Advanced Quantitative Techniques (Week 5)

Gregory M. Eirich

Agenda

1. Comparing texts-- dis/similarity measures
2. Clustering texts to find likely authorship

Quick update on multinomial in Python

Statsmodels does not like categorical outcomes anymore, so instead you need to manually renumber the categories so that 1 is the middle category out of 3:

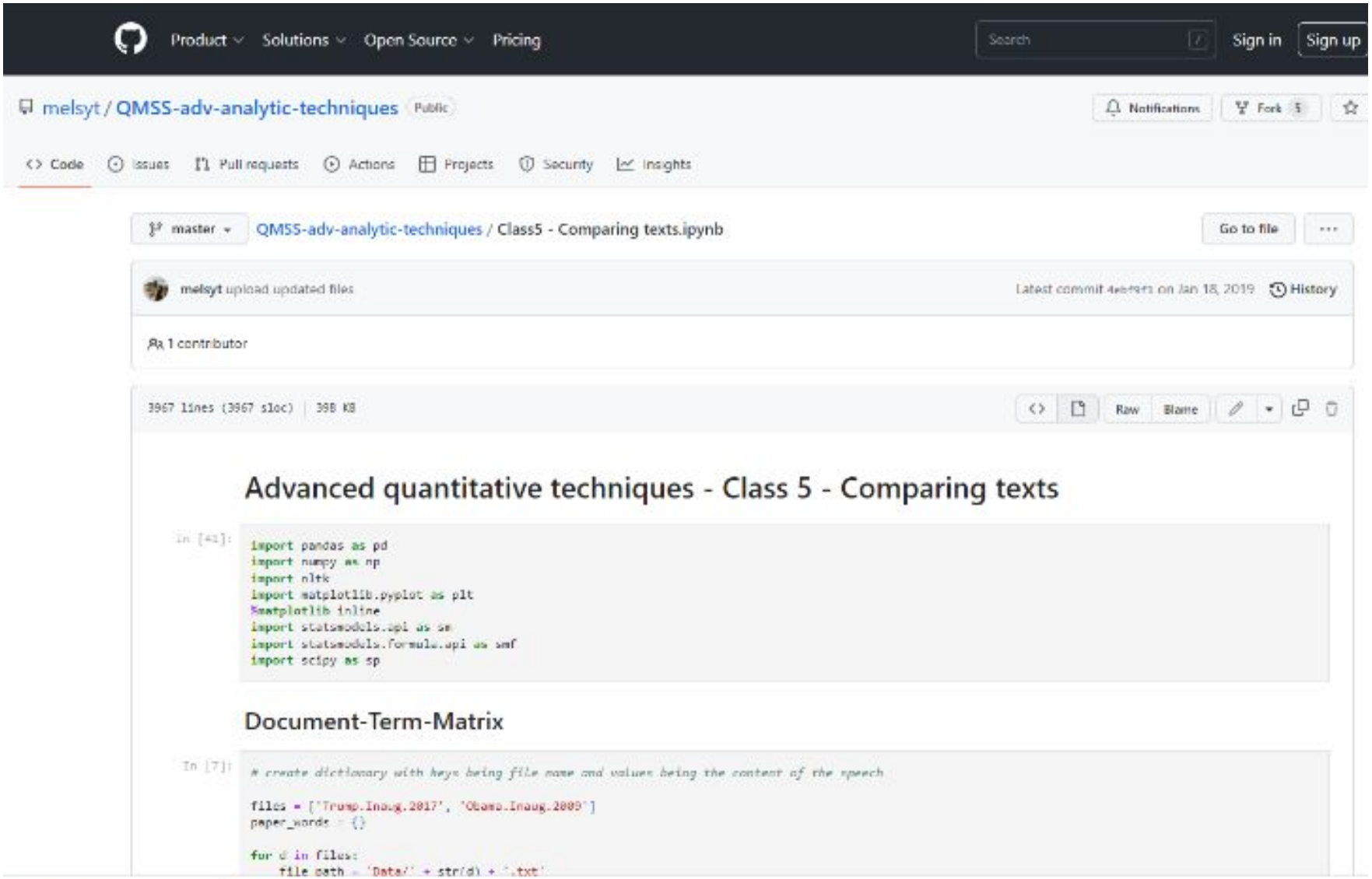
```
gss2['r2']=gss2['rnatchld'].astype('category')
gss2['r2'] = gss2['r2'].cat.reorder_categories([2,1,3])
gss2['r2'] = gss2.r2.cat.codes


# note that r2=1 refers to rnatchld = 1 and
# r2=2 refers to rnatchld = 3
# rnatchld = 2 is the reference against which others are compared
# this is due to smf not being able to handle categorical dependent variables well

ml1 = smf.mnlogit('r2 ~ sex + age + educ + prestg80 + C(region) ', gss2).fit()
ml1.summary()
```

Our example

Comparing inaugural addresses

```
library(tidyverse)
library(tidytext)
library(stringr)
library(SnowballC)

## get Obama SOUs from 2013-2016 ##

paper_words <- data_frame(file = paste0("/Users/gregoryeirich/Downloads/",
                      c("Trump.Inaug.2017.txt", "Obama.Inaug.2009.txt"))) %>%
  mutate(text = map(file, read_lines)) %>%
  unnest() %>%
  group_by(file = str_sub(basename(file), 1, -5)) %>%
  mutate(line_number = row_number()) %>%
  ungroup() %>%
  unnest_tokens(word, text)  %>%
  anti_join(stop_words) %>% ## remove stop words
  mutate(word = wordStem(word)) ## stemming the words
```

Python code is coming …



Obama's 2009 Inaugural Address

The beginning verbatim:

Vice President Biden, Mr. Chief Justice,members of the United States Congress, distinguished guests, and fellow citizens:

Each time we gather to inaugurate a President we bear witness to the enduring strength of our Constitution. We affirm the promise of our democracy. We recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names. What makes us exceptional -- what makes us American -- is our allegiance to an idea articulated in a declaration made more than two centuries ago:
"We hold these truths to be self-evident, that all men are created equal; that they are endowed by their Creator with certain unalienable rights; that among these are life, liberty, and the pursuit of happiness."

Today we continue a never-ending journey to bridge the meaning of those words with the realities of our time. For history tells us that while these truths may be self-evident, they've never been self-executing; that while freedom is a gift from God, it must be secured by His people here on Earth. The patriots of 1776 did not fight to replace the

Trump's 2017 Inaugural Address

The beginning verbatim:

Chief Justice Roberts, President Carter, President Clinton, President Bush, President Obama, fellow Americans, and people of the world: thank you.

We, the citizens of America, are now joined in a great national effort to rebuild our country and to restore its promise for all of our people. Together, we will determine the course of America and the world for years to come.

We will face challenges. We will confront hardships. But we will get the job done.

Every four years, we gather on these steps to carry out the orderly and peaceful transfer of power, and we are grateful to President Obama and First Lady Michelle Obama for their gracious aid throughout this transition. They have been magnificent.

Today's ceremony, however, has very special meaning. Because today we are not merely transferring power from one Administration to another, or from one party to another – but we are transferring power from

Our example

The dataset looks like this

```
> head(paper_words)
# A tibble: 6 × 3
                file line_number      word
               <chr>       <int>     <chr>
1 Obama.Inaug.2009          55     light
2 Obama.Inaug.2009          55   preciou
3 Obama.Inaug.2009          55 uncertain
4 Obama.Inaug.2009          55      call
5 Obama.Inaug.2009          55    answer
6 Obama.Inaug.2009          55     dedic
```

How much of these speeches do we lose if we remove stop words?

All words = 3,538

After removing stop-words = 1,361

Our example

Change the dataset into a dataframe with words as rows and columns as the speeches

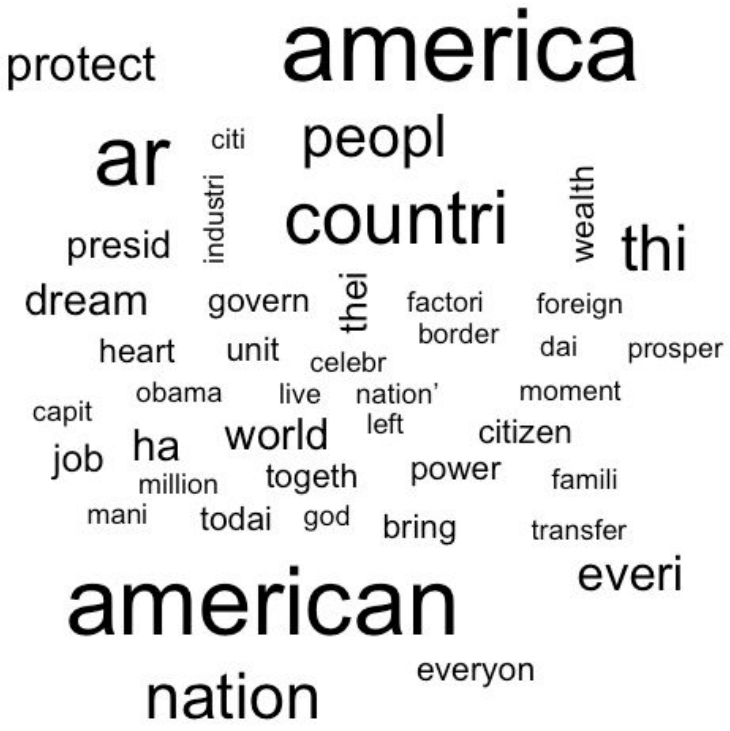
```
library(tm)

## count each word per speech

pw = paper_words[,c("file","word")]
d=  count_(pw, c("file", "word"))

## make a document term matrix ##

pwdtm = d %>%
  cast_dtm(file, word, n)

## make the dtm into a dataframe ##

mpwdtm=as.matrix(pwdtm)
df.mpwdtm=as.data.frame(mpwdtm)

## make the dtm into a tdm instead ##

t.t = t(mpwdtm)
```

