

Wee Kim Wee School of Communication and Information

**H6751 - Web and Text Mining Group Project**

***Twitter Sentiment Analysis***

***of***

***Universities in Singapore***

*Submitted by*

Phu Wai Paing (G1901561C)

Ye Lynn Khant (G1901531J)

Koh Swee Guan (G1802072F)

1. **Introduction**

Nowadays, the world is the village because of the internet. A lot of the information are being shared quickly and widely over social media like Facebook, YouTube, Twitter and, content sharing sites, etc. Social media was designed as a reason for communication and socialization, not only between relatives, friends and small communities, but also for the world to communicate and discuss. Their goal is to support and to satisfy the need of communication and networking between individuals or organizations through the internet.

Among the social media, Twitter is one of the most popular, microblogging platforms. Massive amount of information about any interested topic can be extracted. Tweets can then be analyzed based on people’ thoughts and opinions on any topic of interest which can be positive or negative through the natural language processing termed as ‘sentiment analysis’. Sentiment analysis is also called ‘opinion mining’ that is classifying the opinion or attitude of a speaker as ‘positive’, ‘negative’ or ‘neutral’.

Today, 80% of the world digital data obtained from social media and other sources is unstructured (Cassetta, 2019). Since the information is not managed in any predefined way, it’s difficult to organize and analyze. By pre-processing the unstructured data, sentiment analysis allows us to analyze large sets of tweets and detect the opinion of each tweet automatically. Once properly implemented, the process is efficient, fast and simple. This can result in teams saving invaluable hours and allowing them to focus on other tasks that are more strategic and making a bigger impact to the organization.

Sentiment Analysis tries to gauge and extract opinions, emotions and attitude within a given text and is useful especially in marketing, advertising, political science, sociology and psychology. Through sentiment analysis, companies can understand the customers’ feelings towards the company, how customers respond to their campaigns and develop strategies to meet their goals. Sentiment analysis can also be used to observe the general mood of the blogosphere, monitor the popularity trend of the company and its competitors as well as analyse the social phenomena.

Users can make sense out of data by automating the entire process through sentiment analysis without having to manually indulge with it.

NUS and NTU are ranked as top 25th and 48th under THE World University Rankings (n.d.). SMU is not ranked. UniRank (n.d.) identified NUS, NTU and SMU as the top three universities in Singapore respectively. *Does the ranking of the university result in the similar students’ satisfaction and sentiments?*

The main objective of the project is to conduct comparative sentiment analyses of the top three universities in Singapore based on their tweet datasets to determine if the ranking of the university is related to the satisfaction and sentiments of their stakeholders. Through this project, it is hoped that we can demonstrate the understanding of the using training data to train our model, explore different machine learning classification algorithms and the application of this knowledge in analyzing the sentiments of tweets on the universities in Singapore.

1. **TRAINING DATA**

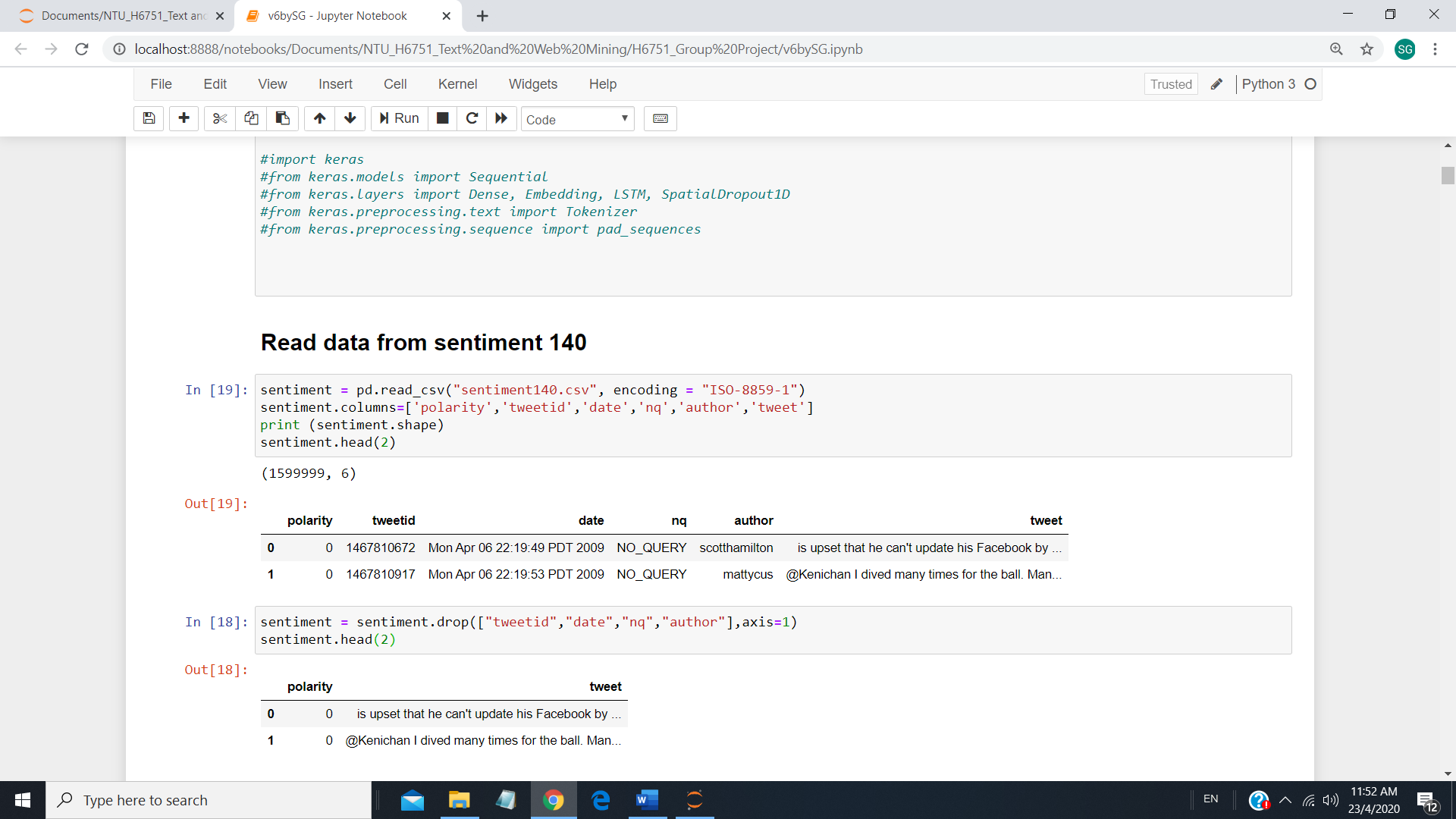
**2.1 Training Data from Kaggle**

For the training and testing data, we will be using the ‘Sentiment140’ data available from Kaggle (n.d.). It was created by Alec Go, Richa Bhayani, and Lei Huang (2009), who were Computer Science graduate students at Stanford University. Dataset was collected for academic research purpose using the twitter API and contains pre-tagged sentiment tweets from multiple hashtags. The dataset has total number of 1,599,999 records with six fields or attributes for each record. They are:

1. Polarity of the Tweet
   * 0: Negative (799999)
   * 4: Positive (800000)
2. Tweet ID
3. Date of Tweet
4. Query: if there is no query, the value is NO\_Query
5. User that tweeted
6. Text of the tweet

**2.2 Checking and Understanding the Data**

The data was loaded and shown to understand the structure of the data. A total of 159, 999 tweets with 6 features were available.



The "tweetid", "date", "nq" and "author" were dropped as we do not need this information. The data was checked for missing data. There were no missing data as shown.

