Visual Text Analysis Hackathon

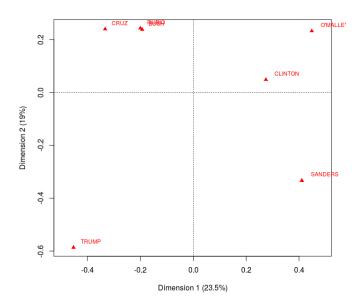
Paul Nulty

February 14, 2018

Supervised and Unsuperised 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

```
timer jihad
                                              stakes
                     Republican
                                          utterly fast
postcard pipeline hispanics abolish fed
   flexibility politicians maker lied steve
    roberts socialized correctness portio arizona lie flat
                                                portion
fill deals field negotiation flexible minor sudden currencies approved dynamic waterboarding currency chapter jame
                                               flexible
afraid da audit nasty badly bid carrier
overturn gdp isil shooting sums 90s jails madam handful blunders blunders burlington isil shooting sums 90s
                   sustainable burlington II k jim
     efficiency
visiting denies poorer departments
            recklessness scientists lester roosevelt judy
                                                  lester
 cayman
 berniesanders.com perpetual deregulation nafta 1990s secreta
                                        secretary's
                       Democrat
  unemployed
                                            speculation
```

Dimension Two

knock Negative unemployed germany cayman hell berniesanders.com judge deregulation disappearing filed horrendous crumbling profitable oil secretary's decent roosevelt fuel speculation total vets destabilize overturn efficiency wealth so-called 1990s perpetual speeches waterboarding nafta shut vietnam negotiation tuition-free carl sachs handful hot denies u.k disgrace devalue currencies goldman da tpp significantly currency nasty scientists plants delay sums pouring fortune carly plants bid badly sudden quagmire e-verify hiring tracking smallest entry lay visa ayatollah apocalyptic vat fill identity aaa marcorubio.com loans carpet fed veto launch palestinians khamenei fbi isil left-wing operating direct ethanol mike mainstream blanket payroll targeted communism mandate disqualified legalization utterly grid unconstitutional law-abiding loan Positive pursue

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

```
left-wing
                             pursue
              Republican
                                     pro
    substantive senator cruz
                                     hire
    negotiation george bush
            hell
                   audience booing
      carl
fled
                republican nominee
                                     kill
    properly
                   seconds
ok
                             sitting
     settle
             bret
                  excuse
           brother
                  fantastic
                             veto
 megyn
            rand
                        ensure
                                  recall
     pleased jorge
                       overall
                              differ
                privacy dealt
 karen
                               list fall
            enormous amount
             so-called last night
 vigorous
             overthrowing
epa accomplished well_john color
        europeans secretary's
 shut
                                   paid
 big money
                                   final
                 Democrat
                           latinos
```

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