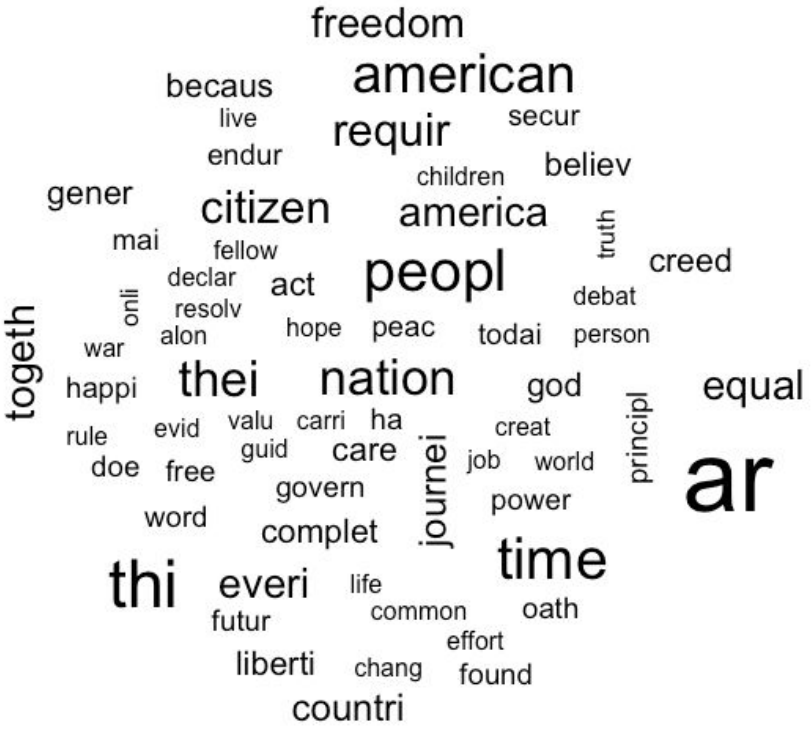
It looks like this

Change the dataset into a dataframe with words as rows and columns as the speeches

```
> head(t.t, 50)
             Docs
Terms         Obama.Inaug.2009 Trump.Inaug.2017
  1776                       1                 0
  40                         1                 0
  400                        1                 0
  abroad                     1                 0
  absolut                    1                 0
  act                        5                 0
  action                     1                 2
  advanc                     2                 0
  affirm                     1                 0
  afford                     1                 0
  africa                     1                 0
  ago                        1                 0
  agre                       1                 0
  allegi                     1                 2
  allianc                    1                 1
  america                    7                17
  america'                   2                 0
```

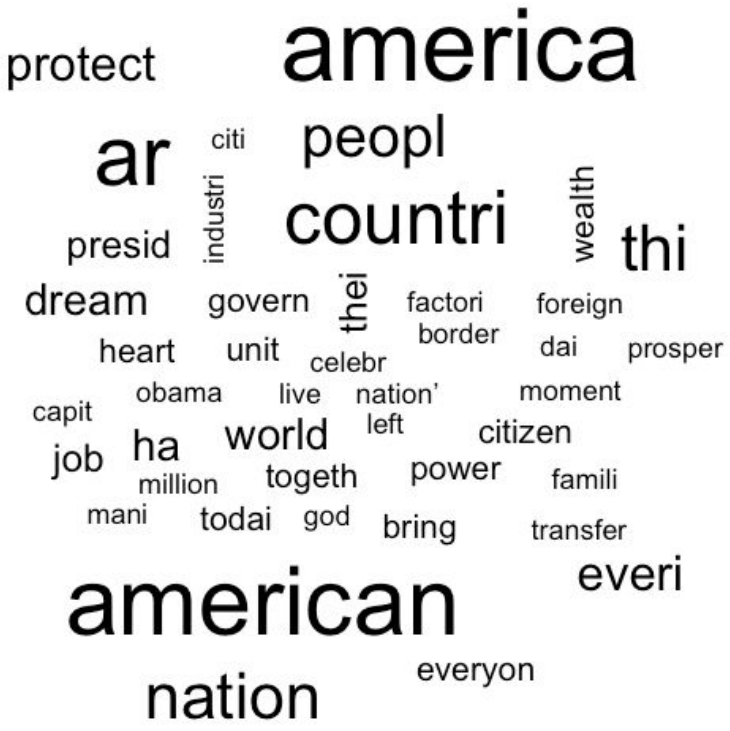
Which one is which?





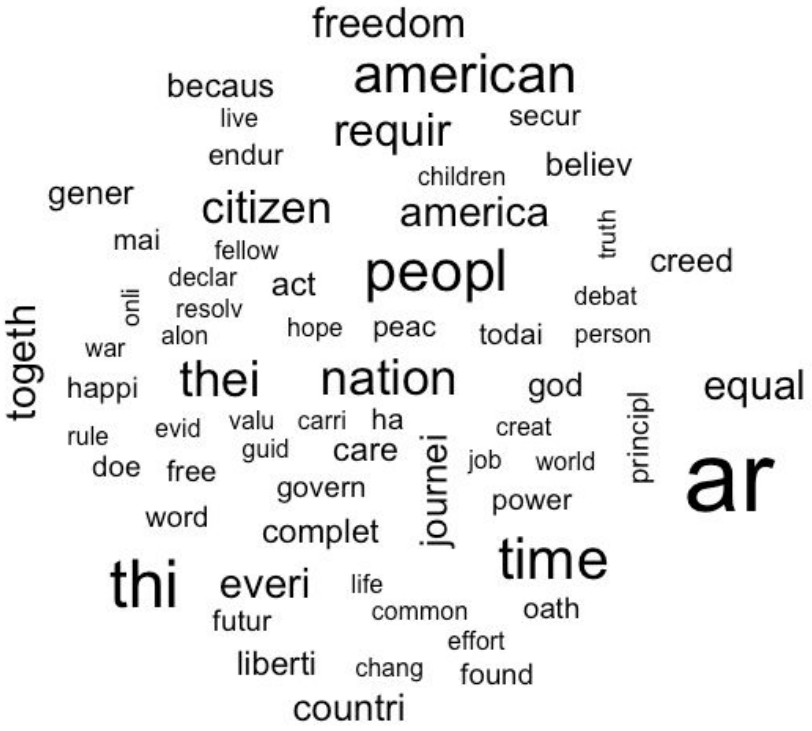
Which one is which?

Trump




```
> sum(total$Trump.Inaug.2017.x, na.rm=T)
[1] 537
```

Obama




```
> sum(total$Obama.Inaug.2009.x, na.rm=T)
[1] 824
```

How did I make these wordclouds?

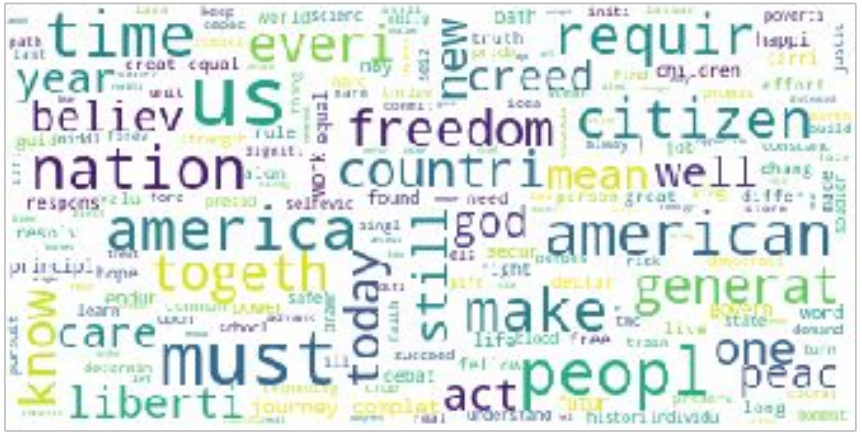
```
library(wordcloud)

wordcloud(df.t.t$names, df.t.t$Obama.Inaug.2009, min.freq=3, random.color=T,
ordered.colors=T)

wordcloud(df.t.t$names, df.t.t$Trump.Inaug.2017, min.freq=3, random.color=T,
ordered.colors=T)
```

This in Python…

Trump


```
Trump = ' '.join(corpus[1])
wordcloud = WordCloud(background_color="white",
max_font_size=40).generate(Trump)
plt.imshow(wordcloud)
plt.axis("off")
```

Obama




```
Obama = ' '.join(corpus[0])
wordcloud = WordCloud(background_color="white",
max_font_size=40).generate(Obama)
plt.imshow(wordcloud)
plt.axis("off")
```

What are the most distinctive words for each speaker?

We can do what Neal Caren had done last week

```
df.t.t = as.data.frame(t.t)

summing = function(x) x/sum(x, na.rm=T)

df.t.t.2 = apply(df.t.t, 2, summing)

df.t.t$names<-rownames(df.t.t)
df.t.t = as.data.frame(t.t)
df.t.t$names<-rownames(df.t.t)
head(df.t.t)

df.t.t.2 = as.data.frame(df.t.t.2)
df.t.t.2$names<-rownames(df.t.t.2)
df.t.t.2 = as.data.frame(df.t.t.2)

total <- merge(df.t.t,df.t.t.2,by="names")

total$obama.over.trump = (total$Obama.Inaug.2009.y) -
(total$Trump.Inaug.2017.y)
sort.OT <- total[order(total$obama.over.trump) , ]
sort.OT[1:30, ]
```

What are the most Obama-esque words, relative to Trump's?

Requirements; time; equality; journey; freedom; act; care

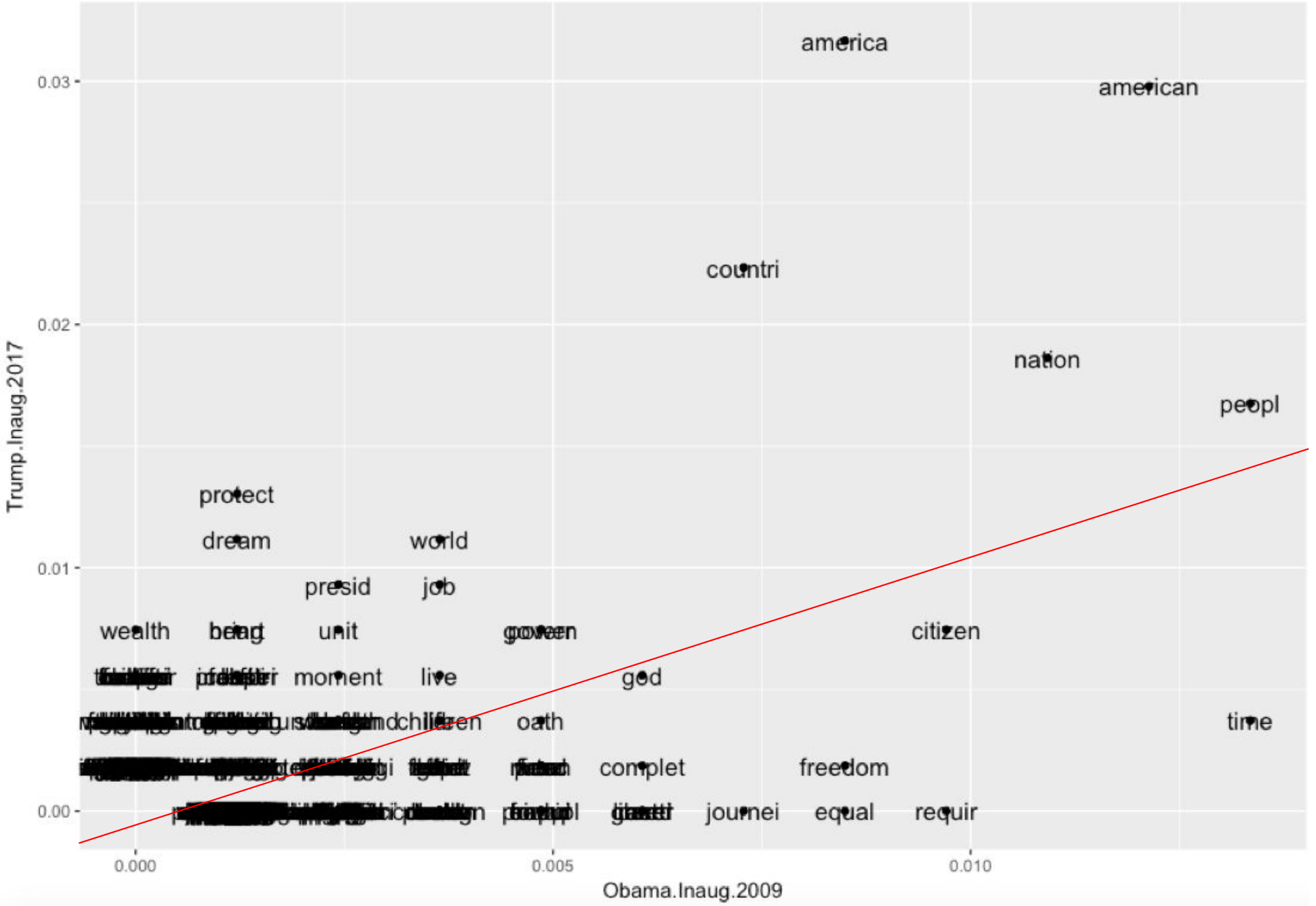
```
sort.OT[1:30, ]
  names Obama.Inaug.2009.x Trump.Inaug.2017.x Obama.Inaug.2009.y Trump.Inaug.2017.y obama.over.trump
 requir                  8                   0        0.009708738        0.000000000      0.009708738
   time                 11                   2        0.013349515        0.003724395      0.009625120
  equal                  7                   0        0.008495146        0.000000000      0.008495146
journei                  6                   0        0.007281553        0.000000000      0.007281553
freedom                  7                   1        0.008495146        0.001862197      0.006632948
    act                  5                   0        0.006067961        0.000000000      0.006067961
   care                  5                   0        0.006067961        0.000000000      0.006067961
  creed                  5                   0        0.006067961        0.000000000      0.006067961
  gener                  5                   0        0.006067961        0.000000000      0.006067961
liberti                  5                   0        0.006067961        0.000000000      0.006067961
  endur                  4                   0        0.004854369        0.000000000      0.004854369
  found                  4                   0        0.004854369        0.000000000      0.004854369
  happi                  4                   0        0.004854369        0.000000000      0.004854369
principl                 4                   0        0.004854369        0.000000000      0.004854369
  secur                  4                   0        0.004854369        0.000000000      0.004854369
complet                  5                   1        0.006067961        0.001862197      0.004205764
  chang                  3                   0        0.003640777        0.000000000      0.003640777
 common                  3                   0        0.003640777        0.000000000      0.003640777
  creat                  3                   0        0.003640777        0.000000000      0.003640777
 declar                  3                   0        0.003640777        0.000000000      0.003640777
   evid                  3                   0        0.003640777        0.000000000      0.003640777
   know                  3                   0        0.003640777        0.000000000      0.003640777
 person                  3                   0        0.003640777        0.000000000      0.003640777
```