**2.3 Data Pre-Processing**

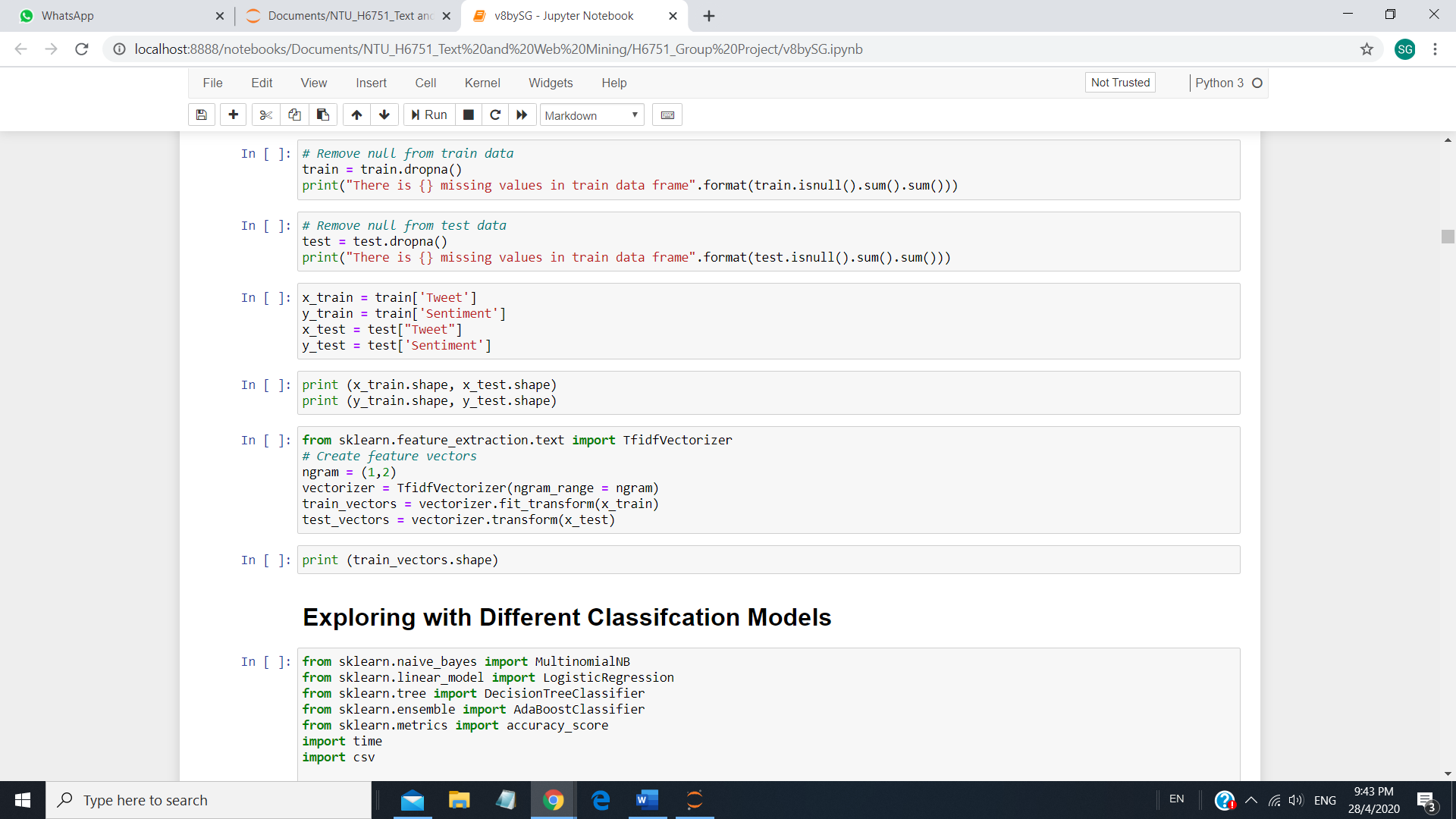
The data was pre-processed to remove punctuation, non-ascii characters, website address and numbers. Stop words were removed, de-contractions, tokenization and stemming were carried out (Monsters, 2018). The pre-processed data were saved via ‘pickle.dump’ so that data cleaning and pre-processing need not be carried out whenever we re-run the program. This helps to save time as it is time-consuming to pre-process 1.6 million records.

**2.4 Validation**

The data was split into test and train data randomly. The train data will be used to train the model to be generalized on other data. The test dataset will be used to test the model’s prediction. The use of train/test split helps validate the predictions of the various models being explored and avoid overfitting (Bronshtein, 2017). The test size of 0.2 is used after exploring test size between 0.15 and 0.3. Test size of 0.2 produces the best testing accuracy.

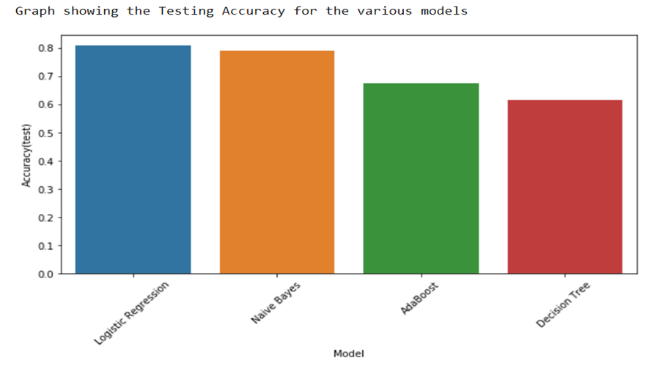
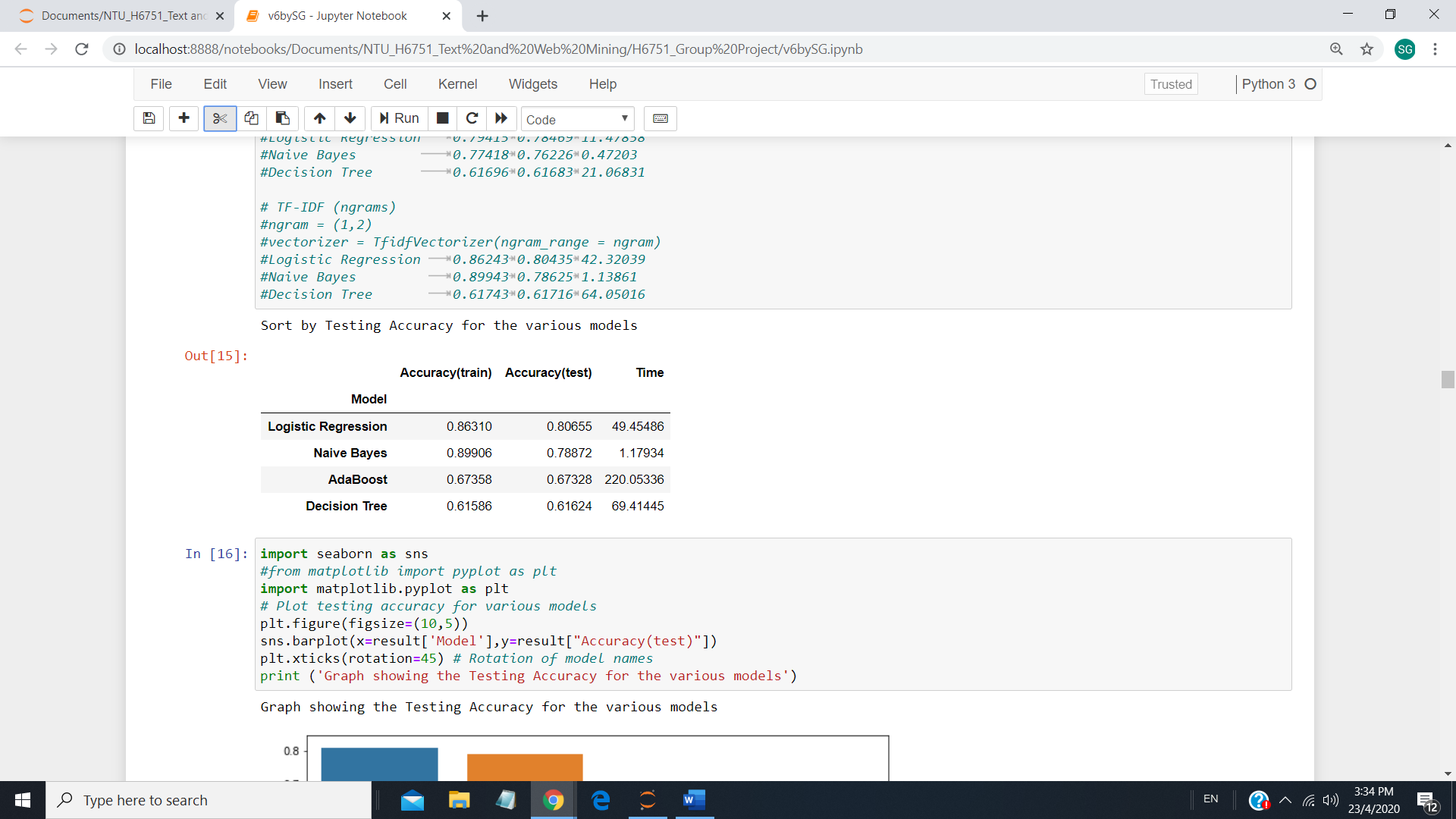
**2.5 Feature Engineering**

For feature extraction, the simplest technique is to use Bag of Words. Bag of Words (BOW) is the making the list of unique words in the document. Term Frequency- Inverse Document Frequency (TF-IDF) is the most widely used technique to process textual data (Tripathi, 2018) and is applied to count how many times a word has occurred in the dataset. TF-IDF vectorizer with n-grams features were used as it produced the highest test accuracy score.



**2.6 Exploring Different Machine Learning Algorithm**

Our project involves classification, which is a is a supervised learning approach in which the model learns from the input data and then uses this learning to classify new observations (Sidana, 2017). Different classifiers were explored to help build understanding on these algorithms. Four different Machine Learning (ML) Algorithm were explored. The four models are Naïve Bayes (NB), Logistic Regression (LR), Decision Tree and AdaBoost. There were attempts to explore other classifiers such as Random Forest, Support Vector Machine and K-Nearest Neighbours. However, the training time for these classifiers took too long and were dropped. The training accuracy, testing accuracy and training time were displayed. The chart and table below show the testing accuracy and results from the various classification models respectively.



The results show the LR algorithm produces the best training score of 0.80655. NB has the highest training accuracy and fastest training time followed by LR. In view of the highest test accuracy score and relatively fast training time, LR is used for our classification of tweets from the universities.

