

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6020 20906 Predictive Analytics**

**Professor Name: Dr. Yevgeniy Frenkel**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: B. 2020**

|  |
| --- |
| **Module 2: Naïve Bayes Model**  **Completion Date:April 19th Due Time:12:00am** |

**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

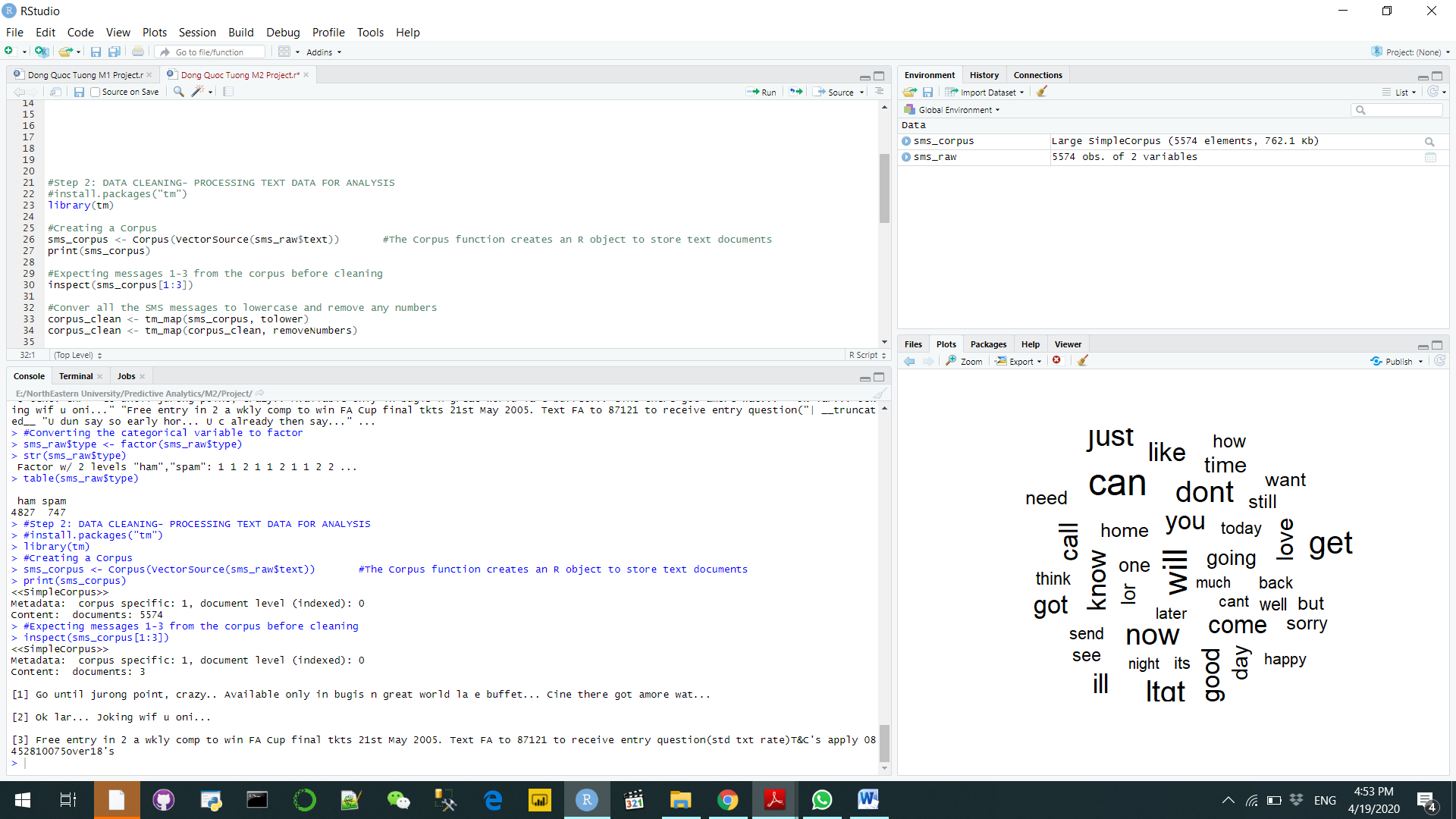
**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