What are the most Trumpian words, relative to Obama?

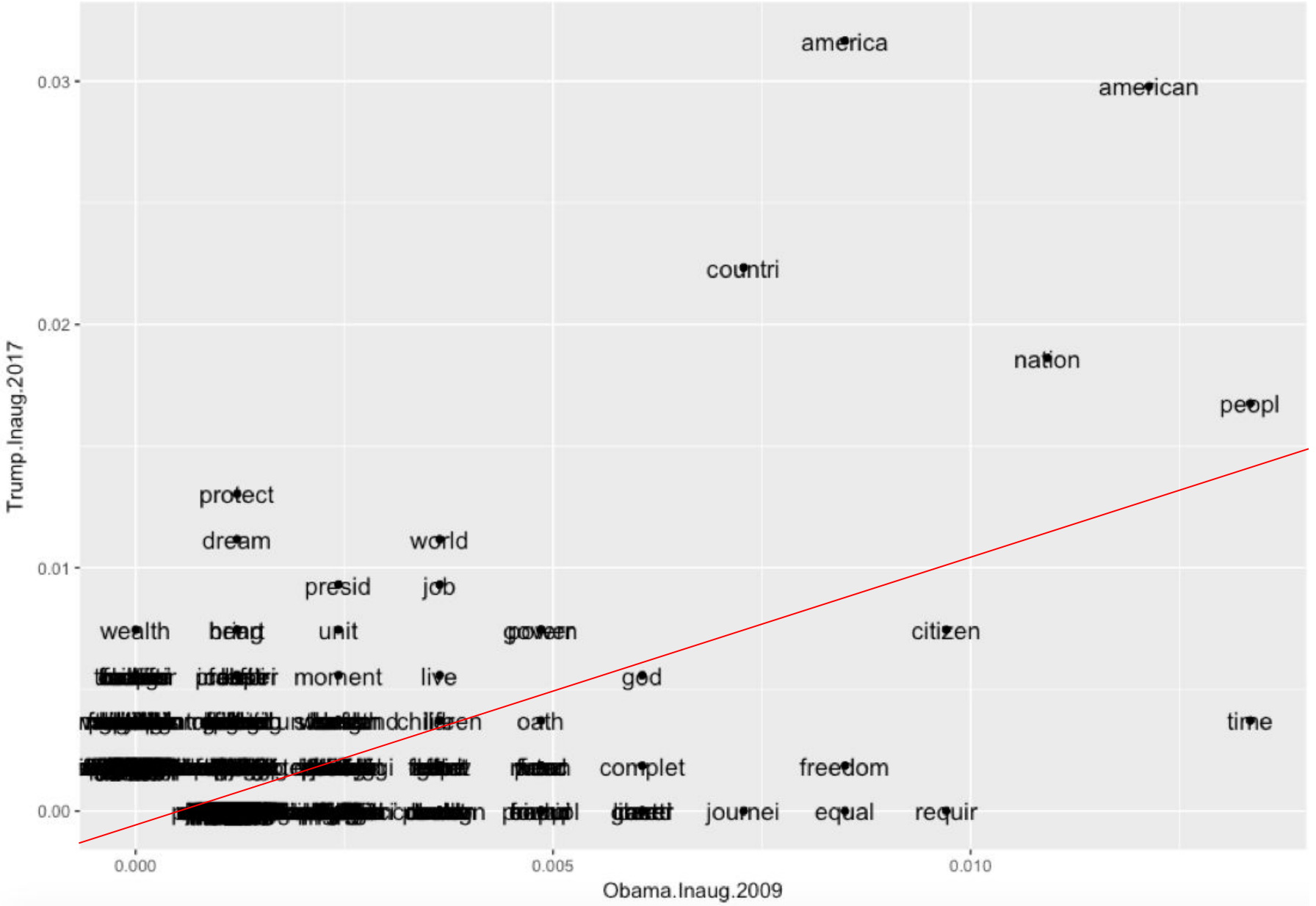
America; Americans; country; protection; dream; nation

```
 sort.OT[1:30, ]
    names Obama.Inaug.2009.x Trump.Inaug.2017.x Obama.Inaug.2009.y Trump.Inaug.2017.y obama.over.trum
 america                   7                  17        0.008495146        0.031657356     -0.0231622
american                  10                  16        0.012135922        0.029795158     -0.0176592
 countri                   6                  12        0.007281553        0.022346369     -0.0150648
 protect                   1                   7        0.001213592        0.013035382     -0.0118217
   dream                   1                   6        0.001213592        0.011173184     -0.0099595
  nation                   9                  10        0.010922330        0.018621974     -0.0076996
   world                   3                   6        0.003640777        0.011173184     -0.0075324
  wealth                   0                   4        0.000000000        0.007448790     -0.0074487
  presid                   2                   5        0.002427184        0.009310987     -0.0068838
   bring                   1                   4        0.001213592        0.007448790     -0.0062351
   heart                   1                   4        0.001213592        0.007448790     -0.0062351
     job                   3                   5        0.003640777        0.009310987     -0.0056702
  border                   0                   3        0.000000000        0.005586592     -0.0055865
   capit                   0                   3        0.000000000        0.005586592     -0.0055865
    citi                   0                   3        0.000000000        0.005586592     -0.0055865
     dai                   0                   3        0.000000000        0.005586592     -0.0055865
 factori                   0                   3        0.000000000        0.005586592     -0.0055865
 foreign                   0                   3        0.000000000        0.005586592     -0.0055865
 million                   0                   3        0.000000000        0.005586592     -0.0055865
 nation’                   0                   3        0.000000000        0.005586592     -0.0055865
   obama                   0                   3        0.000000000        0.005586592     -0.0055865
transfer                   0                   3        0.000000000        0.005586592     -0.0055865
    unit                   2                   4        0.002427184        0.007448790     -0.0050216
```

Viewing the outliers another way



What else does this make you think of?



How did I make this graph?

```
df.t.t = as.data.frame(t.t)

summing = function(x) x/sum(x, na.rm=T)

df.t.t.2 = apply(df.t.t, 2, summing)

df.t.t$names<-rownames(df.t.t)
df.t.t = as.data.frame(t.t)
df.t.t$names<-rownames(df.t.t)
head(df.t.t)

df.t.t.2 = as.data.frame(df.t.t.2)
df.t.t.2$names<-rownames(df.t.t.2)
df.t.t.2 = as.data.frame(df.t.t.2)


q = qplot(Obama.Inaug.2009, Trump.Inaug.2017, data = df.t.t.2)
q + geom_text(aes(label=names), size = 4.5)
```

How do we determine how similar two texts are to each other?

One idea

- What I showed you last week from Sacerdote et al using the methodology, e.g., of MLK (0) to RR (1)

Another idea

- Classic measures of similarity
- The most common measure is cosine similarity. We are comparing two non-binary vectors, where each vector is a document and each attribute represents the frequency with which a particular word occurs in the document. Each document vector is sparse since it has relatively few non-zero attributes. Therefore, the cosine similarity ignores 0-0 matches like the Jaccard measure.

Another idea

- The cosine similarity is defined by the following equation:

$$cos(A, B) = \frac{A \times B}{\|A\|\|B\|}$$

where A is the first work's score vector and B is the second work's score vector. "||A||" represents the Euclidean length of the vector A, which is the square root of the dot product of the vector with itself

Another measure

- Cosine similarity is a close relative of the Pearson product moment correlation for two data vectors.

- The Pearson correlation coefficient represents the angular separation between two *normalized* data vectors measured from the *mean*, while the cosine similarity measures the angular separation of two data vectors measured from *zero*.

- Cosine similarity tends to give higher values than Pearson correlation

Lots more measures possible

- Binary cosine similarity coefficient
- Binary Dice coefficient
- Binary Jaccard coefficient
- Binary overlap coefficient

Check out these sources (where I borrowed liberally from)

Mining Similarity Using Euclidean Distance, Pearson Correlation,and Filtering

Similarity Measurement

Similarity metric is the basic measurement and used by a number of data ming algorithms. It measures the similarity or dissimilarity between two data objects which have one or multiple attributes. Informally, the similarity is a numerical measure of the degree to which the two objects are alike. It is usually non-negative and are often between 0 and 1, where 0 means no similarity, and 1 means complete similarity. [1]

WORDHOARD

Comparing Collocates | Table of Contents | Annotations

Comparing texts

You've seen how WordHoard allows you to compare individual word forms and collocates in texts. WordHoard can also compare two texts by computing a single overall measure of *document similarity*. The more similar the vocabulary for two texts, the higher the value of the similarity

An example

- Which one of Shakespeare's other tragedies is most like Hamlet?
- Least like Hamlet?