**2.7 Hyperparameters Tuning & Cross-Validation**

Hyperparameters are vital to a model as they affect the overall performance of a ML algorithm and can be treated as a search problem to find a best combination of hyperparameters that minimizes a predefined loss function to give better results (Paul, 2018). Grid search can be used to perform hyperparameters optimization and is implemented in scikit-learn under GridSearchCV. Since LR gives the highest accuracy score, fine-tuning their respective hyperparameters may help to further improve the accuracy score. First, the solver for LR. There are different solvers in LR such as 'liblinear', 'lbfgs', 'sag', 'saga'. GridSearchCV recommended the tuned 'solver': 'liblinear' with a score of 0.80279. Next tuning the C value with a value of 0.001, 0.01, 0.1, 1 and 10. The GridSearchCV result shows a tuned C value of 1 with a same score of 0.80279, as the default C value is 1. Fine tuning of C value was carried out between 0.5 to 2.5 with a step of 0.1 (i.e. 0.5, 0.6, ….2.4, 2.5). It was shown that the tuned C value is 1.8 with a score of 0.80401.

Cross-validation can be used to ensure unbiased and generalization of our model. A 10-fold cross-validation was carried out. A standard deviation of 0.001085 was obtained. The low standard deviation indicates that our model is relatively unbiased and is expected to perform when used to make predictions.

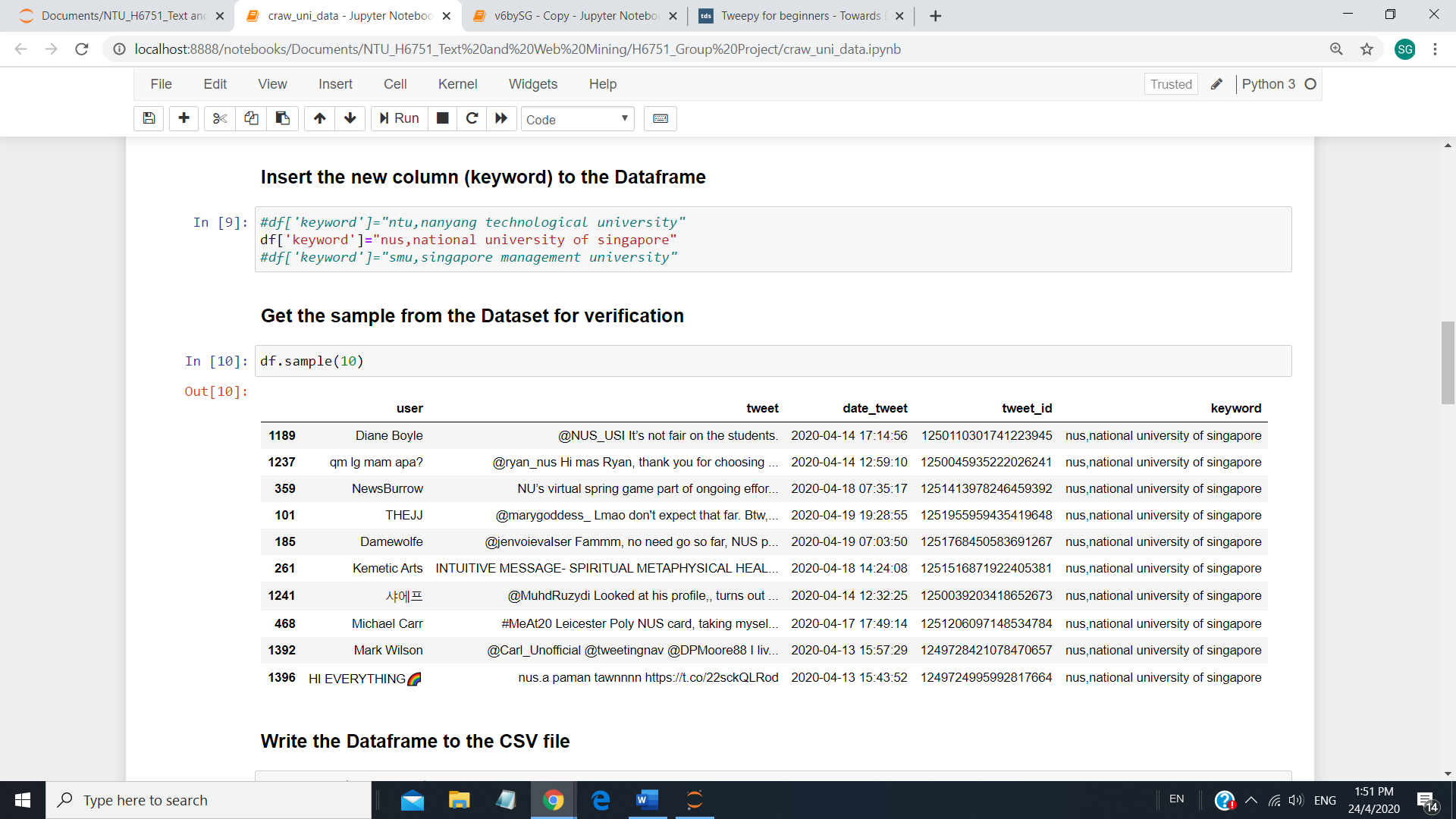
**2.8 Saving the Optimized LR Model**

Using the tuned hyperparameters, the model was re-evaluated. The testing accuracy was increased from 0.80655 to 0.80734, an increase of 1%. The optimized trained LR model was saved using ‘pickle’. This will save processing time as it is time consuming and slow to always train the model before every application.

1. **DATA TO BE ANALYSED**

**3.1 Extracting Tweet Data**

The data, which will be analyzed by using the trained model, is collected on our own by using the twitter API (Srivastava, 2018). Tweepy was used to extract tweets using the search word 'ntu OR "nanyang technological university"', 'nus OR "national university of singapore"' and 'smu OR "singapore management university"' to access tweets on the 3 universtities. A sample of the data extracted from nus tweets is shown below.



The tweets were extracted on two separate occasions. The first set was carried out in February 2020 and the second set was carried out in April 2020. Each set of tweets extracted were limited to 1500 for app authentication by Twitter (n.d.). The tweets from the three universities were saved into csv files for easy access later.

* 1. **Tweets Dataset Pre-processing**

The csv files on the tweets of NTU, NUS and SMU were loaded. The user, tweet\_id and keyword were dropped as it will not be used. The tweets were pre-processed like the training dataset (see Section 2.3). The pre-processed data to be analysed were saved via ‘pickle.dump’ to save time.

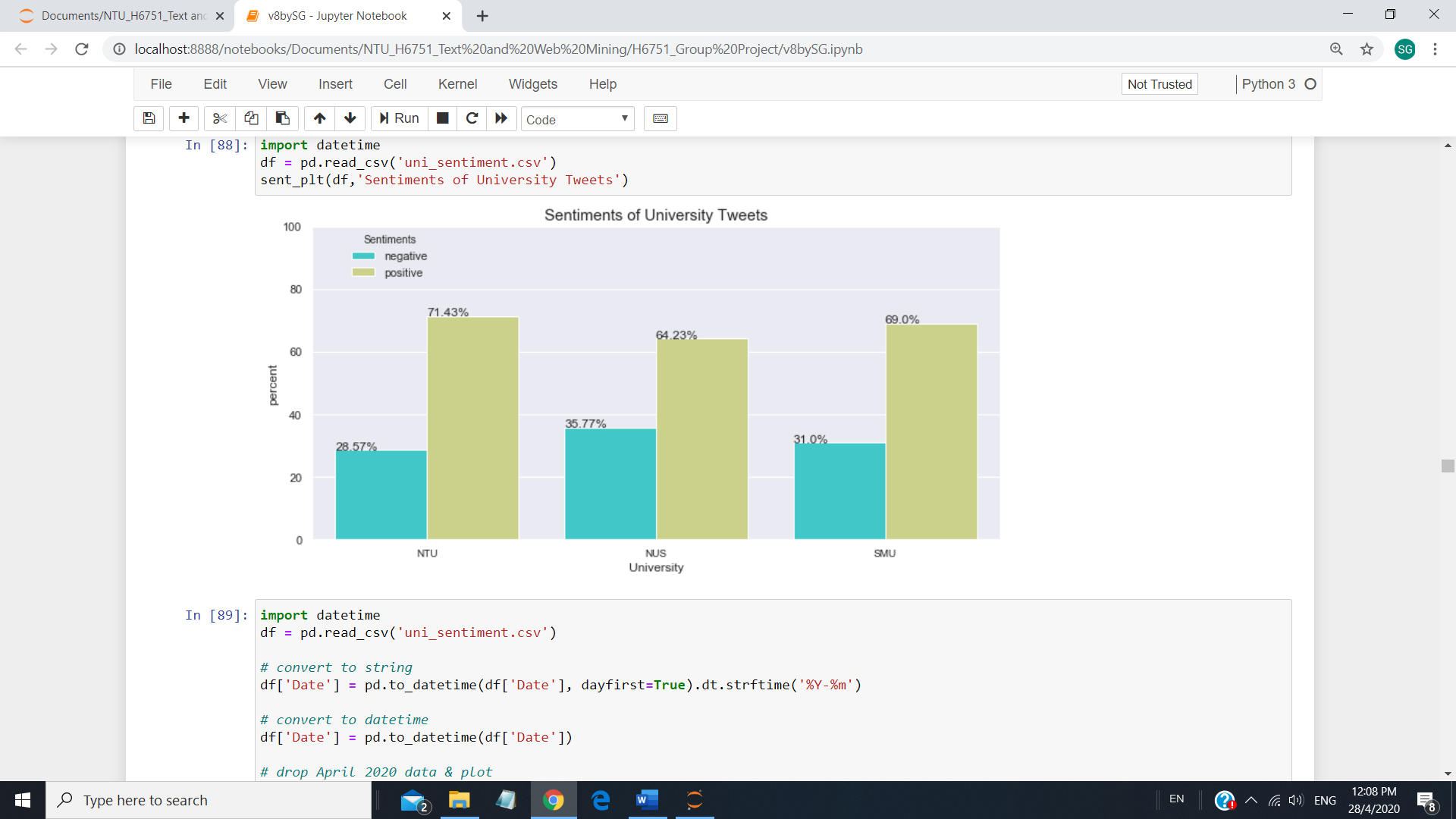
* 1. **Converting Tweet Dataset to Data Frame**

The tweets dataset was converted to data frame and store into csv for later access. The number of records for NTU, NUS and SMU are 2,383, 2818 and 3,000 respectively. Checks were carried out to ensure that there are no missing data by checking for null values. The missing data appeared as a result of pre-processing. Records with missing data were removed. There are three missing values in both NTU and SMU datasets. The number of records for NTU, NUS and SMU are 2,380, 2818 and 2,997 respectively.

