Cmpsci585
Project Proposal
Patrick Carron
Raymond Zhu
11/11/2016

Predicting Political Party Affiliation from Speech

1 Introduction

Description of Problem
Describe Corpus
Describe Baseline Algorithms

2 Dataset

Number of Documents in Training Set
Number of Tokens in Training Set
Size of Vocabulary $|V|$ in Training Set
Discussion of Dev Set?
Discussion of Data Preprocessing (names removed and replaced, spaces inserted, forced to lower case, punctuation remains)
Number of Documents in Test Set

3 Software and Codestructure

Used Scikit learn
Created an analysis Pipeline
-Consistent easy preprocessing for different classifiers
-Easily preform cv and tune hyper parameters
-Write code to create charts and retrieve vocab and token metrics
-Easily add TDIDF

4 Baseline Algorithm

Naive Bayes, assumptions and Results
-Multinomial NB because independent party
-Multinomial NB also allows for smoothing in sclearn
SVD results as well
Discuss accuracies.

5 Timeline

Implement Cross validation code for each classifier 11/19
Write code to create cv graphs over range of hyper parameters 11/26
Draft of Poster and all exhibits 12/3
Draft of Final paper 12/7
Finalize Poster 12/10
Final Draft of Paper 12/10