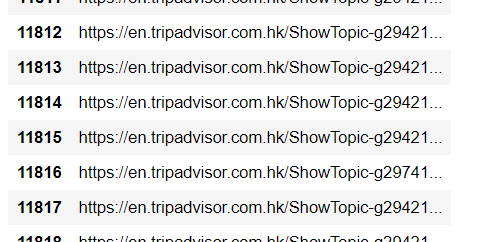
**Hong Kong Airport transport**

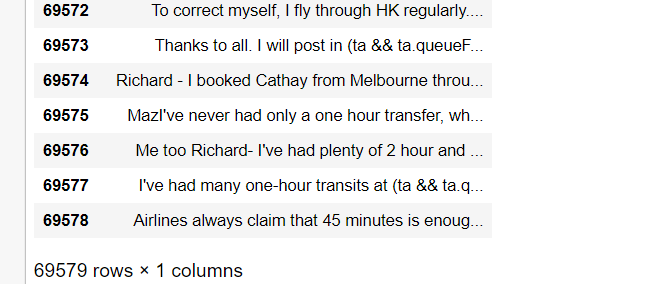
This analysis is done on the data of tripadvisor on the topic Hong Kong Airport transport. The goal of this analysis is to find analytics of travelers and the comments about the Hong Kong airport, China. So, Let’s step in and begging our study on the each part of the analysis.

The first and most important part of this project is to gather data, and the chosen source of data in our case is tripadvisor, since from this source we can not directly collect the data (because it is used as an indirect source of data) so the first step is to scrap the data from the tripadvisor forum using python.

First I have to collect all the links from which I can scrap data, that is first I have to scrap all the link where the desired data is available. Scraped all the pages and found that there are 11820 URLs from which I have to scrap the data related to the Hong Kong airport transport.



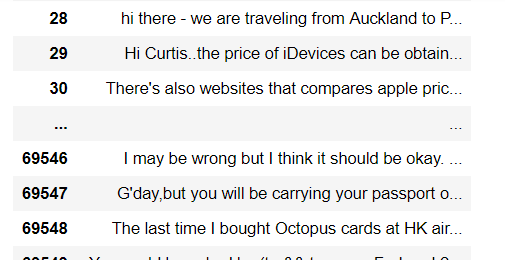
Gathering URLs is just an 10 percent of our task, the main task is yet to begin , So let’s dive the next step is to gather different types of data from each url, and this is time taking task, because I have to scrap 8 different types of data from each url and I have a total of 12 k urls this , to make it faster I’ve decided to scrap each data from certain part of list of urls, that is scraping 1000 url at a time and repeating this for the every next 1000 url. Despite scraping I parts it too lots of my time , but finally I got the dataset as in the provided dataset. The dataset which I obtained from the scrapping of a large number of url (12 k ) is very large , and the obtained dataset has around 60 k rows.



In the obtained dataset, for me the main point of concern is the content of the dataset that is the body or content of the post and their reply on the tripadvisor website. Because using only the content of the posts and the reply’s i can comment on the Hong Kong airport transport, this can only be done after the performing certain operation on the data obtained from the tripadvisor website. Because the data which I have scrapped from the website is in it’s raw form. First I have to perform cleaning operation on the data, only after cleaning and processing the data I can comment on the H K airport transport.

The first step in the cleaning of a data is deleting the duplicate data from our data set, so I have used pandas method to drop duplicate from the dataset, and performing this operation of deleting the duplicates decreased the size of the data.

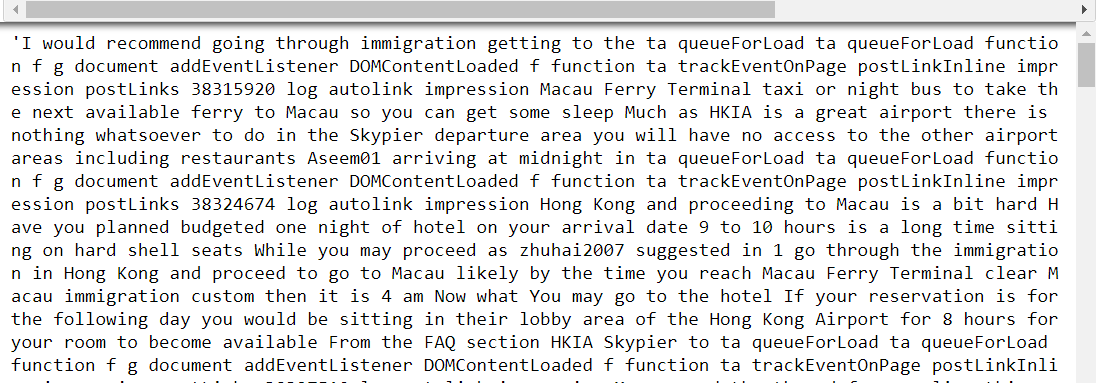
And this shows that there are a lots of duplicate posts on the website and also there are lots of duplicate posts.



