

Visual Text Analysis Hackathon

Paul Nulty

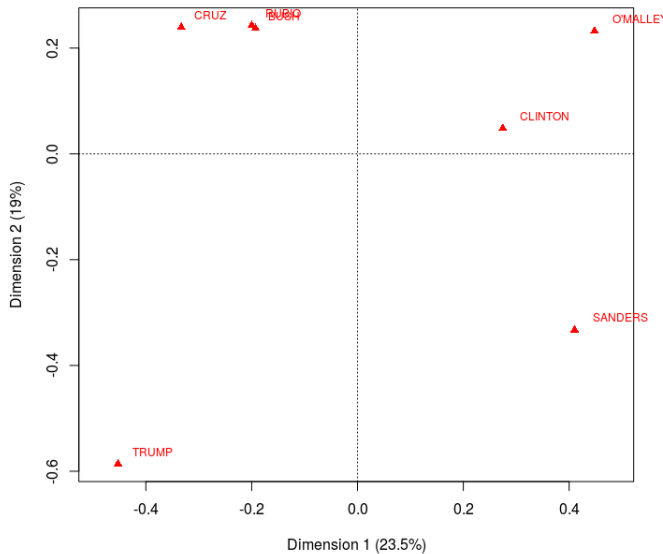
February 14, 2018

S

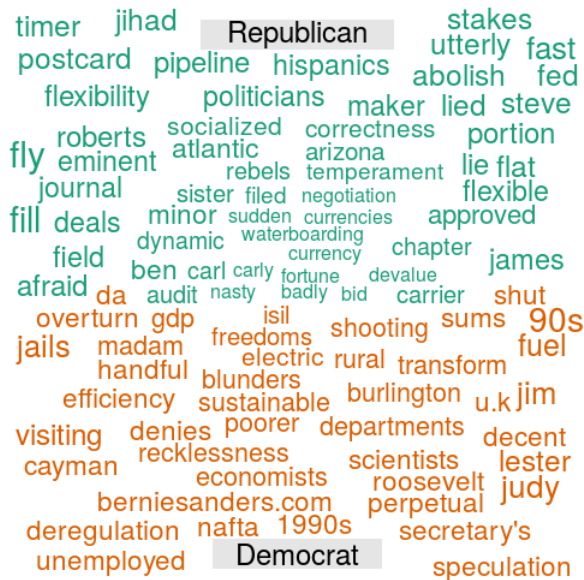
Supervised and Unsupervised Document Scaling

- ▶ Documents converted to bag-of-words/features
- ▶ Pre-processing: tokenization, entity recognition, lemmatization, stopwords removal, feature selection/weighting

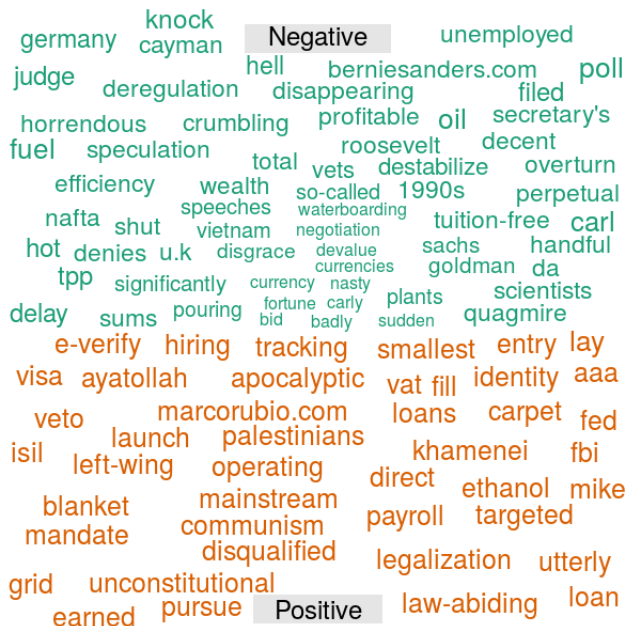
Two dimensional Correspondence Analysis



Dimension One



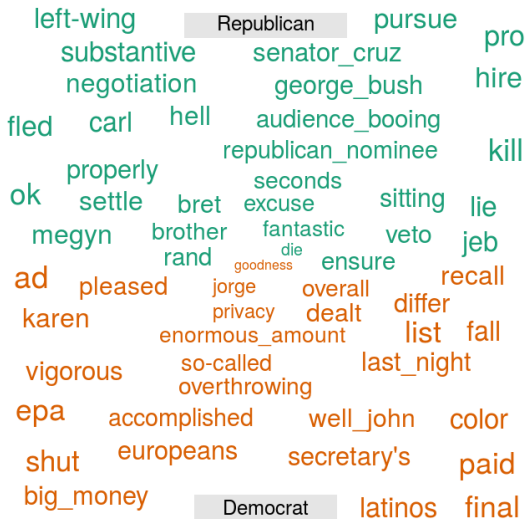
Dimension Two



Alternative: regularised regression, words as predictors

- ▶ glmnet elastic net, trigrams, bigrams and unigrams
- ▶ Keep features occurring more than five times
- ▶ Five-class multinomial logistic model for speaker
- ▶ 3195 parameters (approx. 800 nonzero at min CV error), 1470 documents, cross. val. accuracy: 78.7%
- ▶
- ▶ binomial model for party
- ▶ 3195 parameters (approx. 1481 nonzero at min CV error), 1470 documents, cross. val. accuracy: 89.5%

Party regression model



Tools

- ▶ General R: RStudio, dplyr
- ▶ Models: ca, glmnet
- ▶ Text: quanteda
- ▶ Vis: shiny, wordcloud::comparison.cloud

thank you

- ▶ Thank you for your attention

- ▶ Code available

<https://github.com/pnulty/text-hackathon-2016>