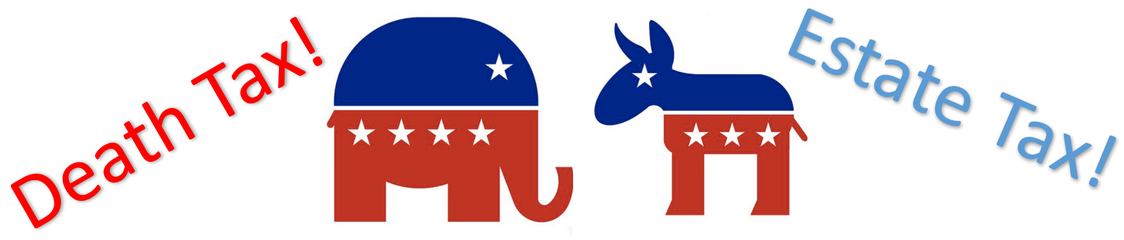
Predicting Party Affiliation from Political Speech

Raymond Zhu **Patrick Carron**

Goals

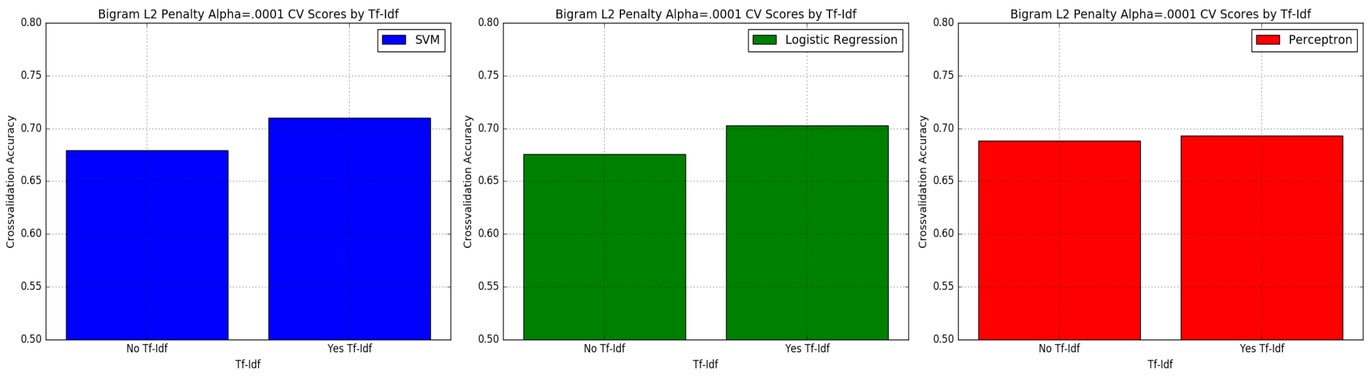
- Determine party affiliation based off Congressional speeches based on unique terms used by each group or differences in frequency of term use.