Now I have text data , and I have to play with the data to predict certain comments about the Hong Kong airport transport.

The most popular way to play with the data is to use NLTK , so, I will also use NLTK library to and follow the methodology to meet the requirements.

According to the methodology the first step is to remove all the stop words from the data. Before this step I have to clean the data at one more level to remove the links and non word character to do this I have used regrex pattern matching and finally got the result as string. the result is as:

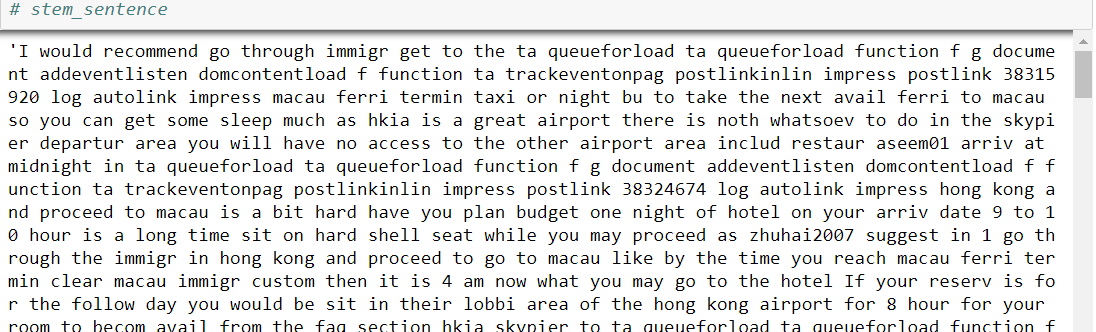


Now I have generated an clean data, now I can proceed further, the next step is to remove the stop word and then also I have to do tockenization.

First I have decided to remove the stop words from the sentence which I have generated in the earlier step. After removing stopwords and doing tockenization, I got the filtered sentence as:

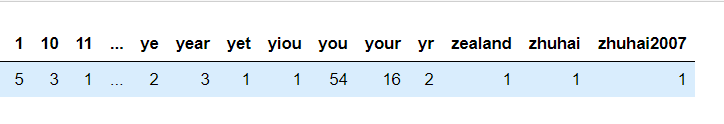


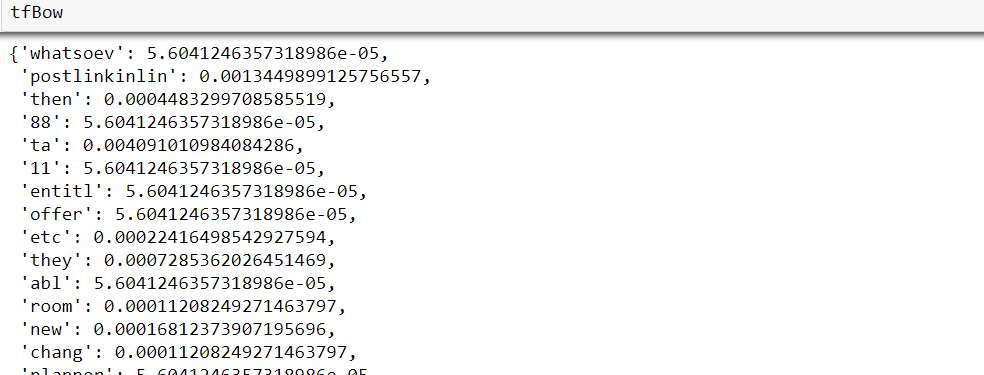
The above task of tockenization and removing the stopword is performed using the popular NLTK library. The next step is leminization and in this step I have to lemenize each words of the final sentence which is obtained earlier, to perform leminization I have used PorterStemmer of nltk. And after leminization I got the sentence as:



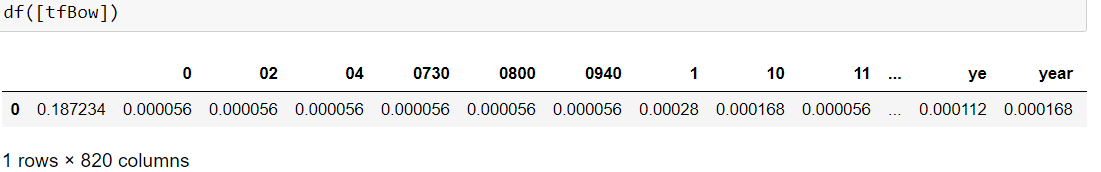
And after doing all this task the next step is to do Term frequency-inverse document frequency (TF-IDF) analysis. This is an analysis used for the frequency counter of an word in an sentence. The results which I got when doing this task is as folloe:



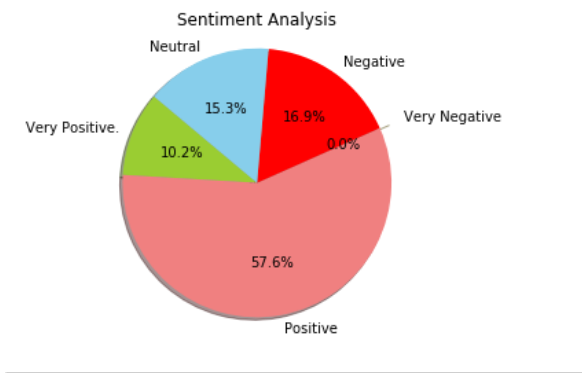




And the final result for the TF-IDF is as:

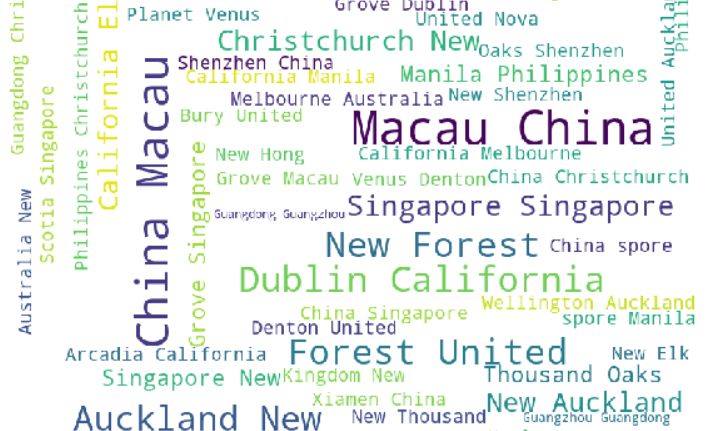


Now, after doing this it is time for the natural language processing, in this I have to calculate the sentiment of the comments made by the people on the Hong Kong airport transport, and to do sentiment analysis I have used textblob. And since it is given in the methodology that the sentiment analysis should must be divided into 5 cateogry keeping this in mind I have done sentiment analysis of the people posts and replies on the website and the result that I got form this analysis is shown with the pi chart as :



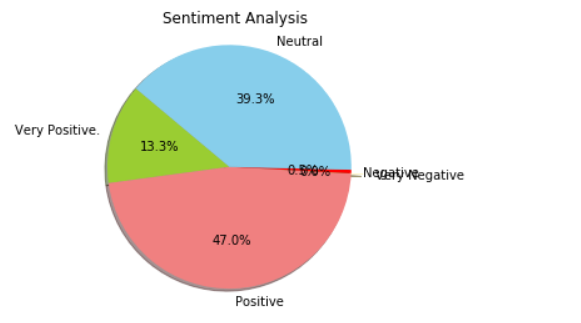
And it is clearly seen in the sentiment analysis of the content present on the website is that the most of the post made on the tripadvisor with Hong Kong airport is positive (56.6 %) content, and very little were negative, and 0.0% with very negative. This refelects good thing about the hong Kong airport transport.

Next I have generated an word cloud too see the point of concern of people on the Hong Kong airport Transport. I have generated 3 wordcloud one to check the location of the people how replies for the post (helpers) and one for the content of the post and the last for the topic of the post. The three word cloud are as follow:



1. Location cloud

At the last I have done sentiment analysis of the topic of the replies and also generated an word cloud for the same. The sentiment analysis for the post topic is as follow:



Got very good result in the sentiment analysis for the post. Now let’s see the word cloud , word cloud will show peoples point of concern. The word cloud is as:

