# CS 583 Project 2: Twitter Data Mining

Xuelong Wang & Shuangxi Zhu

University of Illinois at Chicago

April 5, 2017

# Data Pre-processing

- Data cleaning
  1. Removed ”mixed” category data;
  2. Filtered missing-label data;
  3. Discarded labels out of the range, e.g. !!!.
- Features extraction
  1. Each word is considered as a variable;
  2. Unigram-model is chosen.
- Vectorization:
  1. Transformed the data into numerical feature vectors;
  2. Re-weighted data: tf-idf (lower the influence of those stopping words)

# Classification and Results

- blue represents Obama; red represents Romney

| Algorithms | Results | | | | | | |
|---|---|---|---|---|---|---|---|
| | Positive class | | | Negative class | | | Overall |
| | p | r | F | p | r | F | Acc |
| Naive Bayes | 0.46 | 0.64 | 0.54 | 0.51 | 0.52 | 0.52 | 0.48 |
| | 0.28 | 0.48 | 0.35 | 0.62 | 0.43 | 0.51 | 0.42 |
| Logistic | 0.67 | 0.51 | 0.61 | 0.59 | 0.63 | 0.61 | 0.59 |
| | 0.60 | 0.26 | 0.36 | 0.58 | 0.86 | 0.70 | 0.57 |
| SVM | 0.64 | 0.60 | 0.62 | 0.60 | 0.63 | 0.62 | 0.60 |
| | 0.56 | 0.41 | 0.47 | 0.64 | 0.77 | 0.70 | 0.58 |

# Next Steps

- Data Balance: Tweets voting for Romney is half less than those are against him.
  Solution: Augment data (weighted data)
- Feature Selection: Elastic Net, Forward/Backward feature selection, etc;
- Parameter Tuning: kernel function, penalized parameter
  Solution: Grid search