# CSC 425 525 Project Report Template (Project Title)

Instructor: Dr. Junxiu Zhou

Semester: Fall 2020

Team Members:

Spencer Peace

Katelynn Quinn

Jose Salazar

# 1. Introduction

An email that is received can either be an authentic, solicited or wanted email; or a spam email (unsolicited or unwanted). In this project package, we receive a set of old emails that have been classified either as Spam or Ham (non Spam), and our task is to build a program that detects whether new incoming emails are either Spam or Ham.

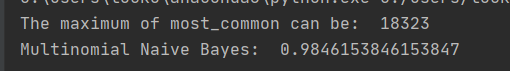
# 2. Implementation Process and Results

To begin, and after having read the training and testing data sets, and classified them as Spam or Ham should the email contain the Ham or Spam flag, the prior probability of an email belonging to a class was calculated by computing the conditional probability of each class based on the features of an email like the number of words and the most common word by generating a feature matrix for each email, as well as a parameterized distribution from the feature matrix for each class.

After that, a new inference was made about an email belonging to either the Spam or Ham classes based on its features like number of words and the frequency of the most common word, considering the prior probability and the commonalities of features like the words in the dictionary.

Finally, an email was classified as either Spam or Ham depending on the highest probability of it belonging to any of the 2 classes by computing the elementwise product for both classes.

The result is shown on the following picture:



# 3. Comparison Analysis

No comparison was made as the only method that was implemented was the Multinomial Naive Bayes.

# 4. Bonus Implementation and Results (if applicable)

[Not applicable]

# 5. Team Members Responsibility Statement

* Spencer:
* Kate:
* Jose:
  + I was responsible the report and initial low-level implementation of the Multinomial algorithm.

# 6. Conclusion

In conclusion, we found that classification algorithms are heavily dependent on historical data and statistics, and one of the easiest and more straight forward ways to conduct classification analysis is to implement a Multinomial Naive Bayes algorithm.

In terms of space complexity, there is not much that can be done to improve it, as it is expected that features are to be extracted from the data set and various matrices are to be filled up with such features as well as their conditional probabilities to belong to a class. However, in terms of the time complexity, there are other algorithms that could reduce the complexity from exponential to linear.