The answer:

Work Parts compared	Cosine	Binary cosine	Binary Dice	Binary Jaccard	Binary Overlap
Hamlet with Othello	0.9643	0.5217	0.5185	0.3500	0.5825
Hamlet with Antony and Cleopatra	0.9710	0.5068	0.5039	0.3368	0.5635
Hamlet with King Lear	0.9748	0.5057	0.5044	0.3372	0.5444
Hamlet with Timon of Athens	0.9634	0.5043	0.4946	0.3285	0.6151
Hamlet with Macbeth	0.9771	0.5003	0.4925	0.3267	0.5974
Hamlet with Romeo and Juliet	0.9520	0.4971	0.4933	0.3274	0.5619
Hamlet with Coriolanus	0.9669	0.4929	0.4904	0.3248	0.5460
Hamlet with Julius Caesar	0.9563	0.4903	0.4747	0.3113	0.6326
Hamlet with Titus Andronicus	0.9536	0.4660	0.4597	0.2985	0.5491

- Using the binary cosine measure, the play most like Hamlet in terms of its word usage is Othello.
- “This isn't too surprising given that both plays feature revenge themes and more interior dialog then usual.”

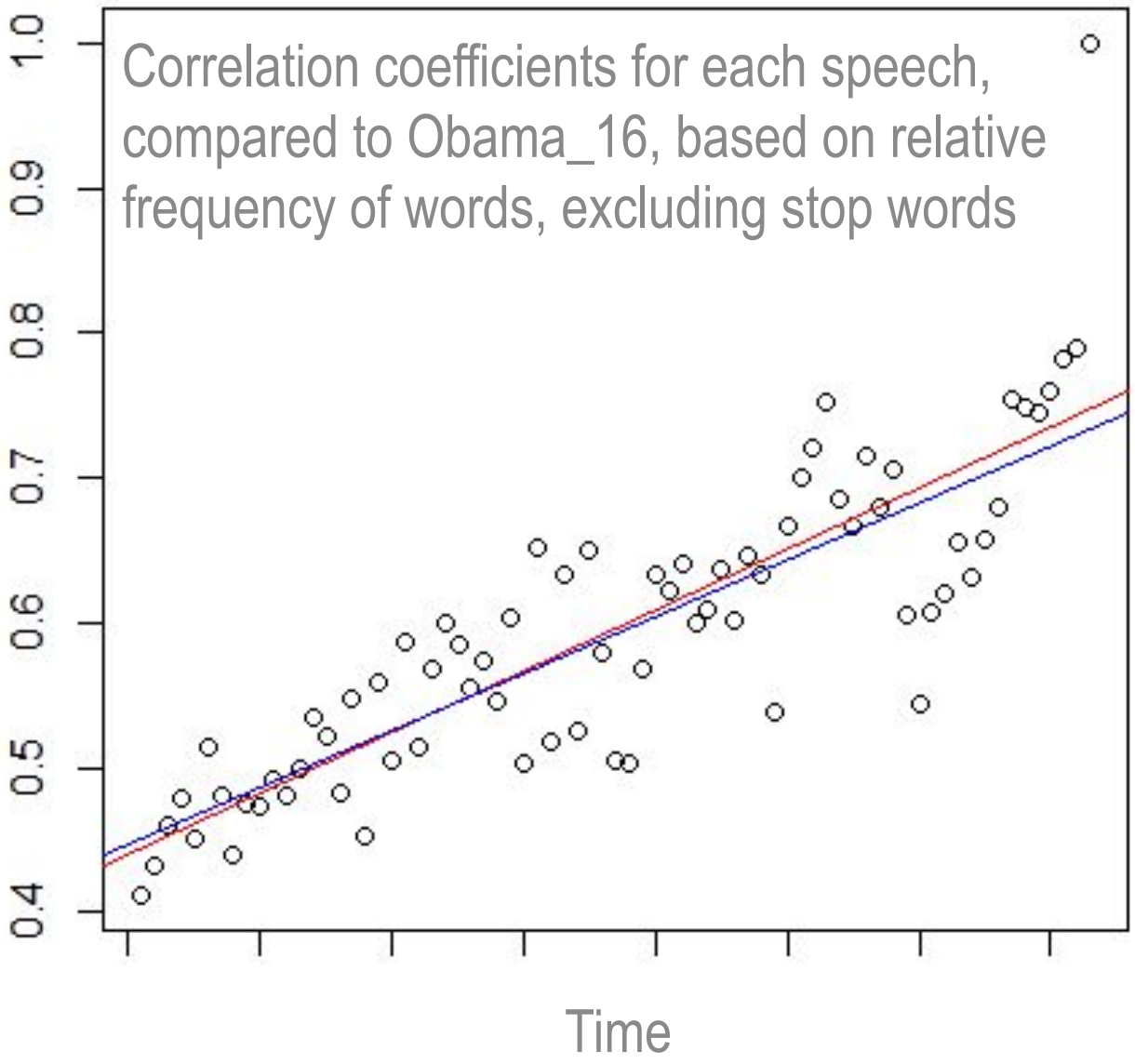
The question

- Can we figure out from word choice which President delivered which States of the Union?

Answer

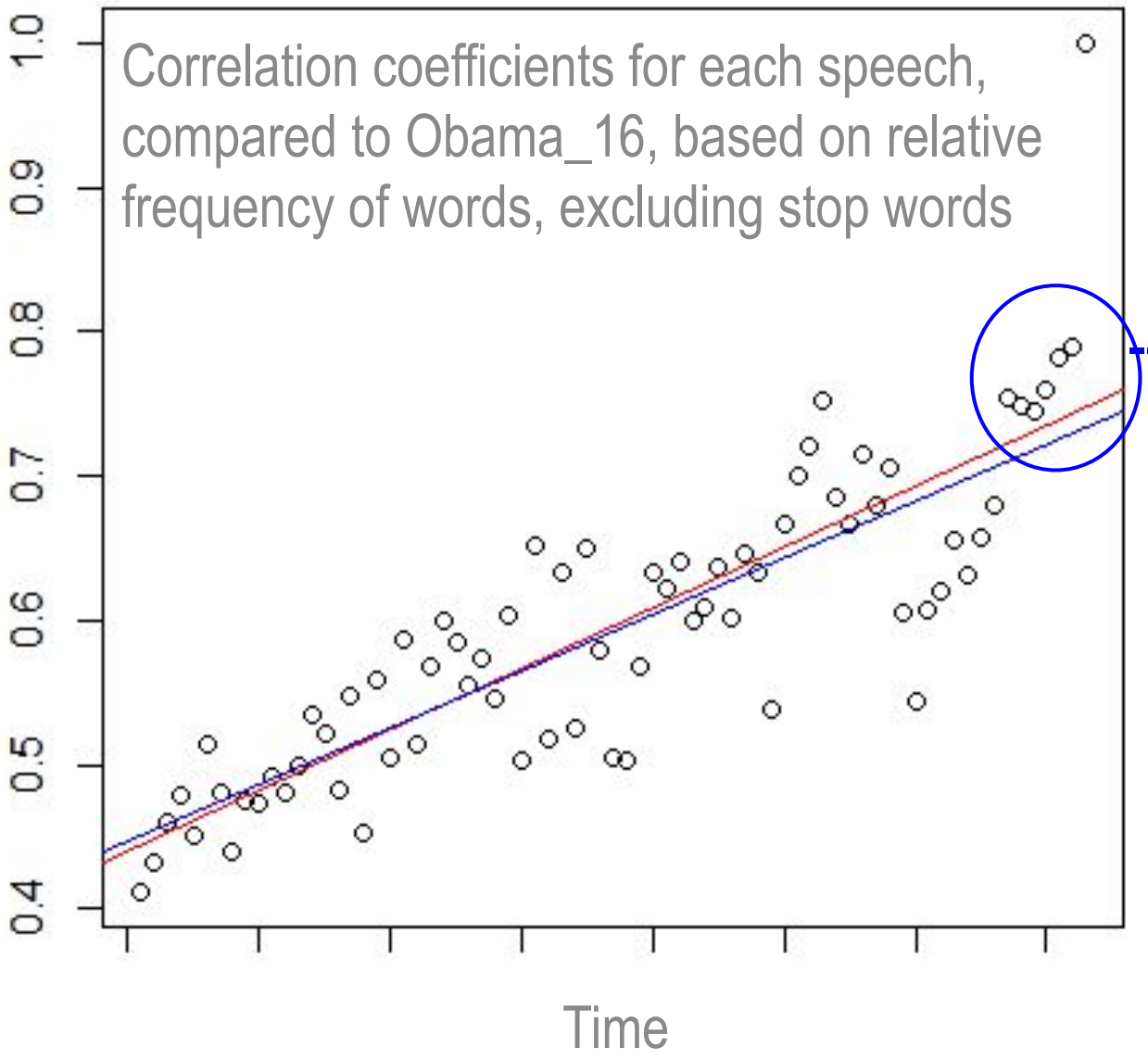
- Let’s use the correlation coefficient to compare

Obama's 2016 State of the Union



Who does Obama in 2016 most speak like?