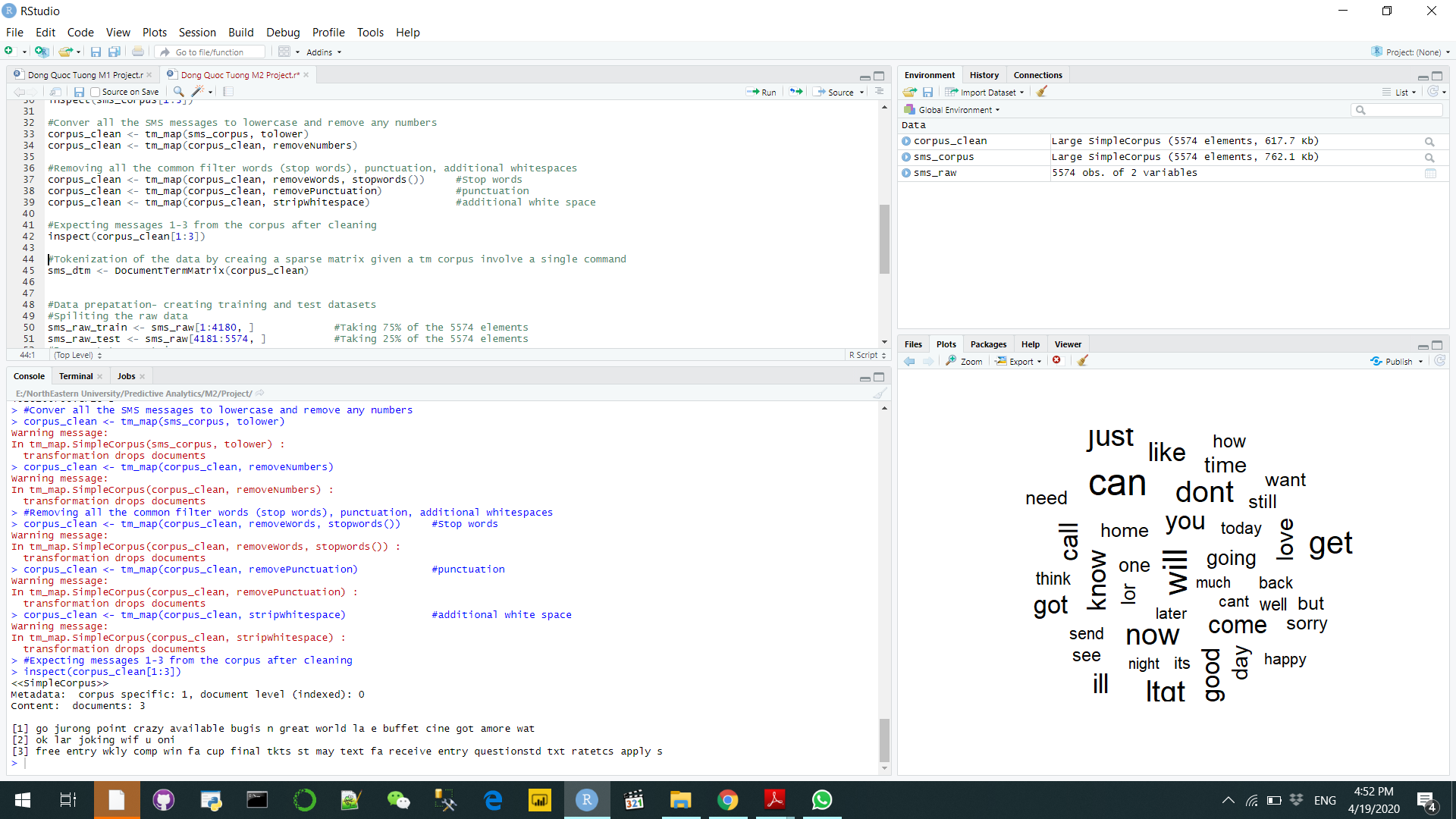
**Executive Summary**

Despite being considered outdated, Naïve Bayes theorem is still very effective in Natural Language Processing. In this exercise, we will try to classify a series of SMS message into either “Harmful” or “Spam”