1. **Analysis of Results**

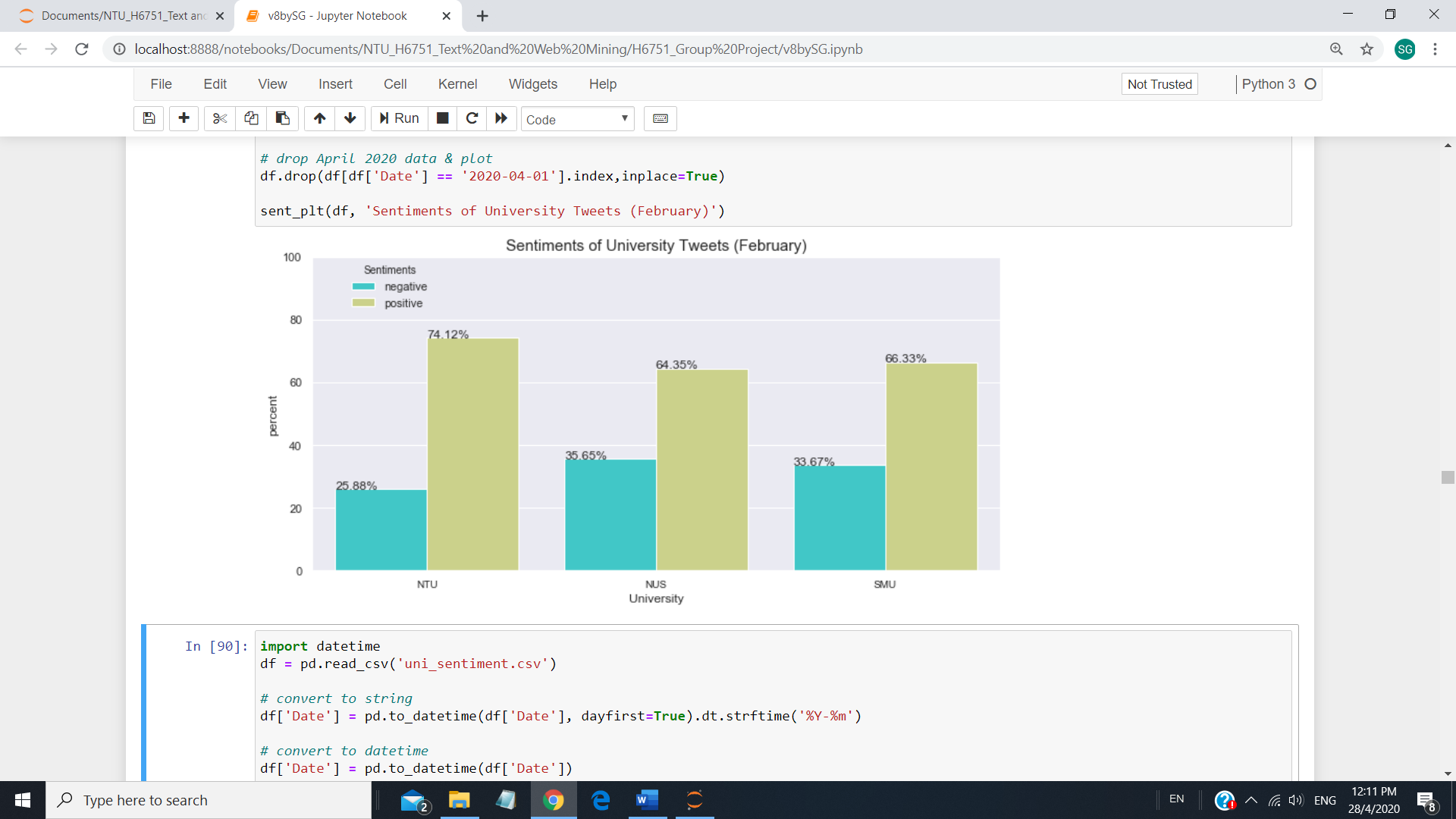
**4.1 Sentiments Analysis through Classification**

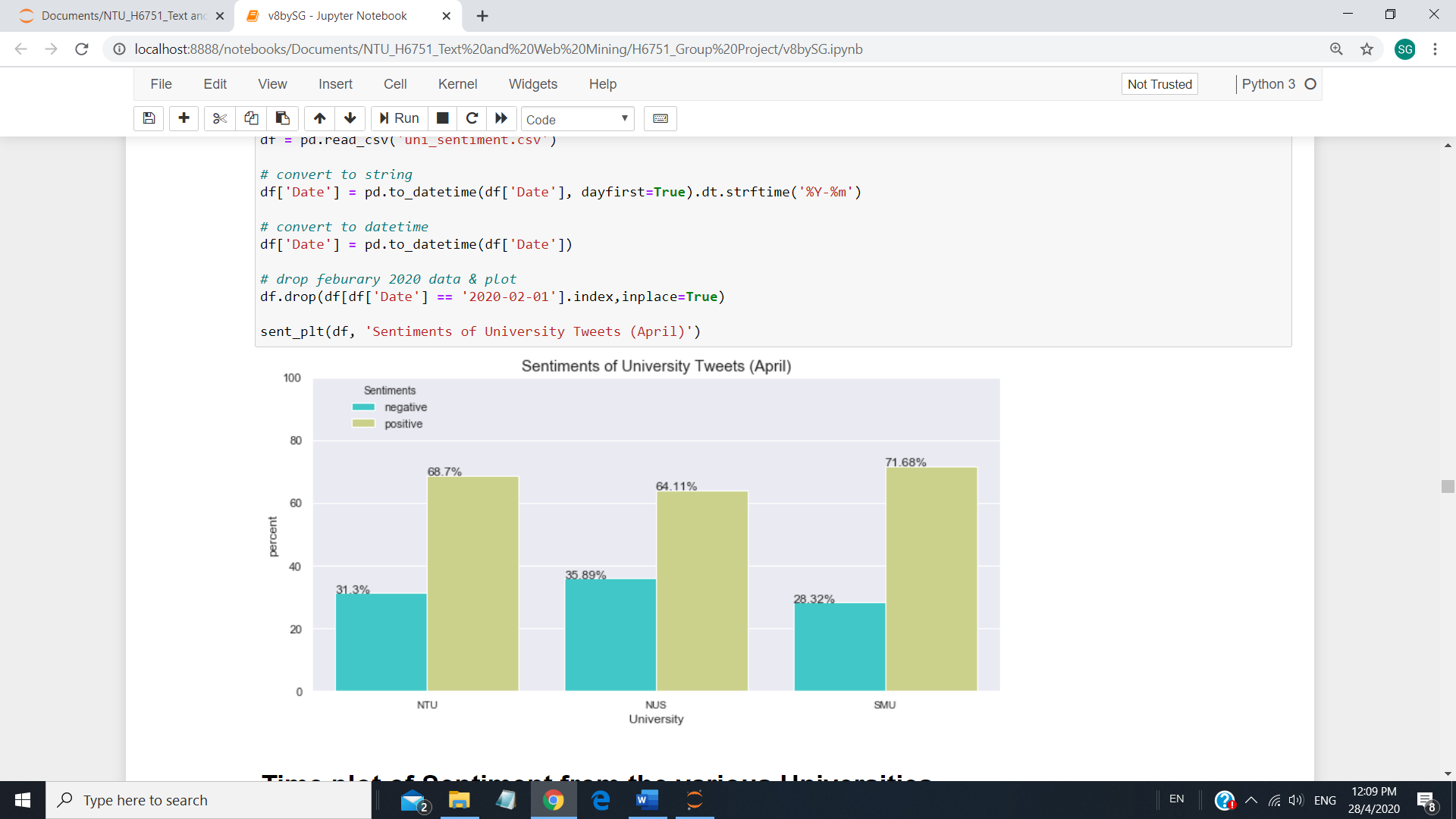
The university tweets were vectorized using TF-IDF vectorizer with n-grams features. The optimized LR model was loaded using pickle to classify the sentiments of the tweet datasets. The sentiments of university tweets from the LR classification for the two periods combined is as shown below:



From the results, NTU has the highest percentage of positive tweets at 71.43% and NUS has the lowest percentage of postive tweets at 64.23%. As this data was collected over two periods, there is a need to look for consistency over the periods.