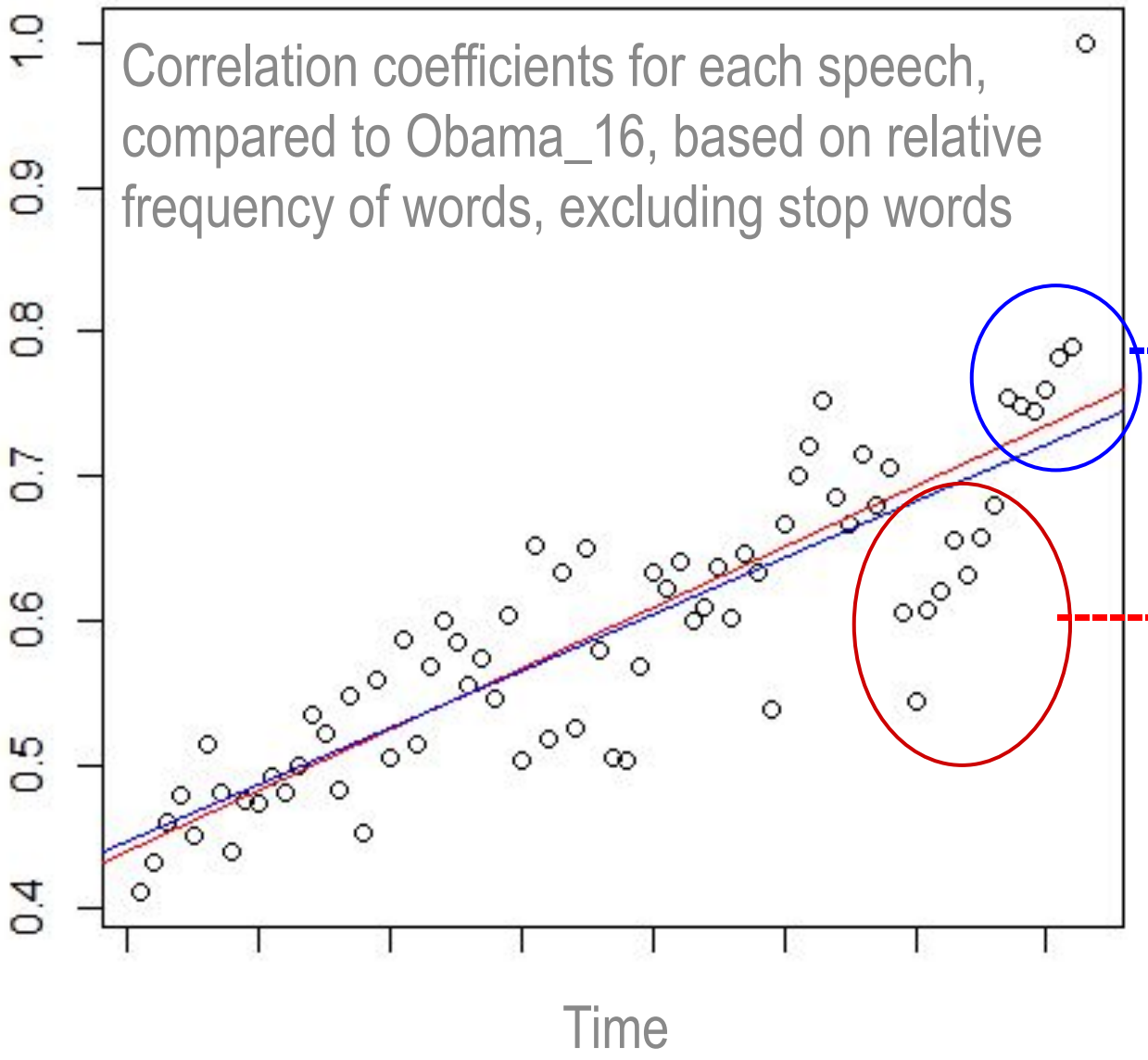
Obama's 2016 State of the Union



Who does Obama in 2016 most speak like?

----------Obama

Obama's 2016 State of the Union

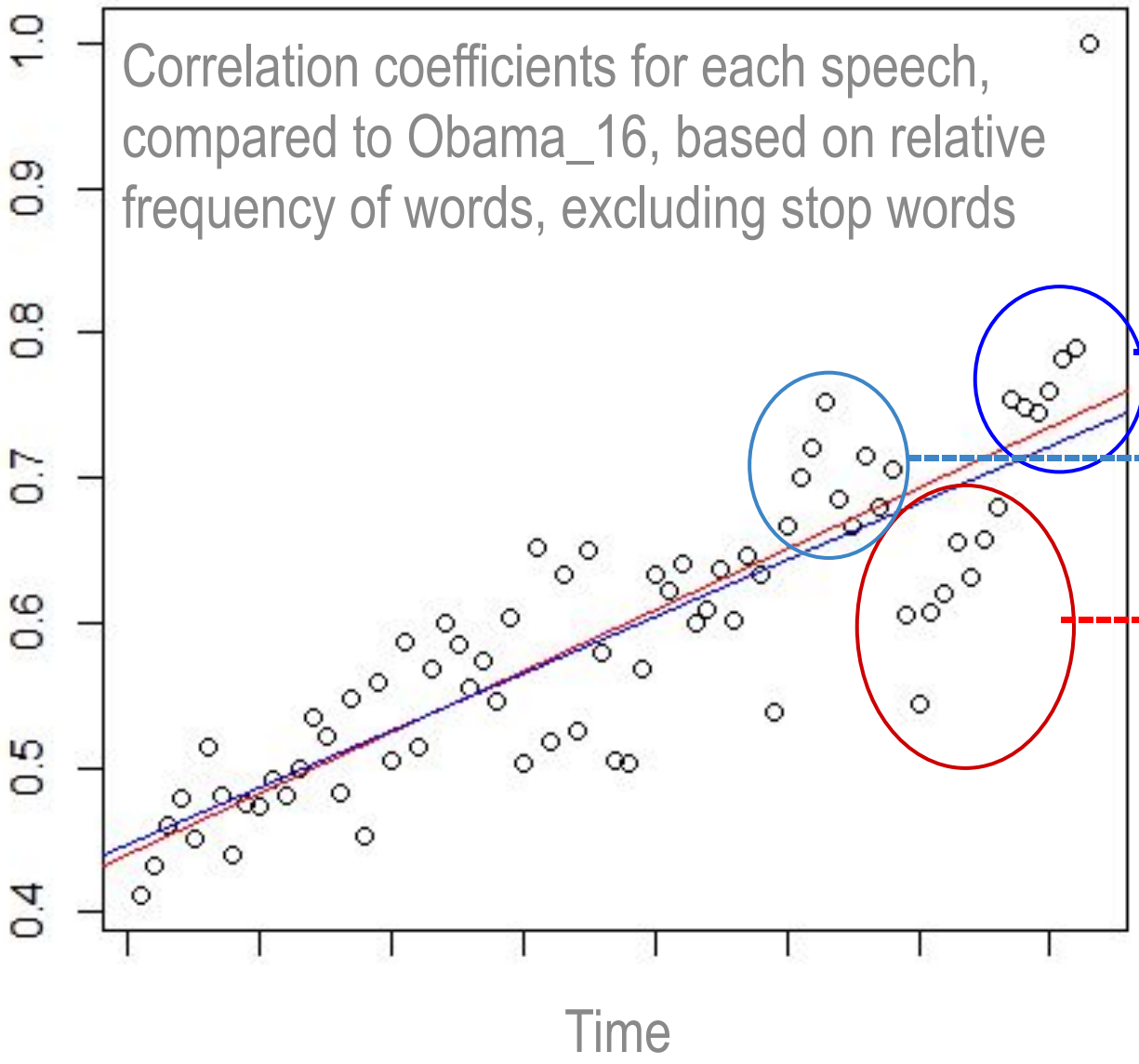


Who does Obama in 2016 most speak like?

----------Obama

----------W. Bush

Obama's 2016 State of the Union



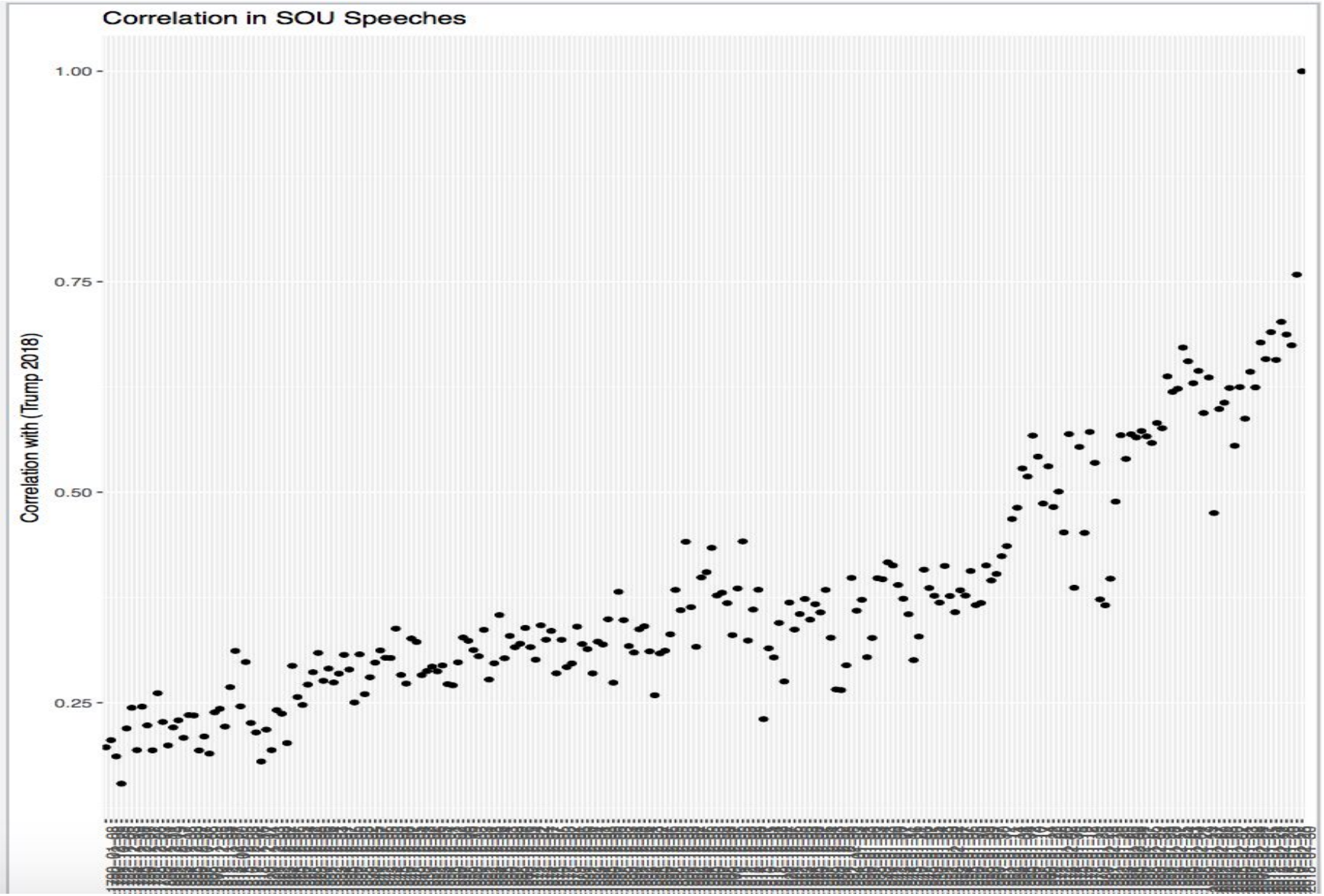
Who does Obama in 2016 most speak like?

Obama

Clinton

W. Bush

BTW: Trump's 2017 State of the Union



Back to our example ...

After stop-words and with stemmed words, the Obama and Trump speeches have a cosine similarity of 0.60

```
> library(lsa)
> cosine(t.t)
                  Obama.Inaug.2009 Trump.Inaug.2017
Obama.Inaug.2009         1.0000000         0.6019748
Trump.Inaug.2017         0.6019748         1.0000000
```

Back to our example ...

After stop-words and with stemmed words, the Obama and Trump speeches have a Pearson correlation coefficient of 0.46

```
> cor(t.t, method="pearson")
                  Obama.Inaug.2009 Trump.Inaug.2017
Obama.Inaug.2009         1.0000000         0.4603077
Trump.Inaug.2017         0.4603077         1.0000000
```

What about a chi-squared test?