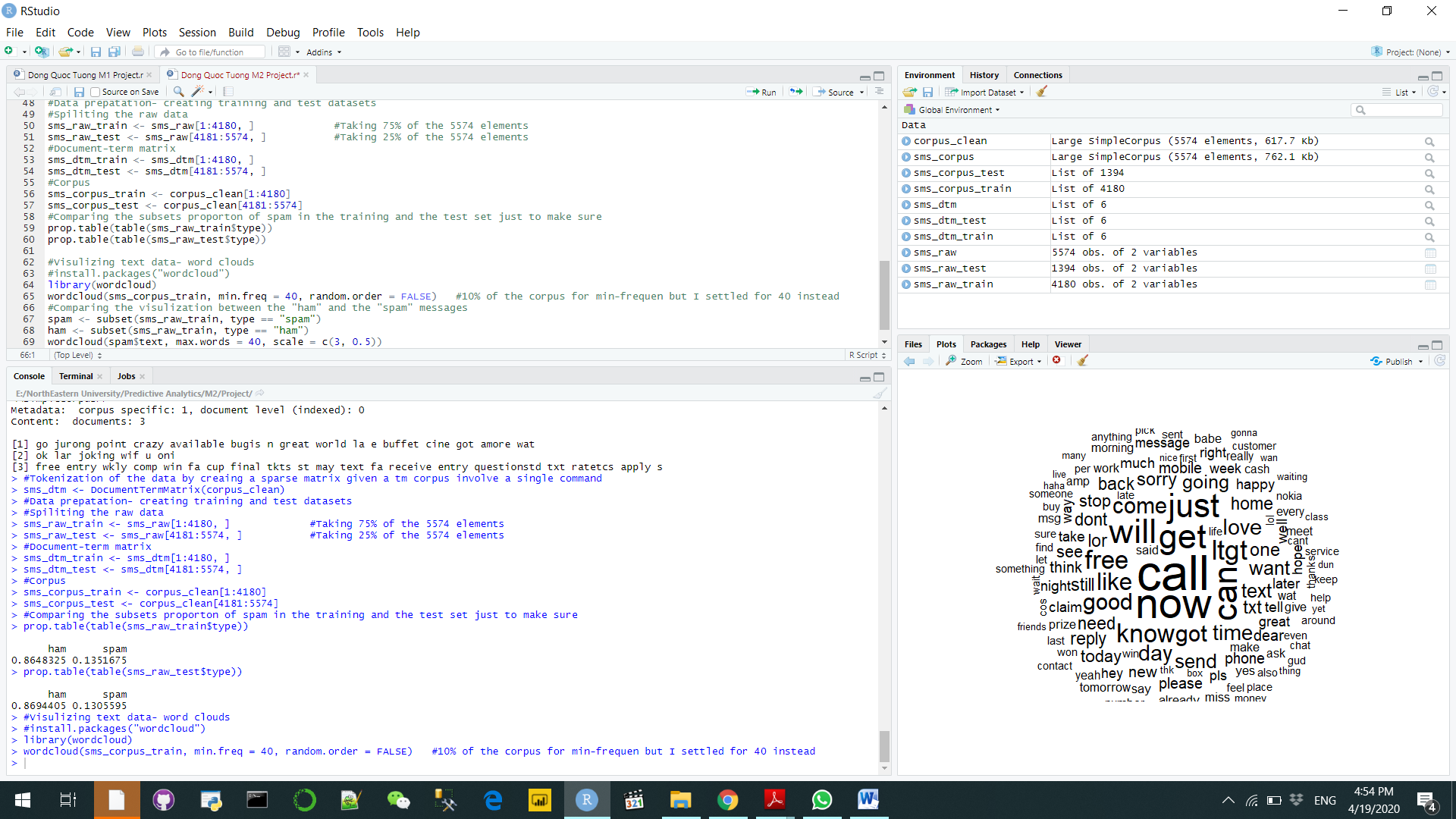
**Analysis**

The dataset consists of 5574 inputs. During the data preparation process we will turn all the words into lower cases, take out the common world filters, extra spaces, punctuations, as well as any words that appear less than 0.1% of the total words count. Therefore, we can see there is a massive difference between the “dirty” data and the “clean” data below