For the period of February 2020, NTU has the highest percentage of positive tweets of 71.43% with NUS having the lowest percentage of positive tweets at 64.23%. SMU is a close second with 69% of the tweets with positive sentiments.



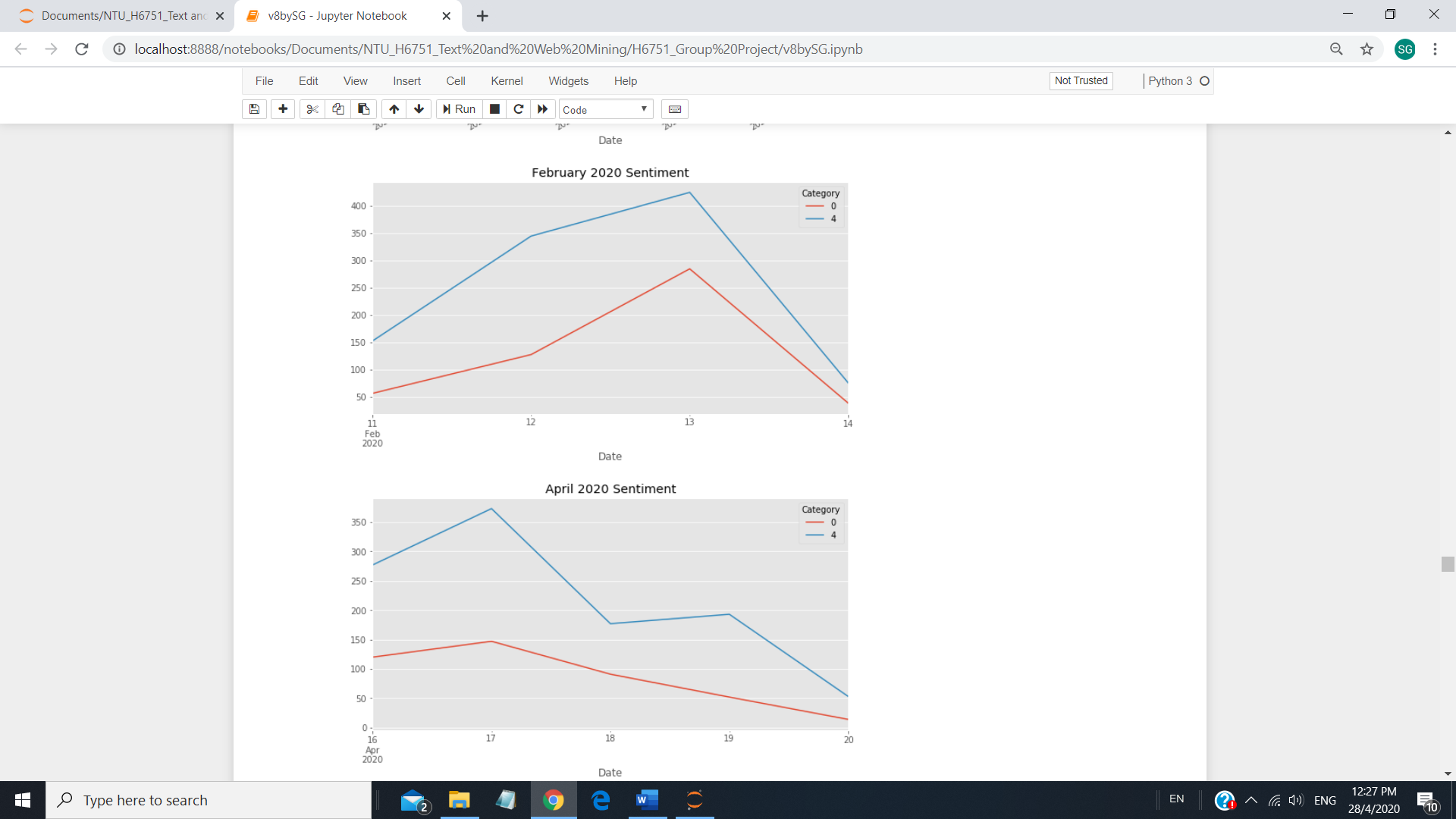
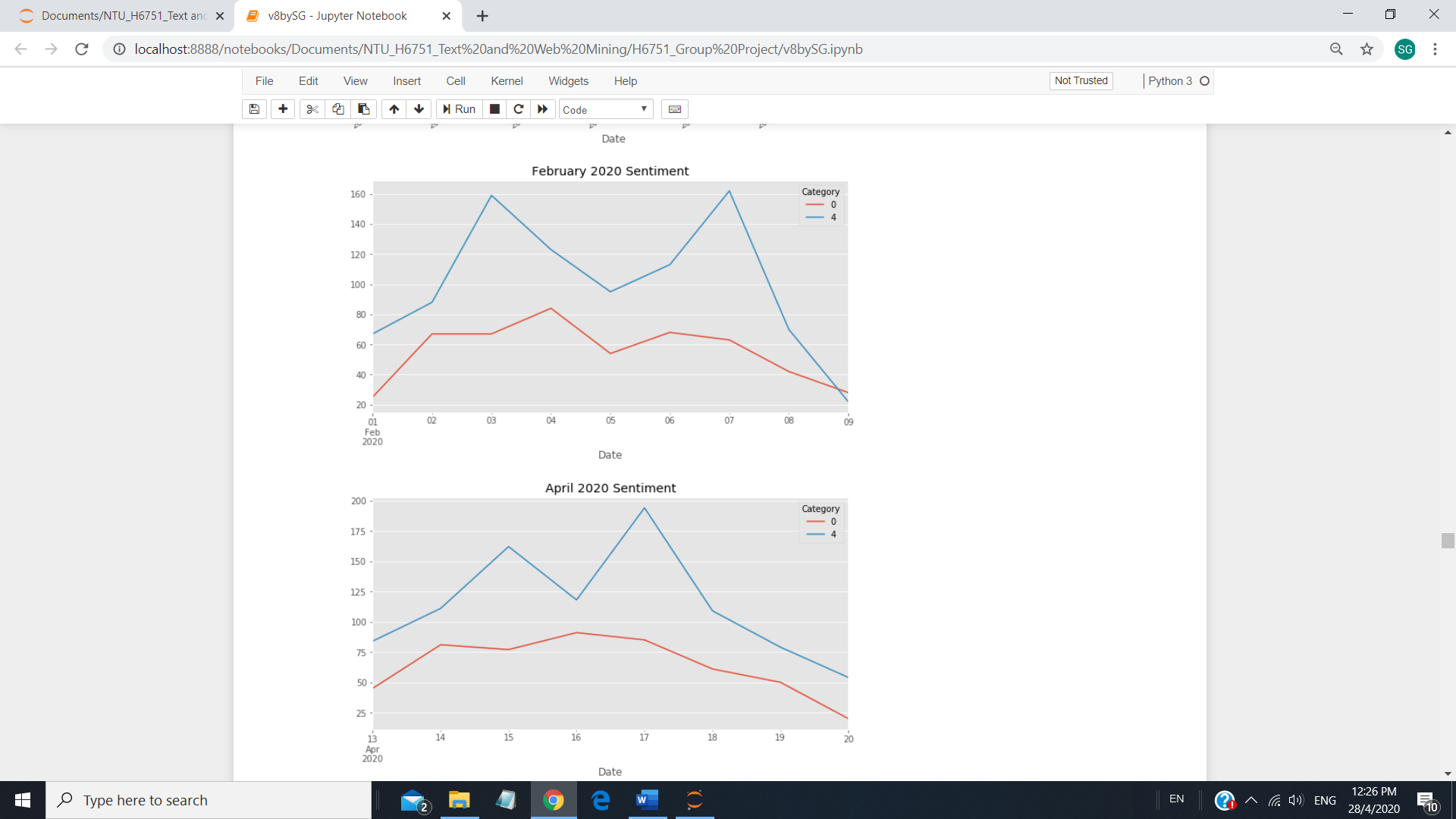
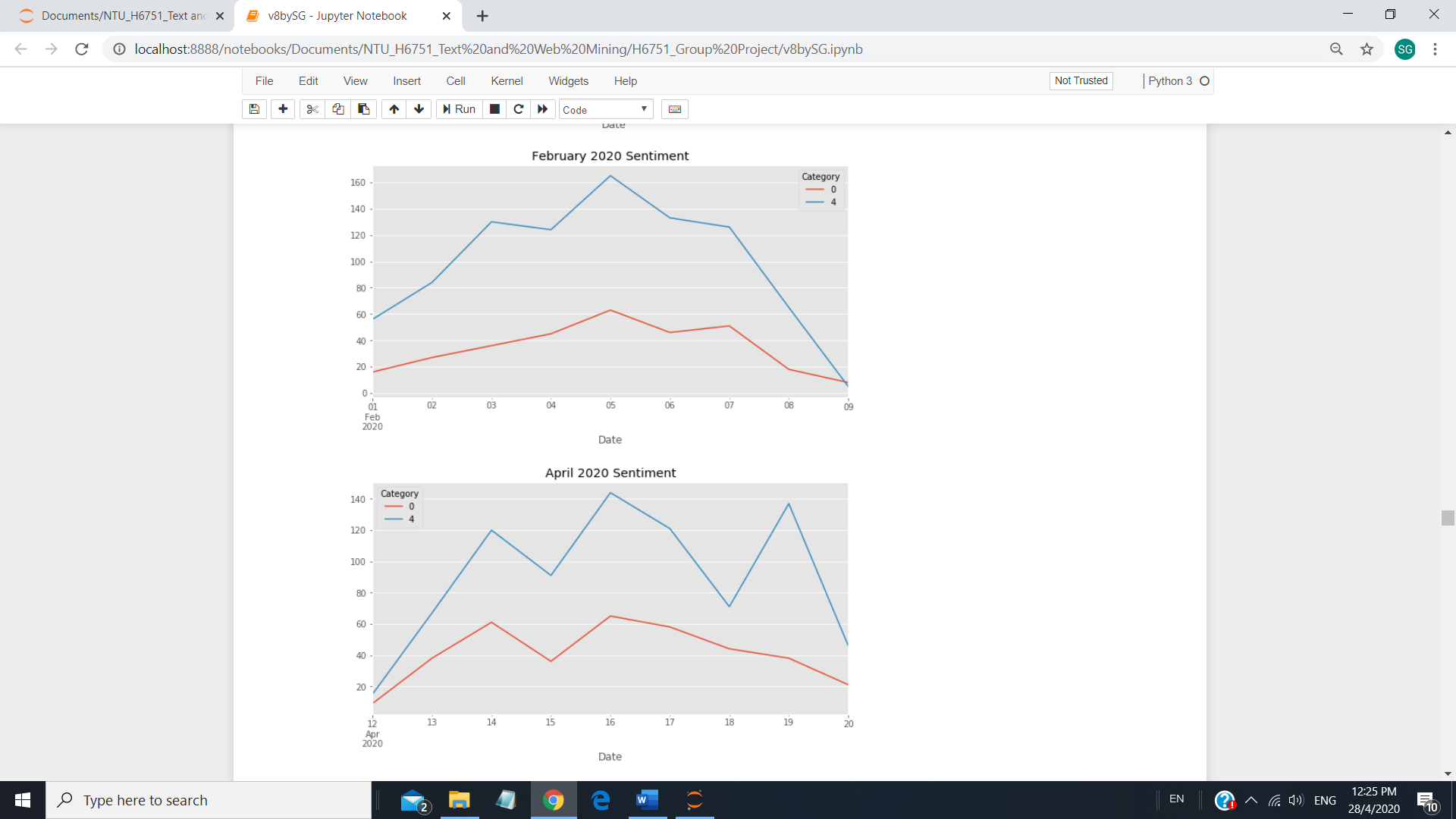