- These guys use it as one stage of their work

Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics

Zubin Jelveh[1], Bruce Kogut[2], and Suresh Naidu[3]

[1]Dept. of Computer Science & Engineering, New York University
[2]Columbia Business School and Dept. of Sociology, Columbia University
[3]Dept. of Economics and SIPA, Columbia University
zj292@nyu.edu, bruce.kogut@columbia.edu, sn2430@columbia.edu

Abstract

Previous work on extracting ideology from text has focused on domains where expression of political views is expected, but it's unclear if current technology can work in domains where displays of ideology are considered inappropriate. We

views itself as a science (Chetty, 2013) carefully applying rigorous methodologies and using institutionalized safe-guards such as peer review. The field's most prominent research organization explicitly prohibits researchers from making policy recommendations in papers that it releases (National Bureau of Economic Research, 2010). Despite these measures, economics' close proximity

Advocates for this chi-squared approach

Expert Systems with Applications

Volume 38, Issue 4, April 2011, Pages 3085–3090

Using chi-square statistics to measure similarities for text categorization

Yao-Tsung Chen[a], Meng Chang Chen[b, 1]

Show more

Choose an option to locate/access this article:

Check if you have access through your login credentials or your institution

Check access

Purchase $39.95

Get Full Text Elsewhere

doi:10.1016/j.eswa.2010.08.100

Get rights and content

Abstract

In this paper, we propose using chi-square statistics to measure similarities and chi-

Variants on the chi-squared approach

Word frequency and key word statistics in historical corpus linguistics

Alistair Baron, Lancaster University
Paul Rayson, Lancaster University
Dawn Archer, University of Central Lancashire

1. Introduction

Frequency-sorted word lists have long been part of the standard methodology for exploiting corpora. Sinclair (1991: 30) noted that "anyone studying a text is likely to need to know how often each different word form occurs in it". Tribble and Jones (1997: 36) outlined a methodology for using texts in the language classroom, proposing that the most effective starting point for understanding a text is a frequency-sorted word list. A frequency list records the number of times that each word occurs in the text. It can therefore provide

Back to our example ...

After stop-words and with stemmed words, a Chi-squared test suggests that the Obama and Trump speeches are not independent

```
> ctable <- table(t.t)
> chisq.test(ctable)

    Chi-squared test for given probabilities

data:  ctable
X-squared = 8785.1, df = 17, p-value < 2.2e-16
```

How about a full regression model?

Let's model this both as a linear relationship: both on the raw frequencies and the relative frequencies

```
m1a = lm(Trump.Inaug.2017 ~ Obama.Inaug.2009, df.t.t)
m2a = lm(Trump.Inaug.2017.y ~ Obama.Inaug.2009.y , total)

##install.packages("stargazer")
library(stargazer)
stargazer(m1a, m2a, type = "text")
```

How about a full regression model?

Knowing which words Obama used frequently (and which ones he didn't) helps us predict Trump's words with up to 20% more accuracy (in Model 1)

```
=> stargazer(m1a, m1b, type = "text", single.row = TRUE)

================================================================
                                Dependent variable:
                    ---------------------------------------------
                                  Trump.Inaug.2017
                                (1)
----------------------------------------------------------------
Obama.Inaug.2009          0.442*** (0.030)

Constant                  0.264*** (0.059)
----------------------------------------------------------------
Observations                  796
R2                           0.212
Adjusted R2                  0.211
Residual Std. Error      1.337 (df = 794)
F Statistic          213.465*** (df = 1; 794)
================================================================
```

How about a full regression model?

The results are equivalent (in the R-sq) between raw frequencies and relative ones

	Dependent variable:			
	Trump.Inaug.2017		Trump.Inaug.2017.y	
	(1)	(2)	(3)	(4)
Obama.Inaug.2009	0.442*** (0.030)	-0.001 (0.051)		
I(Obama.Inaug.2009^2)		0.043*** (0.004)		
Obama.Inaug.2009.y			0.660*** (0.045)	-0.001 (0.076)
I(Obama.Inaug.2009.y^2)				59.375*** (5.708)
Constant	0.264*** (0.059)	0.615*** (0.065)	0.0004*** (0.0001)	0.001*** (0.0001)
Observations	796	796	796	796
R2	0.212	0.307	0.212	0.307
Adjusted R2	0.211	0.305	0.211	0.305
Residual Std. Error	1.337 (df = 794)	1.255 (df = 793)	0.002 (df = 794)	0.002 (df = 793)
F Statistic	213.465*** (df = 1; 794)	175.248*** (df = 2; 793)	213.465*** (df = 1; 794)	175.248*** (df = 2;
Note:				*p<0.1; **p<0.05; ***p<

One more way to think of this:

Term-frequency inverse document-frequency. Weight the matrix by term frequency, inverse document frequency, which seeks to isolate words that are most associated with one document or another

```
d= df.mpwdtm
idf <- log(nrow(d)/colSums(d))
tfidf <- d

for(word in names(idf)){
  tfidf[,word] <- df.mpwdtm[,word] * idf[word]
}

m.tfidf=as.matrix(tfidf)
df.m.tfidf=as.data.frame(m.tfidf)

## make the dtm into a tdm instead ##

t.df.m.tfidf = t(df.m.tfidf)
t.df.m.tfidf=as.data.frame(t.df.m.tfidf)
```

One more way to think of this:

Here are the inverse weightings:

```
> head(t.df.m.tfidf, 25)
         Obama.Inaug.2009 Trump.Inaug.2017
1776            0.6931472        0.0000000
40              0.6931472        0.0000000
400             0.6931472        0.0000000
abov            0.6931472        0.0000000
abroad          0.6931472        0.0000000
absolut         0.6931472        0.0000000
act            -4.5814537        0.0000000
action         -0.4054651       -0.8109302
advanc          0.0000000        0.0000000
affirm          0.6931472        0.0000000
afford          0.6931472        0.0000000
africa          0.6931472        0.0000000
ago             0.6931472        0.0000000
agre            0.6931472        0.0000000
allegi         -0.4054651       -0.8109302
allianc         0.0000000        0.0000000
alon           -1.2163953        0.0000000
alwai          -1.3862944       -1.3862944
america       -17.3943465      -42.2434130
america’        0.0000000        0.0000000
american      -25.6494936      -41.0391897
```

One more way to think of this:

Based on the inverse weightings, the Pearson correlation is .83

```
> cor(t.df.m.tfidf, method="pearson")
                  Obama.Inaug.2009 Trump.Inaug.2017
Obama.Inaug.2009         1.000000         0.828876
Trump.Inaug.2017         0.828876         1.000000
```

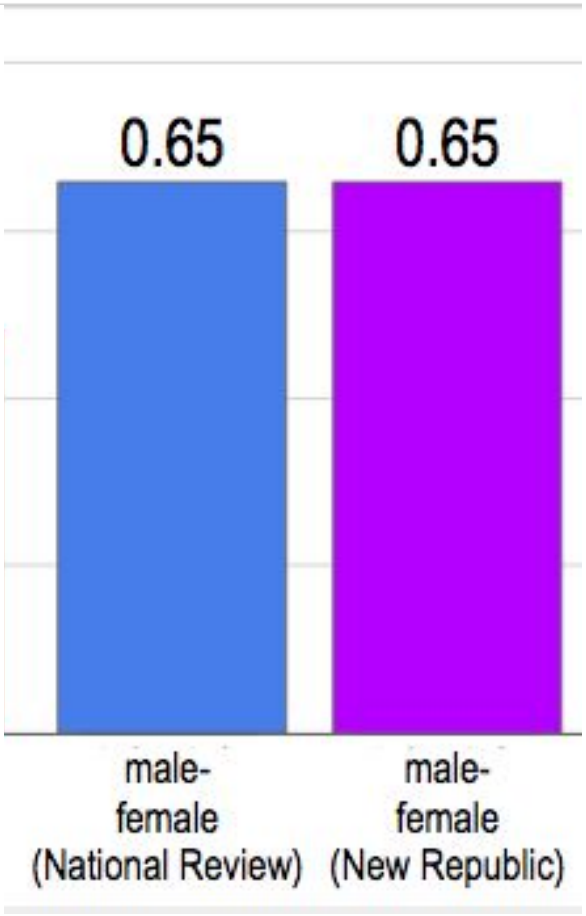
Doubling-back on the example from last week

Cosine similarity for different “bags of words”

The *National Review* writes sentences about male and female individuals in moderately similar ways.

This ***might*** support our hypothesis that male and female sentences are <u>not</u> the same in the conservative magazine (since cosine < 1).

Our hypothesis is not supported

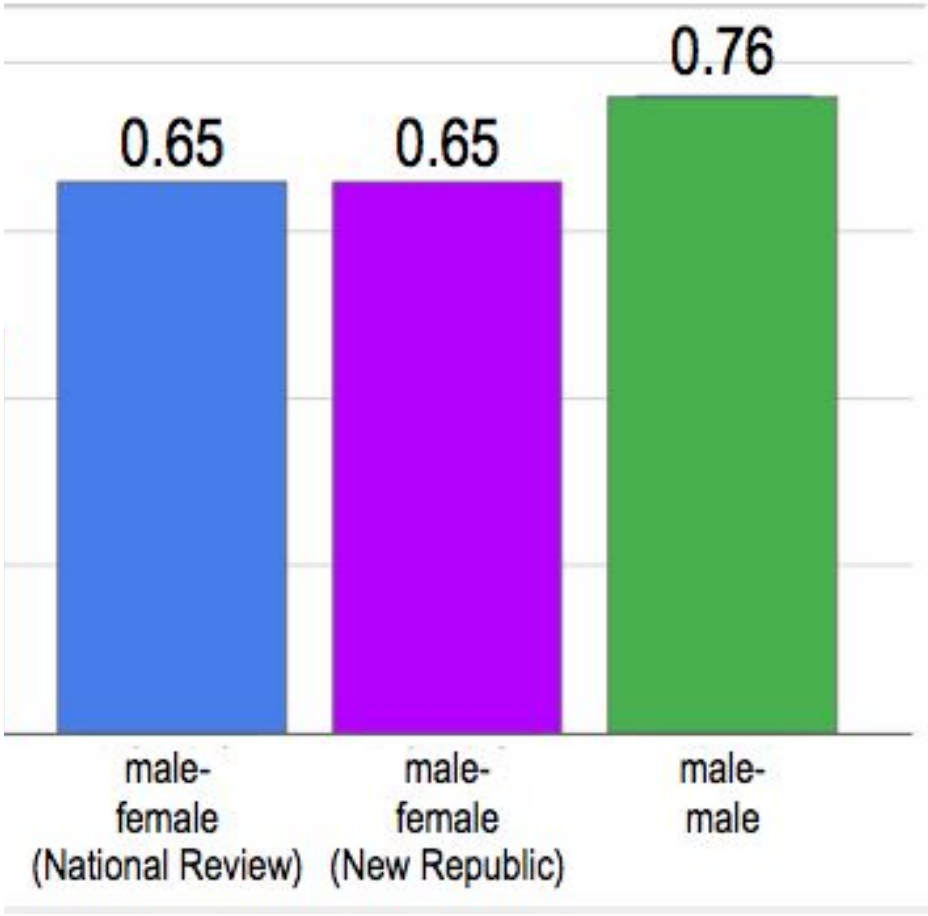


The *New Republic* does no better!

The *New Republic* does not write about males and females in an interchangeable (=egalitarian) way.

Both magazines differentiate between males and females to the exact same degree.