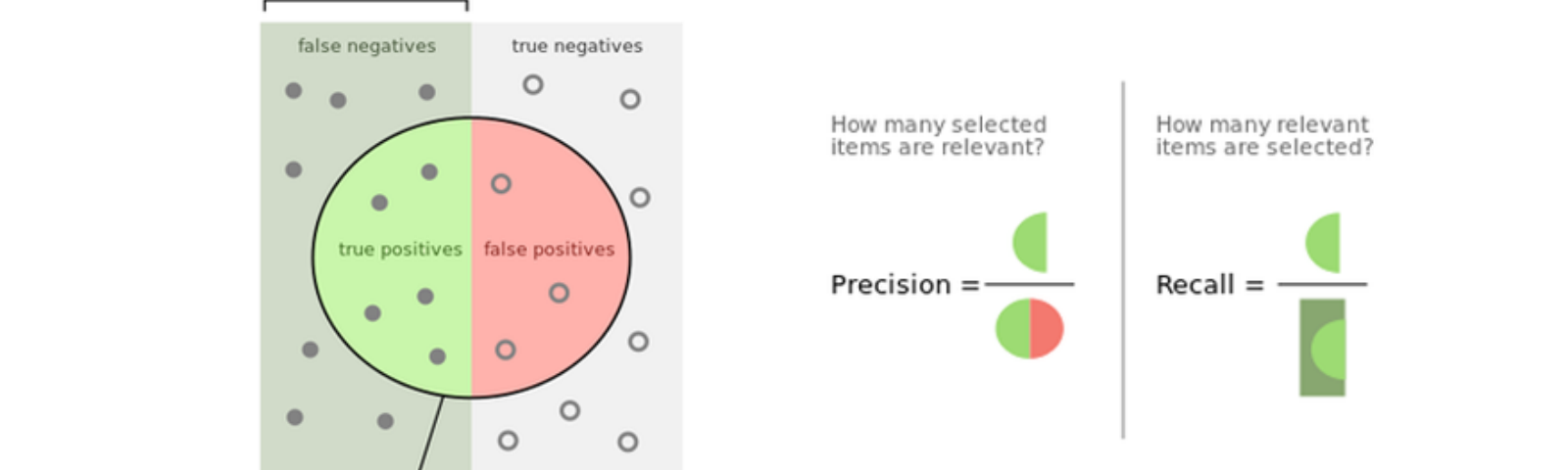
Finally, we created the Words Cloud below for the total of our dataset.



