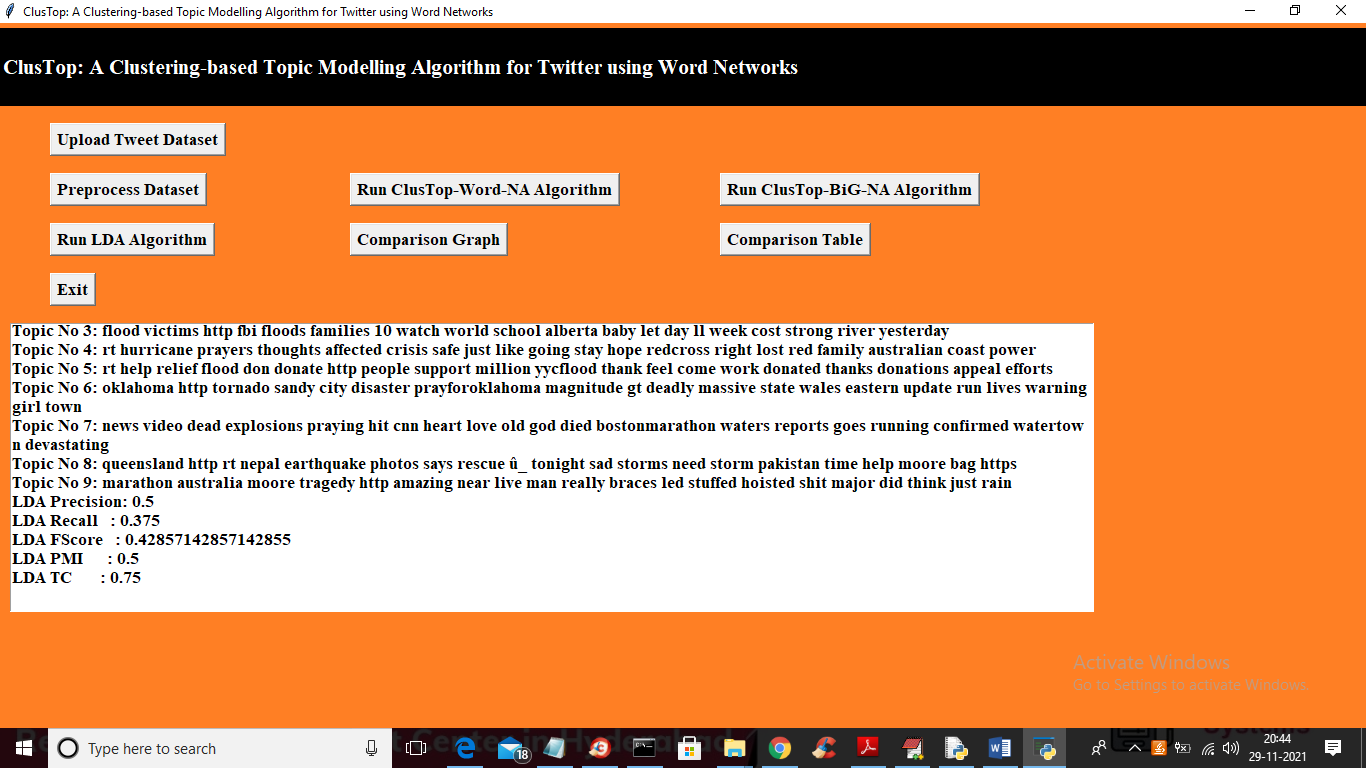
In above screen in first line I am displaying TWEET and in second line displaying detected topic as ‘flood’ and we you can scroll above output screen to see all detected topics as dataset contains 37000 tweets and we detected topic for all tweets