For the period of April 2020, SMU positive sentiments increases to 71.68% while NTU positive sentiments drops to 68.7%. The sentiments in NUS is consistently lowest at about 64%. The results show that being ranked the top university in Singapore does not equate to the most positive sentiments from their stakeholders.

**4.2 Analysis over time frame**

Next we look at the analysis over the days in February where tweets data were extracted. The figures below compare the results. NTU and NUS tweets extracted in February cover 1 Feb to 9 Feb 2020 while SMU tweets were extracted from 11 Feb to 14 Feb 2020. In April, tweets extracted from the NTU, NUS and SMU postings were from 12, 13 and 16 April respectively to 20 April 2020. SMU have much higher tweets posting with about 1500 tweets in four days while NTU and NUS takes about eight days to have similar number of tweets. This is observed both in February and April 2020.

The charts show that the daily ratio of positive over negative sentiments remain relatively constant as shown by the similar shapes of the red and blue lines. It is observed that most of the tweets were posted on weekdays except on 19 April 2020 (Sunday) where NTU has a surge in positive posting. Upon checking the tweets, there was extensive tweet posting on ‘King Chody is the Legendary NTU SCSE SG Top Graduate’ tweets and King Chody's Almighty Natural Language Processing Project for the module CZ4045 Natural Language Processing which may resulted in substantial posting on a Sunday.



**NTU NUS SMU**

**4.3 Sentiment Intensity Analyzer - VADER**

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool that is very useful for sentiment analysis of social media posting and not only tells whether it is positive or negative sentiment but also how positive or negative a sentiment is. VADER advantages over traditional methods of sentiment analysis includes suitability for social media type text, fast enough to be used online with streaming data and does not require any training data as it is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon (Pandey, 2018).

The tables below summaries the top ten positive and negative words from the tweets of the three universities using VADER. Noticed that the word ‘lmao’, which is an internet slang appears in both NUS and SMU top ten positive words and successfully identified by VADER. ‘lmao’ or ‘Laughing my ass off” is indeed an expression of positive emotions and sentiments. The results show that the lingo used by the tweeters of the three universities are similar.

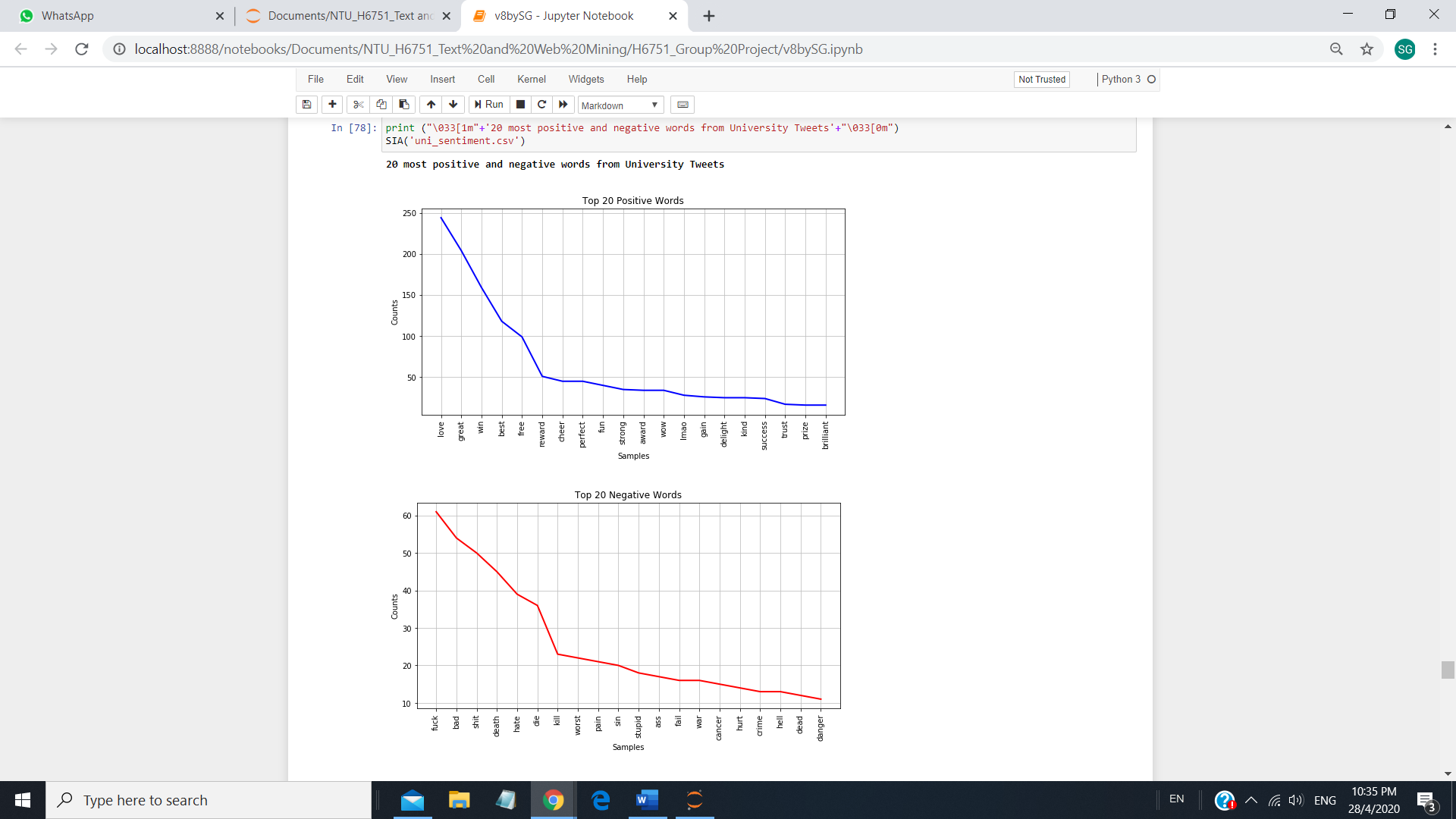
**Table of Top 10 Positive Words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Positive Words Ranking | Nanyang Technological University (NTU) | National University of Singapore (NUS) | Singapore Management University (SMU) | Three  University  Tweet Datasets |
| 1 | love | great | win | love |
| 2 | great | love | love | great |
| 3 | reward | best | great | win |
| 4 | cheer | free | free | best |
| 5 | best | delight | best | free |
| 6 | perfect | award | strong | reward |
| 7 | free | wow | fun | cheer |
| 8 | fun | honour | lmao | perfect |
| 9 | win | gain | wow | fun |
| 10 | success | lmao | rich | strong |