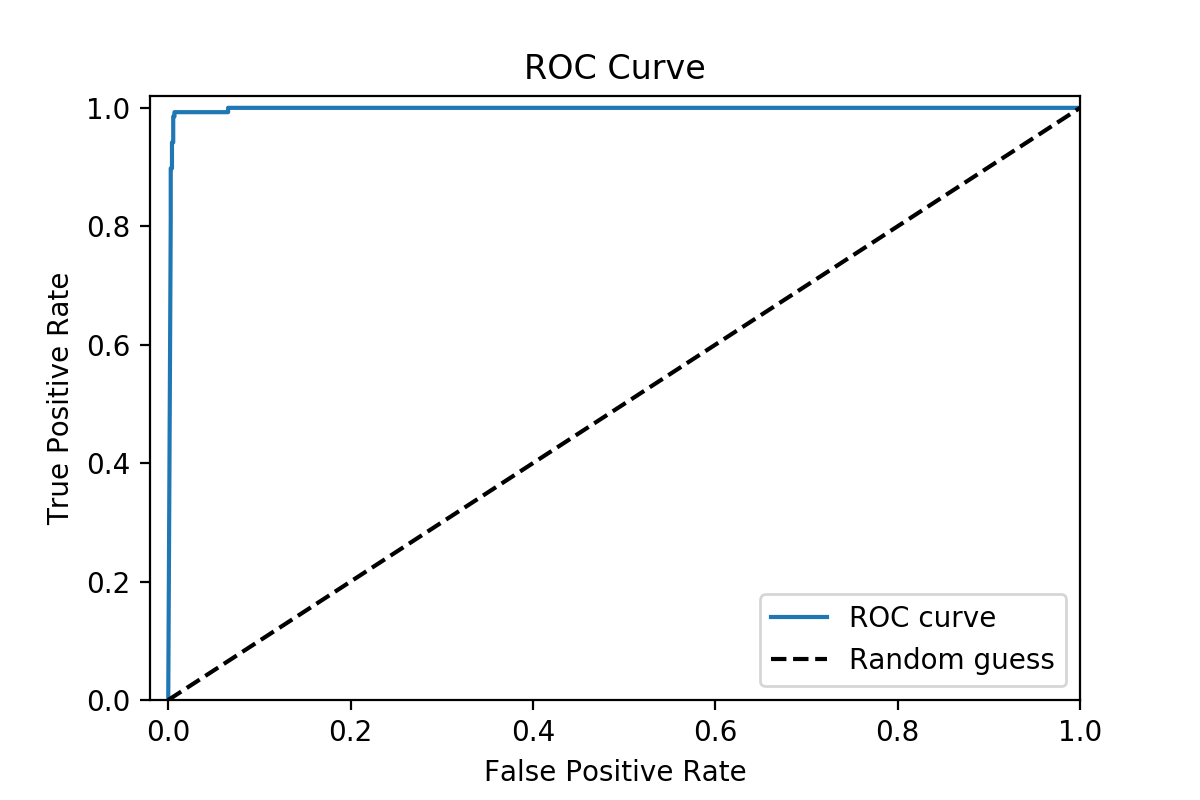
**Q1**

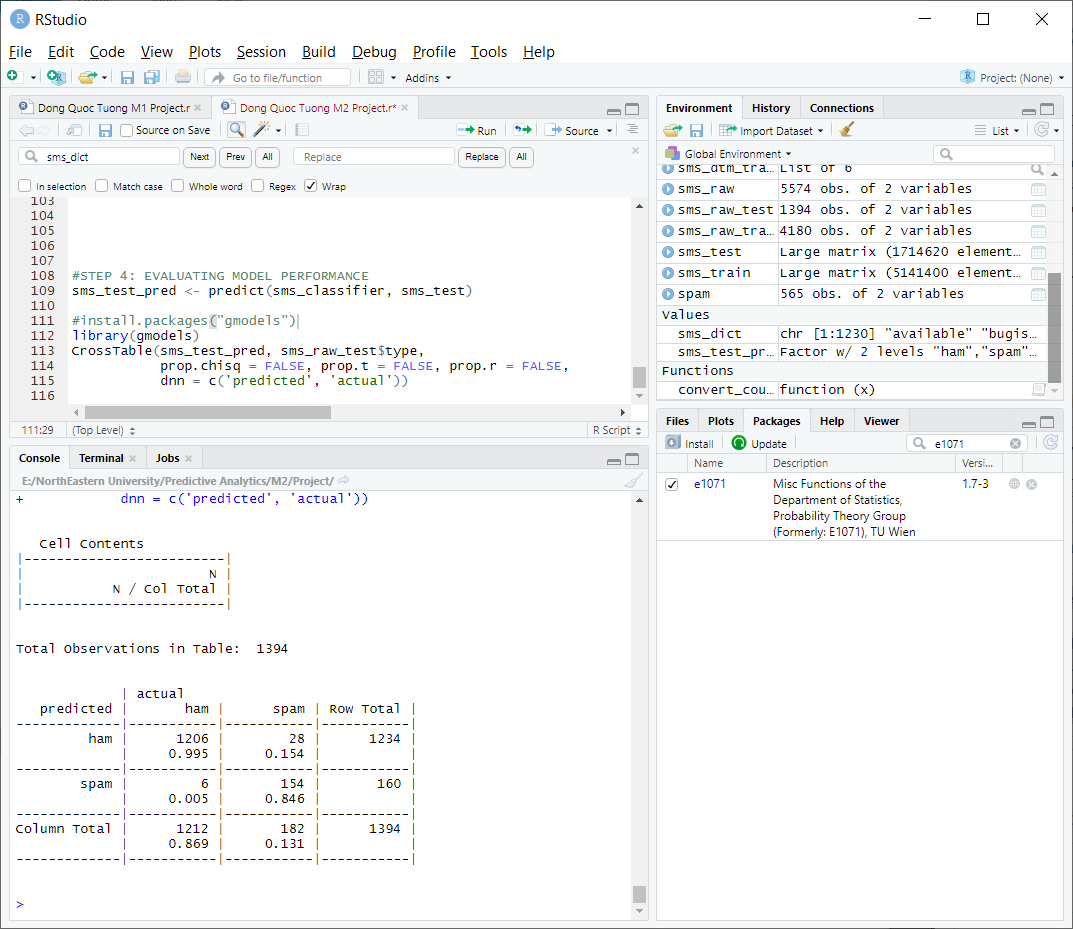
Like KNN model, Naïve Bayes also have multiple similar metrics to judge the performances of a classification for the dataset. They are the Precision and Recall and ROC (receiver operating characteristic) curve .