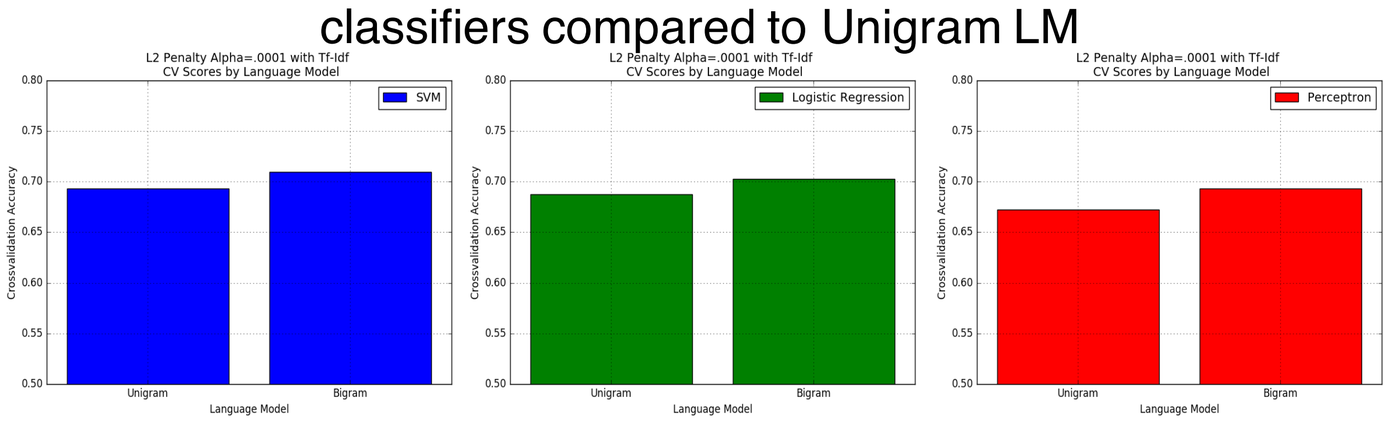
- Find a way to parse the filename and create a sparse vector representation.
- Compare classification algorithm and find optimal hyper-parameters.

Models

- We used Multinomial NB and Stochastic Gradient Descent.
- We looked at the smoothing parameters for Multinomial NB. We picked Multinomial NB because there was a chance the features were independent.
- Stochastic Gradient Descent:
  - Loss Functions Considered:
    - Hinge,Logistic, Perceptron
  - Regularization Penalties Considered:
    - L1, L2, None.
  - Range of Alphas Searched.
- Used 3-fold cross validation In order to find optimal hyper-parameters.
- Tf-Idf Increased accuracy for all classifiers.

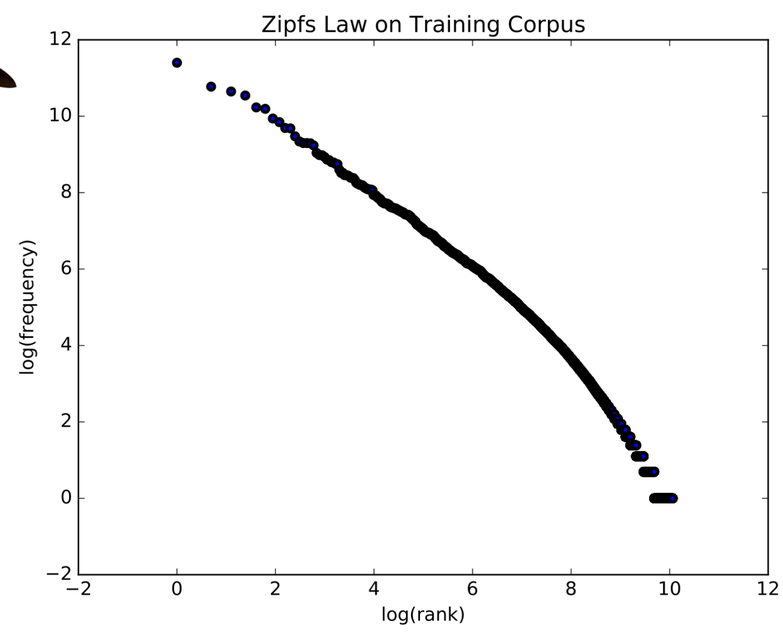


- Bigram LM improved accuracy for all classifiers compared to Unigram LM



Data

- Congressional Speech data set by Lillian Lee from Cornell.
- Training set is 5660 documents, with 1.3 million tokens.
- Single terms found:
  - herzog
  - surfrider
  - hamburglar
  - blazed



Results

- Highest test accuracy was **74.9%** using Stochastic Gradient Descent with Hinge loss with a L2 penalty, Bigram LM, Tf-Idf, and Alpha=.0002.
- Highest NB accuracy was **70.6**% accuracy. Alpha of 0.05.
- Our accuracy of predicting party affiliation beat their accuracy by 3% of trying to predict voting outcomes for speakers.