**Table of Top 10 Negative Words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Positive Words Ranking | Nanyang Technological University (NTU) | National University of Singapore (NUS) | Singapore Management University (SMU) | Three  University  Tweet Datasets |
| 1 | bad | fuck | death | fuck |
| 2 | sin | shit | bad | bad |
| 3 | die | hate | fuck | shit |
| 4 | hate | bad | shit | death |
| 5 | shit | die | ass | hate |
| 6 | fuck | pain | kill | die |
| 7 | crime | cancer | hate | kill |
| 8 | evil | worst | worst | worst |
| 9 | hell | war | idiot | pain |
| 10 | death | worsen | stupid | sin |

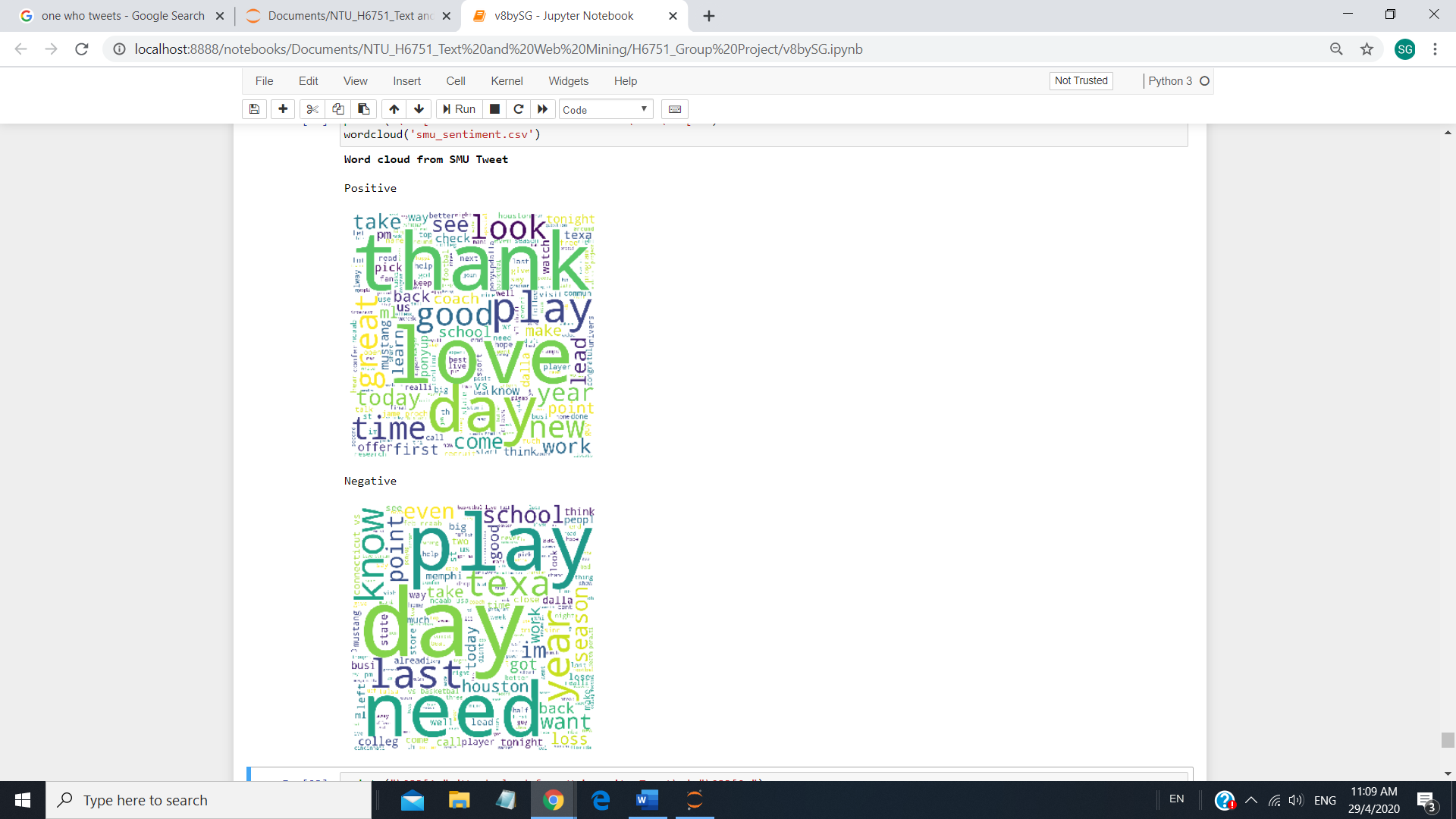
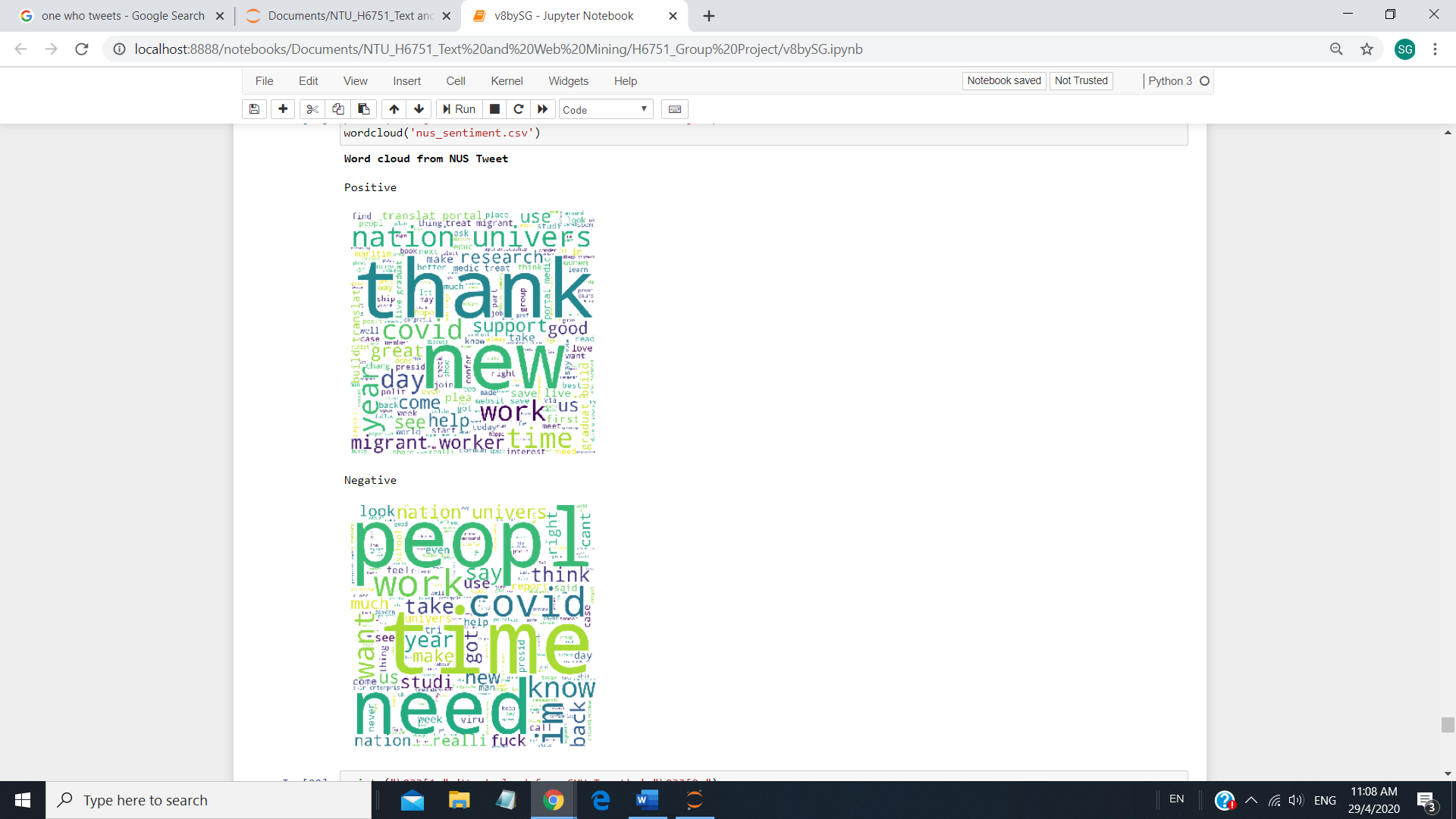
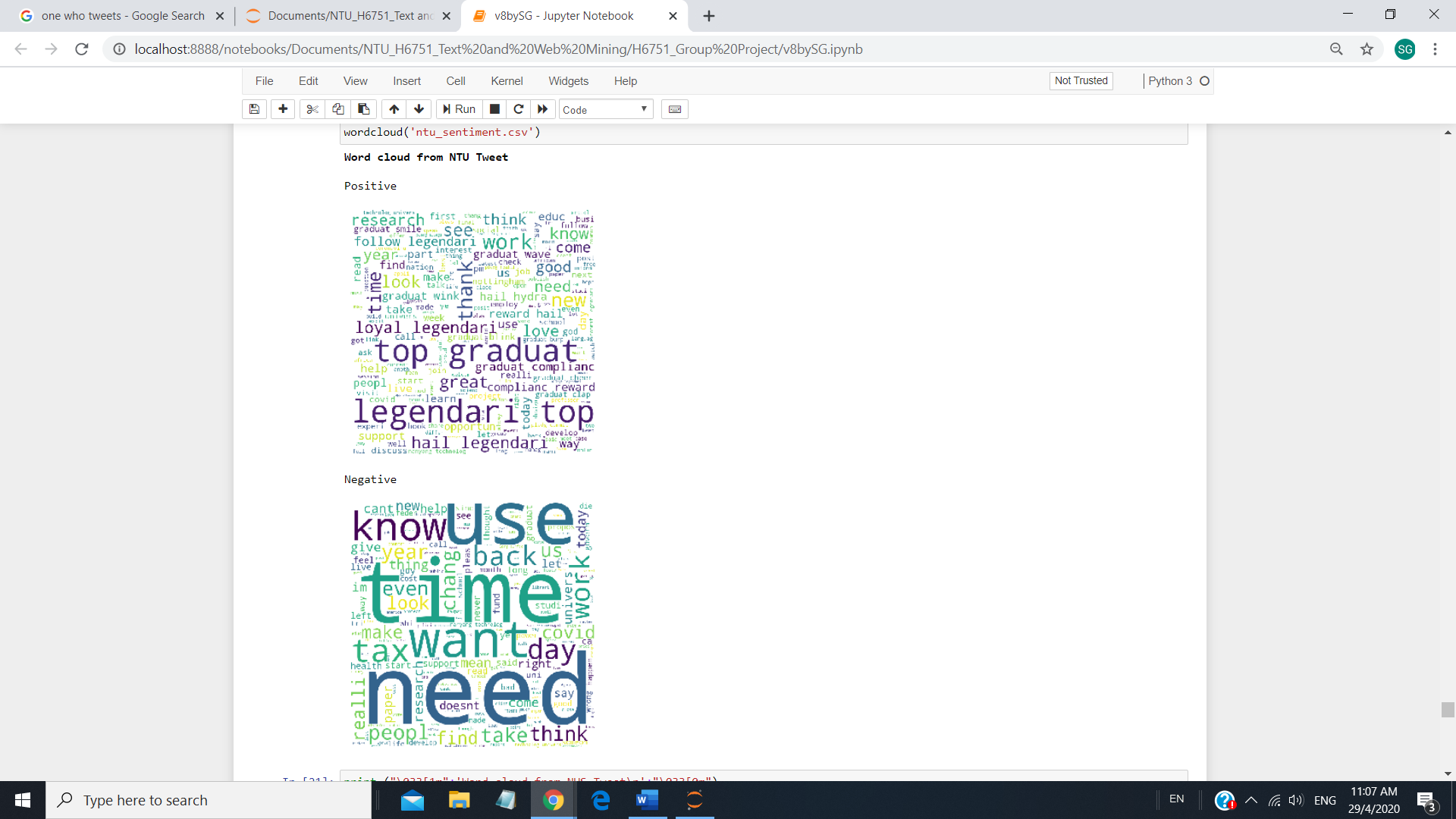
The figures below show the top 20 positive and negative word ranking and count from the three university datasets combined.



**4.4 Word Cloud**

Word cloud is an image comprising words such that the size of each word shows its importance or frequency. The more often a specific word appears in the tweets, the bigger and bolder it appears in your word cloud making visualization of important words easy (Parker, 2019). Word Cloud is used to represent the importance of tweets’ words for positive tweets and negative tweets from the three universities.

From the word cloud, all three universities positive tweets focus on ‘thank”, ‘good’ and “love”. NTU positive tweets are mainly overshadowed by ‘King Chody is the Legendary NTU SCSE SG Top Graduate’ tweets as highlighted in Section 4.2. ‘need’, ‘want’, ‘time’ and ‘day’ are featured prominently on all the negative tweets. Due the Covid-19 circuit breaker, all on-site examinations were cancelled and replaced with assessments (Clark, 2020), resulting in students needing/wanting more time (day) to complete the additional assignments and projects. The situation for the students is less acute as their mid-term examinations were cancelled due to Covid-19 concerns (Chen, 2020).



It is noted that ‘covid’ is featured prominently in both NUS positive and negative tweets indicating active discussion and comments on this topic among NUS tweeters. This could be due to increase awareness, impacts and concerns in their university due to an infected case of a NUS lecturer (Teng, 2020).

1. **DISCUSSION AND FUTURE WORK**

The project provides a basic sentiment analysis using Logistic Regression classifier using data from Kaggle as training dataset and successfully classify the tweets to better understand the sentiments at the three universities. VADER and word cloud were also successfully introduced to help gain insights to the positive and negative tweets.

Further work to improve the accuracy of the sentiment analysis can be incorporated. Emojis were created as symbolic representation of emotions and are used to represent body language and tone of voice in text-based communication. It can thus play an important role in representing a sentiment. Thus, the accuracy of our sentiment analysis could be further improved if the analysis of tweets incorporates emoji and emoticons (Chen et al., 2019).

The tweets extracted from the tweeters of the three universities were limited and collected over 2 separate periods due to tweeter API and time constraints. It would be more informative and revealing if tweets are collected continuously over a long period. This can helps provide a more accurate and generalised satisfaction level of the respective tweeters. The collection of tweets continuously would also allow for good understanding of the sentiment of tweeters during specific events such as examination, open house, career fair etc. This can help universities gauge the success of an event, identify gaps and take proactive actions.

1. **CONCLUSION**

In conclusion, our sentiment analysis shows that the ranking of the university does not correspond correspondingly to students’ satisfaction and sentiments. NUS is ranked as the top university in Singapore by both UniRank and The World University Rankings. NUS satisfaction level is the lowest among the top three university in Singapore based on the tweet datasets. THE World University Rankings (2020) methology does not include staff and students’ satisfaction but are based on teaching, research, citations, industry income and international outlook. Maybe the university rankings should include staff and students’ well-being and satisfaction level as these are the key talents and resources that make up the university.

The project provides the opportunity to explore and learn how to make use of Kaggle training datasets, extracting data from tweets, explore different classification algorithms, tuning of hyperparameters, cross validations and tools such as VADER and word cloud. The approaches to sentiment analysis are vast and diverse. Through continual learning, exploring and experience, users can develop better skills and knowledge in building an accurate and reliable sentiment analysis tool.

*The code is available at* [*https://github.com/yelynn1/universitydatasingapore*](https://github.com/yelynn1/universitydatasingapore)

**References:**

Bronshtein, A. (2017). *Train/Test Split and Cross Validation in Python*. Retrieved from <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>

Cassetta, M. (2019). Putting Structure Around Unstructured Data. Retrieve from <https://www.dataversity.net/putting-structure-around-unstructured-data/>

Chen, J. (2020). *SMU Cancels Exams While NUS & NTU Shift Large Classes Online Due To Covid-19 Concerns*. Retrieved from <https://mustsharenews.com/covid-19-smu-exams/>

Chen, Z., Cao, Y., Lu, X., Mei, Q. & Liu, X. (2019). *SEntiMoji: An Emoji-Powered Learning Approach for Sentiment Analysis in Software Engineering*. Retrieved from <https://arxiv.org/pdf/1907.02202.pdf>

Clark, T (2020). *Provost Message.* Retrieved from <https://crisisresponse.smu.edu.sg/sites/crisisresponse.smu.edu.sg/files/2020-03/Key%20changes%20to%20UG%3AJD%20teaching%20and%20exams%20for%20current%20semester.pdf>

Go, A., Bhayani, R. & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 1(2009), p.12.

Kaggle (n.d.). *Sentiment140 dataset with 1.6 million tweets. Sentiment analysis with tweets.* Retrieved from: <https://www.kaggle.com/kazanova/sentiment140>

Monsters, D. (2018). *Text Preprocessing in Python: Steps, Tools, and Examples*. Retrieved from <https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908>

Pandey, P. (2018). *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)*. Retrieved from

<https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f>

Parker, P. (2019). *Creating word clouds with python.* Retrieved from <https://towardsdatascience.com/creating-word-clouds-with-python-f2077c8de5cc>

Paul, S. (2018). *Hyperparameter Optimization in Machine Learning Models*. Retrieved from <https://www.datacamp.com/community/tutorials/parameter-optimization-machine-learning-models>

Sidana, M. (2017). *Types of classification algorithms in Machine Learning*. Retrieved from <https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14>

Srivastava, J. (2018). How to Extract Tweets from Twitter in Python. Retrieved from <https://medium.com/@jayeshsrivastava470/how-to-extract-tweets-from-twitter-in-python-47dd07f4e8e7>

Teng, A. (2020). *Coronavirus: All classes at affected NUS professor's faculty to be conducted online*. Retrieved from <https://www.straitstimes.com/singapore/health/coronavirus-all-classes-at-affected-nus-professors-faculty-to-be-conducted-online>

The World University Rankings (2020). *THE World University Rankings 2020: methodology.* Retrieve from <https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2020-methodology>

The World University Rankings (n.d.). *National University of Singapore.* Retrieve from <https://www.timeshighereducation.com/world-university-rankings/national-university-singapore>

Tripathi, M. (2018). *How to process textual data using TF-IDF in Python*. Retrieved from <https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/>

Twitter (n.d.). *Developer.* Retrieve from <https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline>

UniRank (n.d.). *Top Universities in Singapore.* Retrieve from <https://www.4icu.org/sg/>