* As stated in the previous exercise, the Precision is defined as the percentage of the percentage of your model’s results that is indistinguishable to the test set. In addition, recall allows us to know the percentage of the total relevant results that has been classified by the model. The spectrum of these two numbers is between 0 and 1, the nearer to 0 these numbers are, the more incorrect they become. In this exercise, if the precision is nearly 1, it means that the model has done a great job in separating the Spam and Ham messages. (Saxena, 2018)

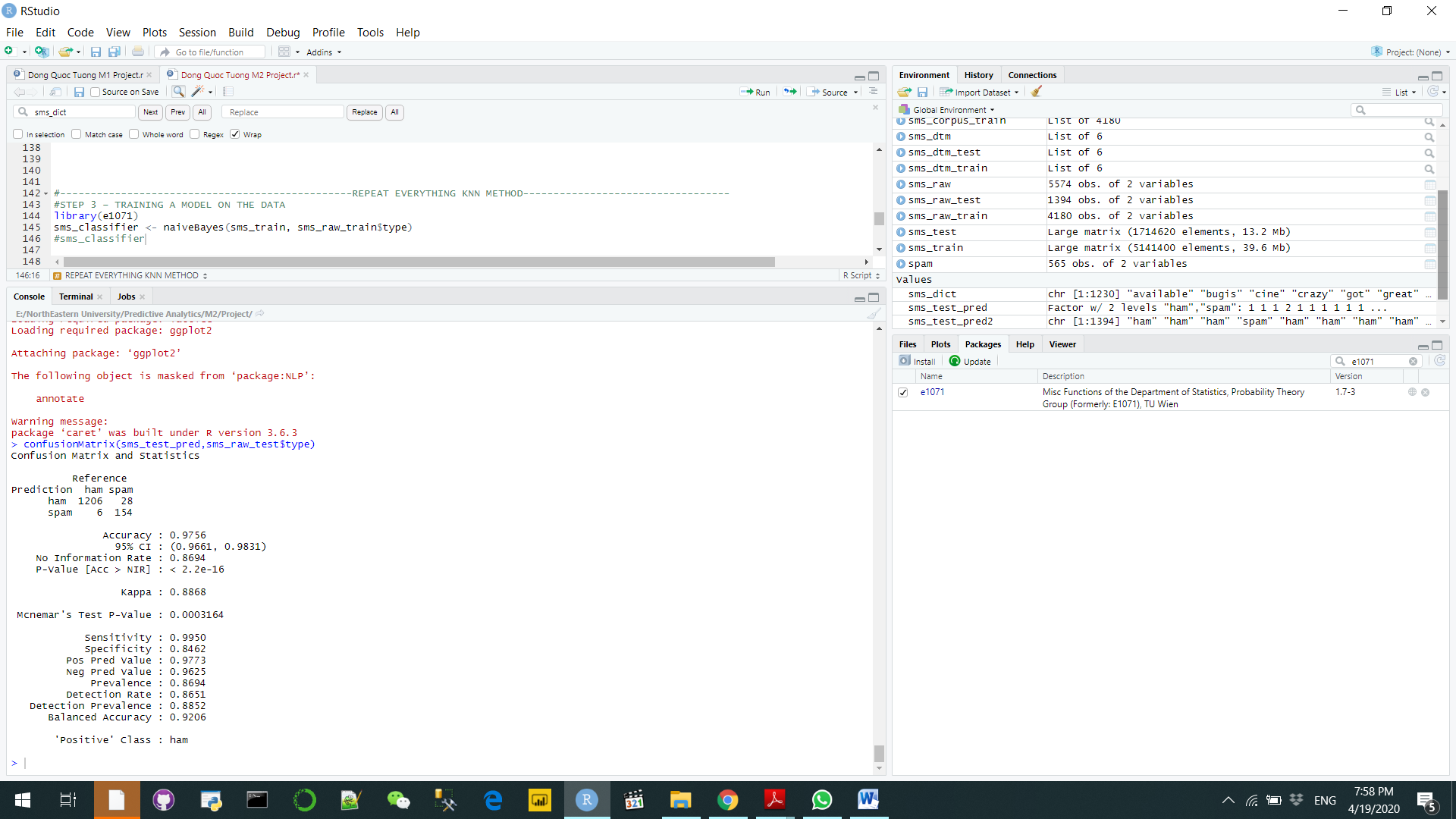


* ROC Curve is a more visual way to measure the performance of the binary classifier. It is established by plotting the true posiutive rate (TPR) against the Type 1 error rate. Then we need to define what is the threshold to be calssified, usually it is 0.5. One advantage of the ROC curve is it can visualize how the performance of the classifier changes as we vary the threshold