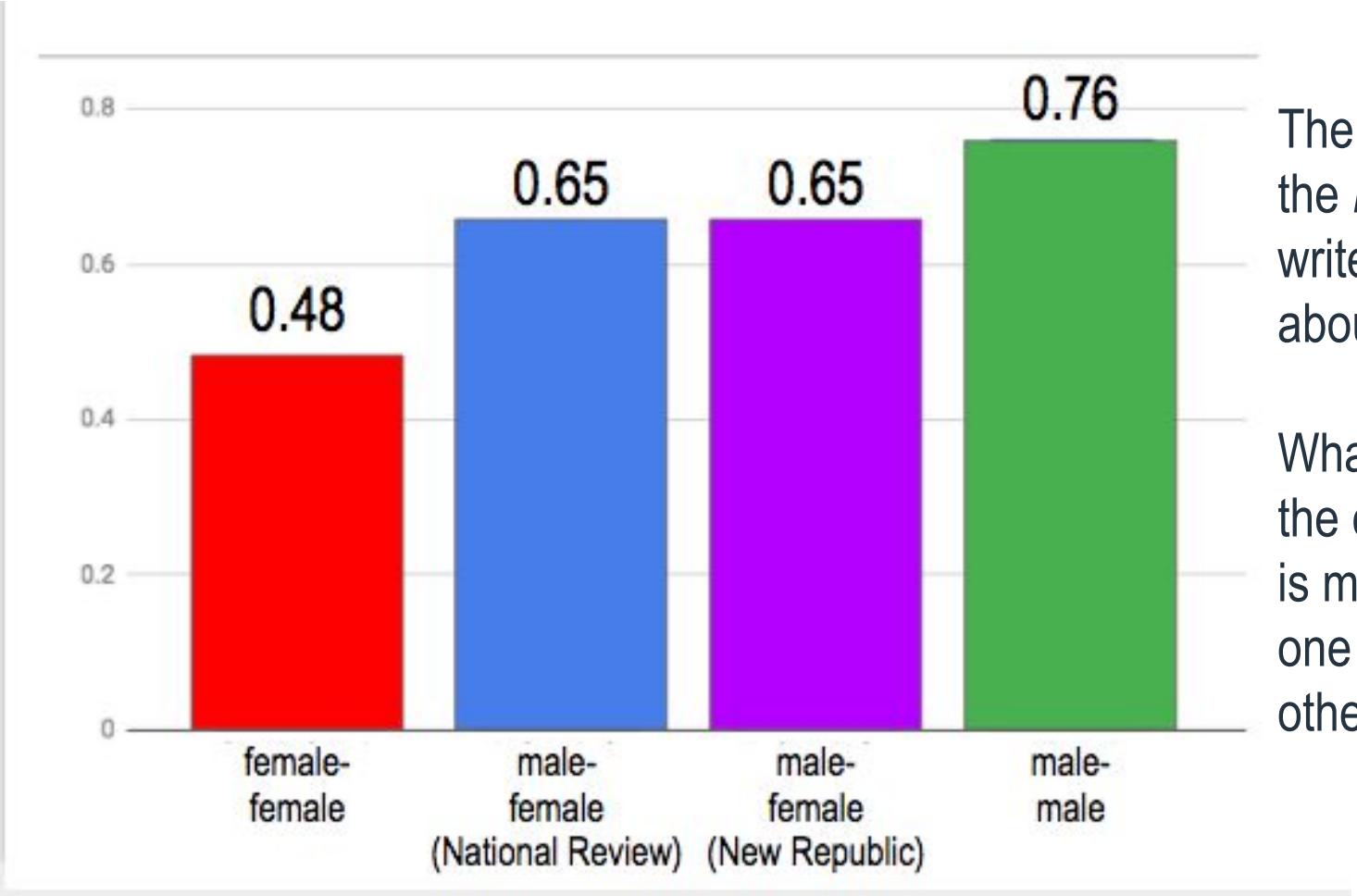
There is a narrower “script” around men



Interestingly, the *New Republic* writes about males in a way that is ***more*** similar to how the *National Review* writes about males, and vice versa.

Potentially, gender overrides editorial (and/or ideological) voice. This is quite remarkable.

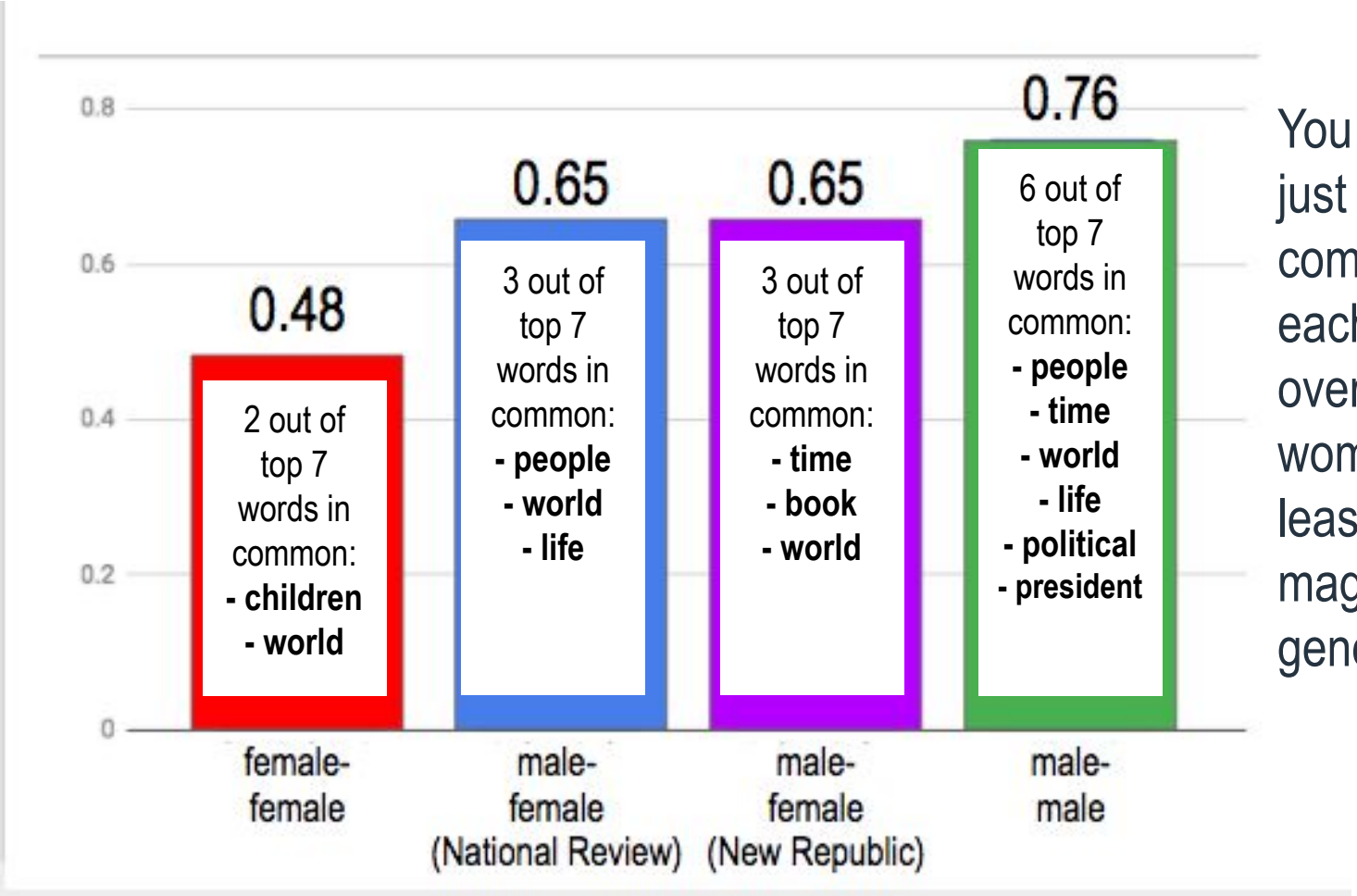
There is a looser “script” around females



The *New Republic* and the *National Review* write pretty differently about females.

What is talked about in the context of females is more variable from one publication to the other.

How could this be?



You can see that if we just look at the 7 most common words for each group, men overlap the most, and women overlap the least, with each magazine (across genders) in the middle.

Another example, exclusively in **tm**

Comparing these 3 texts

- The “About Us” + “FAQ” pages for QMSS, A3SR, and AQR

QMSS

- Quantitative Methods in the Social Sciences (QMSS) is an innovative, flexible, interdisciplinary social science Master of Arts degree program at Columbia University that focuses on quantitative research techniques and strategies. The program integrates the perspectives and research methods of six social science disciplines: Economics, History, Political Science, Psychology, Sociology, and Statistics.
- QMSS provides students with rigorous training in quantitative research, with an emphasis on written and oral communication about research techniques and findings. These skills prepare QMSS graduates to enter (or further) an analytical or research career or to continue their education in a PhD program.
- Etc.

A3SR - Applied Statistics MS at NYU

- **Learn advanced statistical methods and use them to address important social issues.** In this master of science program, you will build a strong foundation in statistical research techniques and apply them in contemporary social, behavioral, and health science research.
- According to the National Center for Education Statistics, **the number of jobs in statistics and data science has nearly tripled in the past five years** – and the American Statistical Association says there still aren't enough statistics graduates to meet this demand. As a graduate of our program, you'll acquire highly marketable skills for a wide range of challenging and rewarding data-intensive careers in the social sciences.
- Etc.

AQR - Applied Quantitative Research

- The Master's Degree program in Applied Quantitative Research (AQR) is designed to be innovative, flexible, and interdisciplinary—training students in cutting-edge quantitative techniques and strategies as applied in contemporary social science research. The program brings together a group of students with quantitatively-focused interests and career ambitions to provide them with rigorous training in quantitative techniques and applied statistical analysis. The structure of the program is designed to improve students' understanding of quantitative research and prepare them for a number of possible careers in the private or public sector, as well as for further academic study.
- Etc.

