# Logistic Regression Spam Filtering
## Csci 183: Final Project

R. Johnson, G. Nguyen, R. Young
*Santa Clara Univeristy*
May 10, 2017

**Abstract**

Spam messages are a nuisance to everyone. In this day and age, people encounter short messages, such as SMS or other messaging applications on a day to day basis. Spam must also take on a short form in order to be deliverable on these platforms. As a result, classifying a message as spam may present a more difficult task as there are less words in the message to work with. The goal of our project is to use logistic regression in order to determine if a short message should be classified as spam or not. This will involve training our algorithm to identify which terms are most likely to be associated with spam (i.e. 'FREE', 'CASH', links, etc). Then, we will use our trained algorithm to categorize a test set of messages as spam or not spam.

*Keywords:* spam , sms , email , filtering , classification , logistic regression

## Goal

To predict whether a message is spam or not.

Training Phase: Get data from email messages that are already categorized spam/ham. Train data based on hypothesis function $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$. If $h_\theta(x)$ is close to 1, then it would belong to the spam class. Otherwise, we will classify it as ham.

Test Phase: Given an email message, predict whether it is spam or ham. Check if the hypothesis function is closer to 0 or 1. Use this to classify the data.

## Proposed Plan

**Language:** Python
**Data Source:** https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
**Algorithm(s):** Logistic Regression
**Worload Distribution:**

Raya: Data Munging and Cleaning
Ryan: Algorithm Development
Grace: Final Project Development

## Proposed Timeline

May 10th: Project Proposal Due (await feedback from Dr. Manna).

May 11th- 15th: Finish up preliminary research, data gathering, finalizing which libraries/algorithms to use.

May 15th- 24th: Each member works on their portion of code.

May 24th- 30th: Combine code and test algorithm on test data, finalize code.

May 31st-June 4th: Prepare project write-up and presentation.

June 5th- 9th: Project due. Presentation is prepared by this time.

## Contact

Ryan Johnson: rtjohnson@scu.edu
G. Nguyen: gnguyen@scu.edu
Raya Young: rlyoung@scu.edu

## References

1. Spam Filtering for Short Messages
https://pdfs.semanticscholar.org/d457/4461f72712c025df14d1a3d4d73ef86ed23b.pdf