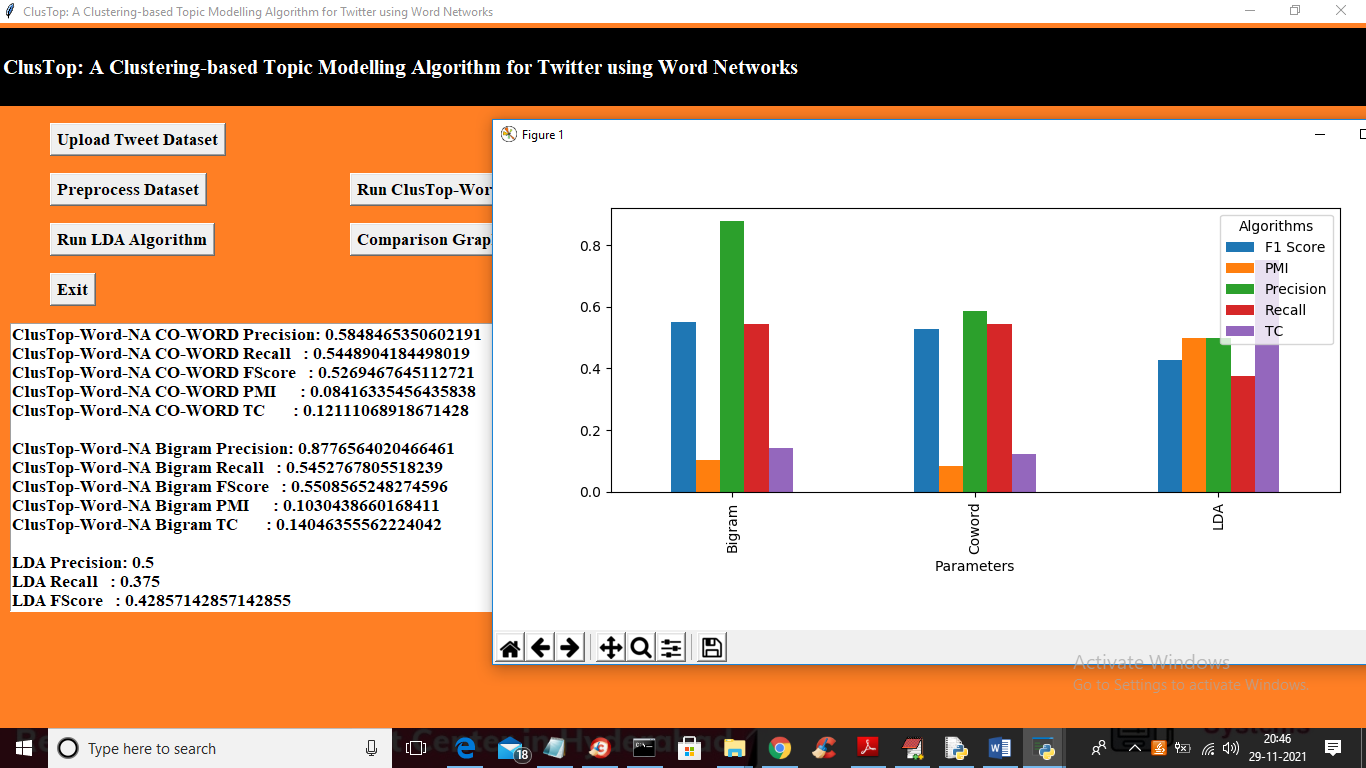
Now click on ‘Run ClusTop-BiG-NA Algorithm’ to detect topic using BIGRAM clustering



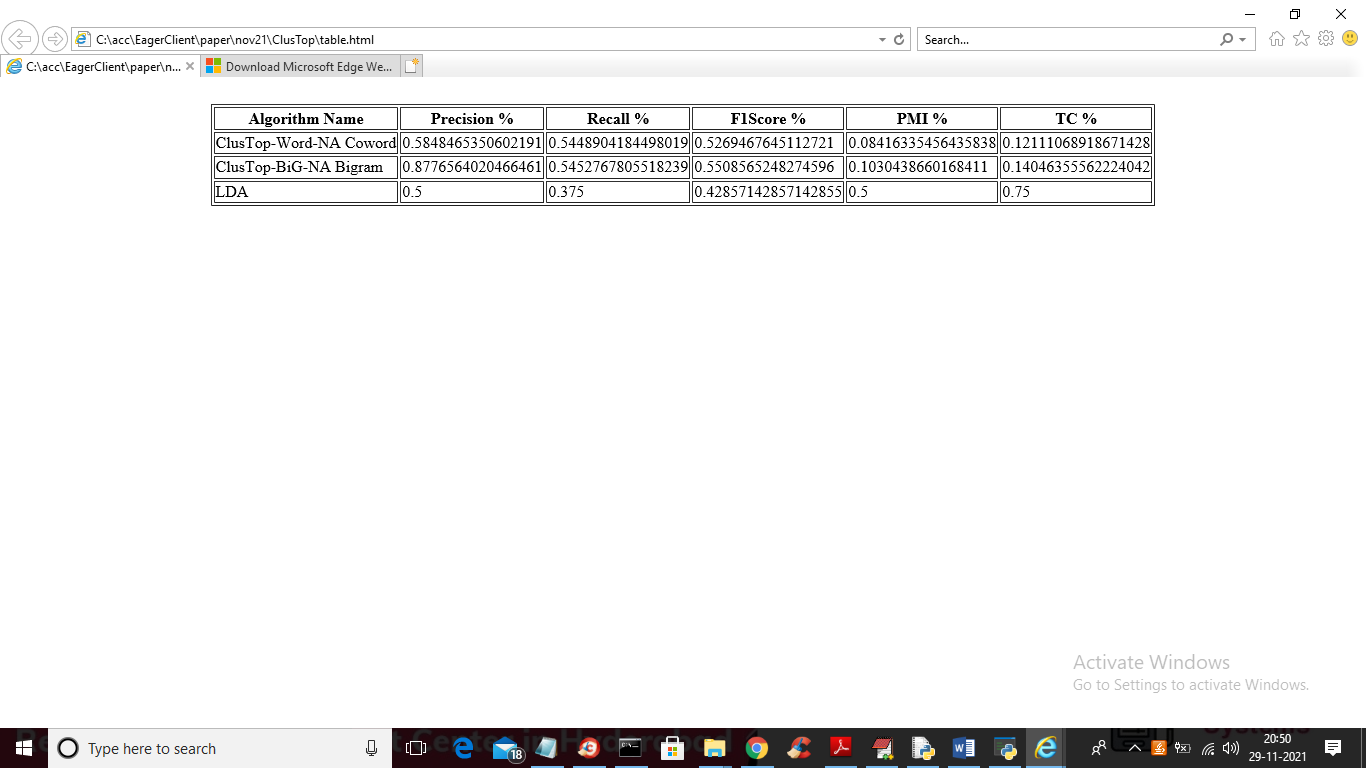
In above screen we detected topic with BIGRAM technique both algorithms are same but due to bigram technique this algorithm will have little good performance. Now click on ‘Run LDA Algorithm’ to find topics



In above screen LDA displaying all topics detected from all tweets and we can see its output is not as much clear as this paper algorithms and we got LDA precision and recall values and now click on ‘Comparison Graph’ button to get below comparison of all algorithms

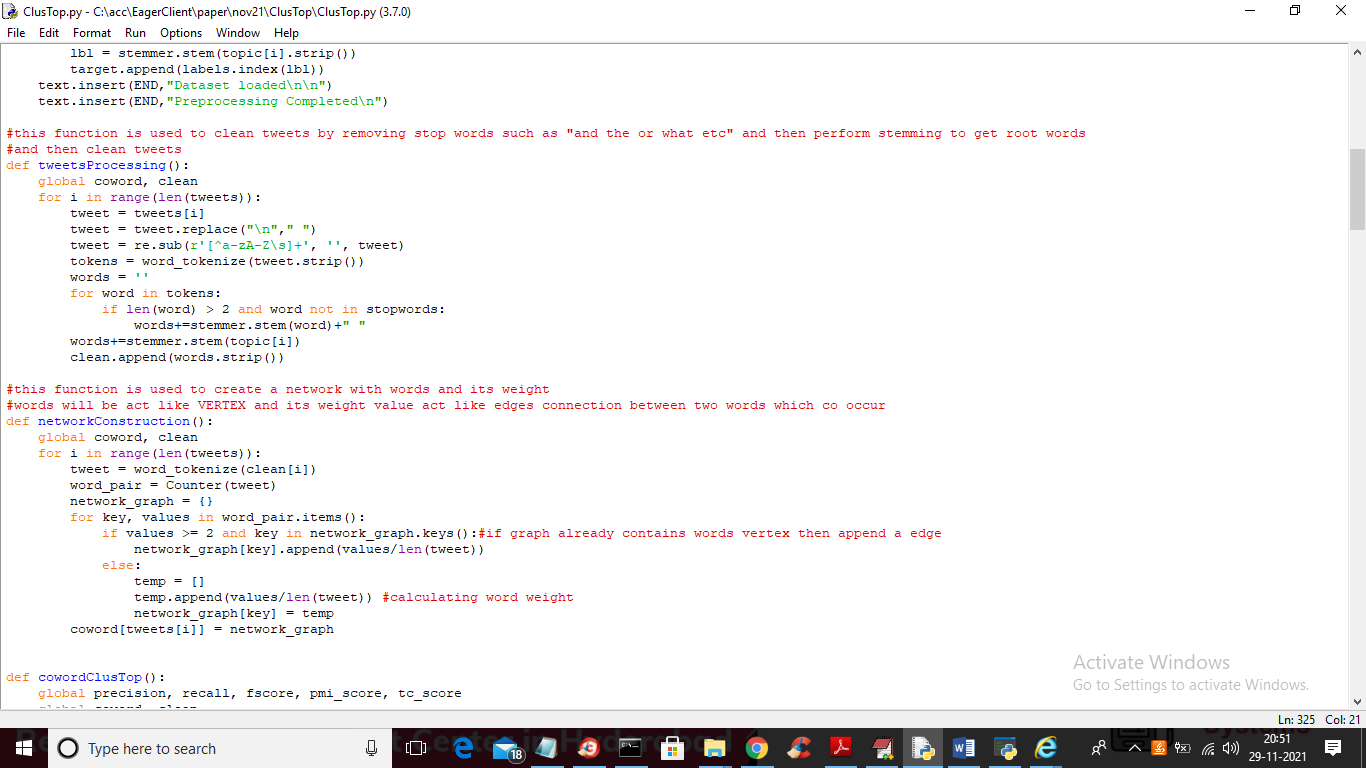


In above screen CO-WORD precision is 0.58 and bigram precision is 0.87 and LDA precision is 0.5 and in graph we are plotting different metrics such as precision, recall, fscore etc for each algorithms in different colour bar and from above graph we can conclude that propose paper BIGRAM algorithm got high performance compare to other algorithms. Now close above graph and then click on ‘Comparison Table’ button to get below comparison values of all algorithms



In above table you can see precision, recall and other values for each algorithm

In below screens I am showing code for algorithms implementation



In above screen read red colour comments to know about algorithm implementation