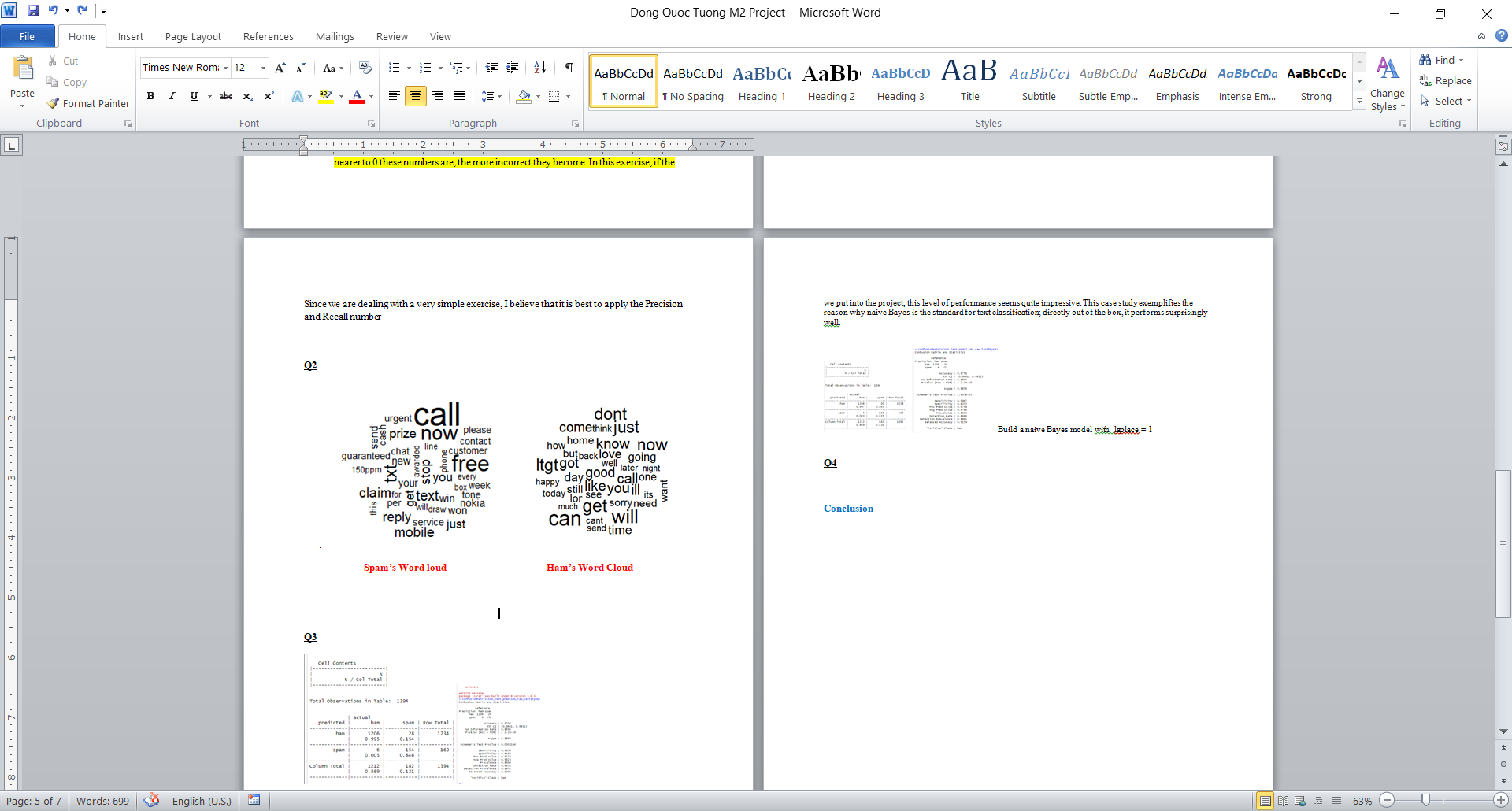


Looking at the table, we can see that 6 out of 1212 ham messages were incorrectly identified while spam messages detection failed by 28 cases. This level of performance is quite impressive and demonstrates why Naïve Bayes Model is still useful when Natural Language Processing. Right out of the gate, it performs surprisingly well



The Recall number (Sensitivity), which is 0.995, resonates the same sentiment to us with almost no faulty in our prediction. Then we also have a very high precision value, at 0.97, indicates that our models ‘ prediction are highly accurate when compared to the test set. Therefore, we can conclude that Precision and Recall is the metric that we will choose for our model

**Q2**



If we look a little bit closer, we see that spam message contains the word “free” while no ham messages have any of those words. On the other hand, quite a few ham messages mention the specific date of the week while spam ones have none. Furthermore, the Ham messages use more verbs while spam messages are filled with anything related to phone. It is clear that Spam messages try to make you spend more money on phone services while the Ham messages are quite neutral. We saw that the Ham messages accounts for 87% of the messages (4872) while only 1 out of ten messages are spam.



**Q3**



According to Laplace’s rule of succession, it is much better that α equals 1 (in which case the term add-one smoothing is also used), although a much smaller value is chosen in practice. It is proven when trying to use different Laplace values, the best number is 1 because the greater Laplace is the greater the wrong % number.

**Conclusion**

While Neural Network dominate the more complex text analyzing tasks, Naïve Bayes theorem approach is more suitable with simple text analyzing like this exercise. It is much better that we have a deep look on how we can also reduce dimensionality to increase success rate

**References**

Ferreira, H. (2018, April 4). Confusion matrix and other metrics in machine learning. Retrieved from <https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a>

Saxena, S. (2018). Precision vs Recall. Retrieved from <https://towardsdatascience.com/precision-vs-recall-386cf9f89488>