Classifing HRC's Emails by Sender

Pranay Singal, Kensen Tan, Renee Sweeney, Abigail Chaver

December 2, 2016

Introduction

In this project, our goal is to classify the sender of emails found on Hillary Clinton's email server. We will use a word frequency table as our primary source of predictors, and will use 3 modeling approaches: Random Forest, Support Vector Machine, and K-means Clustering. We will compare the results of our supervised methods to select our final model.

Data Processing

The initial data consisted of a tab separated file with two columns: the sender code (labeled 1-5) and the string of the email.

After reading in the tsv, we converted the string to a character vector. We then wrote those vectors to a corpus so that we could use the R package `tm` to remove punctuation, numbers, and stop words; convert to lower case; and stem.

Next, we removed common words that were seen in every email, such as "subject", and "US".

We also explored other possible features that could predict sender, including number of words, average word length, and rate of use for ampersands, semicolons, question marks, and upper case. We found that the only variable which differed significantly between groups was number of words per email, so we added that as a feature to our matrix.

Steps	Terms
Raw	38066
Stop Words	37966
Stemmed	27504
Frequency Selection	9177
Additional Features	9178

Random Forest Model

Explanation...

Step	Number of Features Used	Accuracy	Accuracy by Class
RF	0	0	0
RF with Feature Selection	0	0	0

Top Ten Features

Rank	Feature
1	a

Rank	Feature
2	b
3	c
4	d
5	e
6	f
7	g
8	h
9	i
10	j

SVM Model

Explanation...

Table 4: Table continues below

Step	Number of Features Used	Accuracy
SVM	0	0
SVM with Feature Selection	0	0

Accuracy by Class
0
0

Top Ten Features

Rank	Feature
1	a
2	b
3	c
4	d
5	e
6	f
7	g
8	h
9	i
10	j

K-means Clustering

Model Selection

Compare MSEs here.

Predict Classifications on Test Set

Using our best model, we used `predict` to generate classifications for each email in the test set. # Conclusion