Reading in these texts

- I confess: I just copied and pasted these 3 texts into individual text files and saved them individually as "A3SR.txt", "AQR.txt", and "QMSS.txt"

- You could certainly use RCurl and other packages to get them into R directly (more on that later today and next week)

Getting started

Set up the tm package and related packages and read in the .txt files

```
Needed <- c("tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust",
"cluster", "igraph", "fpc")
install.packages(Needed, dependencies=TRUE)

install.packages("Rcampdf", repos = "http://datacube.wu.ac.at/", type = "source")

library(tm)
library(SnowballC)
library(Rcampdf)

cname <- file.path("~", "Desktop", "texts") ## for Macs

##cname <- file.path("/Users/gregoryeirich/Downloads", "texts") ## for PCs do this ##

> cname
[1] "~/Desktop/texts"

> dir(cname)
[1] "A3SR.txt" "AQR.txt"  "QMSS.txt"
```

Making a corpus

Turn your text files into something tm can use:

```
> docs <- Corpus(DirSource(cname)) ## you may get a Warning message, don’t worry ##

> summary(docs)
A corpus with 3 text documents

> inspect(docs[1])
A corpus with 1 text document

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create date creator
Available variables in the data frame are:
  MetaID

$A3SR.txt
Learn advanced statistical methods and use them to address important social issues. In
this master of science program, you will build a strong foundation in statistical
research techniques and apply them in contemporary social, behavioral, and health
science research.
According to the National Center for Education Statistics, the number of jobs in
statistics and data science has nearly tripled in the past five years - and the
American Statistical Association says there still aren't enough statistics graduates
to meet this demand. As a graduate of our program, you'll acquire highly marketable
skills for a wide range of challenging and rewarding data-intensive careers in the
social sciences.
```

Cleaning up the corpus

Perform a handful of standard procedures

```
docs <- tm_map(docs, removePunctuation)

for(j in seq(docs))
{
  docs[[j]] <- gsub("/", " ", docs[[j]])
  docs[[j]] <- gsub("@", " ", docs[[j]])
  docs[[j]] <- gsub("\\|", " ", docs[[j]])
}

docs <- tm_map(docs, removeNumbers)

docs <- tm_map(docs, tolower)

# For a list of the stopwords, see:
# length(stopwords("english"))
# stopwords("english")
docs <- tm_map(docs, removeWords, stopwords("english"))

#docs <- tm map(docs, removeWords, c("department", "email"))
# Just replace "department" and "email" with words that you would like to remove.
```

Cleaning up the corpus, cont'd

Perform a handful of standard procedures

```
##for (j in seq(docs))
##{
  ##docs[[j]] <- gsub("qualitative research", "QDA", docs[[j]])
  ##docs[[j]] <- gsub("qualitative studies", "QDA", docs[[j]])
  ##docs[[j]] <- gsub("qualitative analysis", "QDA", docs[[j]])
  ##docs[[j]] <- gsub("research methods", "research_methods", docs[[j]])
##}

library(SnowballC)
docs <- tm_map(docs, stemDocument)

docs <- tm_map(docs, stripWhitespace)

docs <- tm_map(docs, PlainTextDocument)

> inspect(docs[1])
A corpus with 1 text document

[skip]

$A3SR.txt
learn advanc statist method use address import social issu master scienc program will
build strong foundat statist research techniqu appli contemporari social behavior
health scienc research
accord nation center educ statist number job statist data scienc near tripl past five
year american statist associ say still arent enough statist graduat meet demand
graduat program youll acquir high market skill wide rang challeng reward dataintens
career social scienc
```

Now, the document-term matrix

Here it is:

```
> docs <- Corpus(VectorSource(docs))
> dtm <- DocumentTermMatrix(docs)

> dtm
A document-term matrix (3 documents, 796 terms)

Non-/sparse entries: 1242/1146
Sparsity           : 48%
Maximal term length: 45
Weighting          : term frequency (tf)
```

What's the document-term matrix?

We going to get this into a dataframe format now too

```
> m.dtm = as.matrix(dtm)

> df.dtm = as.data.frame(m.dtm)

> head(df.dtm)
         abil abl abroad academ acceler accept access accord accredit achiev acquir addit addre
A3SR.txt    2   2      0      4       2      1      1      1         1      2      1     5
AQR.txt     0   0      1      4       0      1      0      0         0      0      1     3
QMSS.txt    2   5      0      9       0      4      0      0         1      0      1     0
         administ admiss admit advanc advis advisor advocaci afternoon agenc agre aid aim algeb
A3SR.txt        0     20     0      7     1       5         1         0     0    0   9   0
AQR.txt         1      2     2      3     0       0         0         1     1    0   3   1
QMSS.txt        0      5     2      0     1       0         0         0     0    1   0   0
         algorithm ali align allow almost along alp also alter although alumni ambit american a
A3SR.txt         1   2     1     2      0     0   0    6     1         1      1     0        2
AQR.txt          0   0     0     3      1     0   0    3     0         0      0     1        1
QMSS.txt         0   0     0     0      0     1   1    1     0         0      1     0        2
```

Looking at some words

What are the most common words in this corpus

```
> findFreqTerms(dtm, lowfreq=20)
 [1] "admiss"   "appli"    "applic"   "aqr"      "can"      "cours"    "graduat"
"nyu"      "offer"
[10] "program"  "qmss"     "quantit"  "requir"   "research" "scienc"   "score"
"statist"  "student"
[19] "will"     "work"
```

What are the most common words in each document:

```
> findFreqTerms(dtm[1,], lowfreq=20)
[1] "admiss"  "applic"  "program" "scienc"  "statist" "student"

> findFreqTerms(dtm[2,], lowfreq=20)
[1] "aqr"     "program" "student"

> findFreqTerms(dtm[3,], lowfreq=20)
[1] "applic"  "program" "qmss"    "student"
```

Document-term matrix -> dataframe

Transpose the dataframe so columns are documents ...

```
> t.df.dtm=t(df.dtm)

> head(t.df.dtm)
        A3SR.txt AQR.txt QMSS.txt
abil           2       0        2
abl            2       0        5
abroad         0       1        0
academ         4       4        9
acceler        2       0        0
accept         1       1        4
```

Measures of similarity

Correlation:

```
> cor(t.df.dtm)
          A3SR.txt   AQR.txt  QMSS.txt
A3SR.txt 1.0000000 0.6688018 0.5793769
AQR.txt  0.6688018 1.0000000 0.6764536
QMSS.txt 0.5793769 0.6764536 1.0000000
```

Cosine:

```
install.packages("lsa")
library(lsa)
> cosine(t.df.dtm)
          A3SR.txt   AQR.txt  QMSS.txt
A3SR.txt 1.0000000 0.7213635 0.6468832
AQR.txt  0.7213635 1.0000000 0.7189285
QMSS.txt 0.6468832 0.7189285 1.0000000
```

Measures of similarity

Chi-square test:

```
> ctable <- table(t.df.dtm)
> chisq.test(ctable)

     Chi-squared test for given probabilities

data:  ctable
X-squared = 20148.62, df = 28, p-value < 2.2e-16
```

Conclusion so far

These documents are similar, though different :)

Making the matrix more dense

This further finds only frequently occurring words

```
> dtms <- removeSparseTerms(dtm, 0.05) # This makes a matrix that is 5% empty space, maximum.
> inspect(dtms)
A document-term matrix (3 documents, 129 terms)

Non-/sparse entries: 387/0
Sparsity           : 0%
Maximal term length: 10
Weighting          : term frequency (tf)

          Terms
Docs       academ accept acquir admiss also american analysi analyt appli applic appropri ask a
  A3SR.txt      4      1      1     20    6        2       7      2    18     21         2   1
  AQR.txt       4      1      1      2    3        1       6      1    10     10         1   1
  QMSS.txt      9      4      1      5    1        2       1      1     3     22         1   1
          Terms
Docs       background can career choos class communic complet countri cours deadlin decis degre
  A3SR.txt          5  18      5     2     6        1      12       1    19       3     6    10
  AQR.txt           2   4      4     2     1        4       3       1     3       2     1     5
  QMSS.txt          1   6      2     1     3        1       2       1     7       2     2     1
          Terms
Docs       design direct educ effect elect emphasi encourag english enter evalu even experi fie
  A3SR.txt      2      2    4      2     3       1        2       2     2     1    1      5
  AQR.txt       2      1    1      1     1       2        1       2     3     2    1      4
  QMSS.txt      3      4    4      1     1       1        2       1     2     2    2      6
          Terms
Docs       first flexibl fulltim graduat gre help hous howev import improv increas innov instit
  A3SR.txt     1       3       5      19   4    3    5     3      2       1       1     1
  AQR.txt      1       2       2      15   7    1    2     5      1       1       1     1
```

Weighting the matrix by tf-idf

Weight by term frequency, inverse document frequency

```
> terms <-DocumentTermMatrix(docs,control = list(weighting = function(x) weightTfIdf(x,
normalize = FALSE)))
```

```
> inspect(terms)
A document-term matrix (3 documents, 796 terms)

Non-/sparse entries: 855/1533
Sparsity           : 64%
Maximal term length: 45
Weighting          : term frequency - inverse document frequency (tf-idf)

          Terms
Docs           abil      abl   abroad academ  acceler accept   access   accord  accredit   achie
  A3SR.txt 1.169925 1.169925 0.000000      0 3.169925      0 1.584963 1.584963 0.5849625 3.1699
  AQR.txt  0.000000 0.000000 1.584963      0 0.000000      0 0.000000 0.000000 0.0000000 0.0000
  QMSS.txt 1.169925 2.924813 0.000000      0 0.000000      0 0.000000 0.000000 0.5849625 0.0000
          Terms
Docs       acquir    addit   address adminis admiss    admit   advanc     advis  advisor advoca
  A3SR.txt      0 2.924813 2.3398500 0.000000      0 0.000000 4.094738 0.5849625 7.924813 1.584
  AQR.txt       0 1.754888 0.5849625 1.584963      0 1.169925 1.754888 0.0000000 0.000000 0.000
  QMSS.txt      0 0.000000 0.0000000 0.000000      0 1.169925 0.000000 0.5849625 0.000000 0.000
          Terms
Docs       afternoon    agenc     agre      aid      aim  algebra algorithm      ali    align
  A3SR.txt  0.000000 0.000000 0.000000 5.264663 0.000000 1.584963  1.584963 3.169925 1.584963 1
  AQR.txt   1.584963 1.584963 0.000000 1.754888 1.584963 0.000000  0.000000 0.000000 0.000000 1
  QMSS.txt  0.000000 0.000000 1.584963 0.000000 0.000000 0.000000  0.000000 0.000000 0.000000 0
```

Weighting the matrix by tf-idf

Most common words on this basis are:

```
> terms <-DocumentTermMatrix(docs,control = list(weighting = function(x) weightTfIdf(x,
normalize = FALSE)))
```

```
> findFreqTerms(terms[1,], lowfreq=4)
 [1] "advanc"              "advisor"             "aid"                 "assist"
 [5] "candid"              "close"               "colleg"              "comput"
 [9] "credit"              "data"                "detail"              "elig"
[13] "extend"              "faculti"             "fee"                 "financi"
[17] "form"                "found"               "health"              "infer"
[21] "inform"              "internship"          "linear"              "list"
[25] "mail"                "mathemat"            "may"                 "member"
[29] "model"               "nyu"                 "offcampus"           "onlin"
[33] "page"                "parttim"             "polici"              "practic"
[37] "prior"               "problem"             "process"             "project"
[41] "relat"               "result"              "scholarship"         "solv"
[45] "steinhardt"          "steinhardtasrnyuedu" "still"               "summer"
[49] "supplement"          "three"               "websit"              "youll"

> findFreqTerms(terms[2,], lowfreq=4)
 [1] "aqr"     "award"   "citi"    "locat"   "nyu"     "tailor"  "test"    "tool"
"version" "world"

> findFreqTerms(terms[3,], lowfreq=4)
 [1] "back"     "columbia" "continu"  "maintain" "qmss"     "qualiti"  "second"   "top"
"vari"
[10] "write"
```

Straight relative frequencies, too

Develop relative frequencies and use those instead, a la Neal Caren

```
> terms <-DocumentTermMatrix(docs,control = list(weighting = function(x) weightTfIdf(x,
normalize = FALSE)))
```

```
> summing = function(x) x/sum(x, na.rm=T)

> t.df.dtm.df_new = apply(t.df.dtm, 2, summing)

> head(t.df.dtm.df new)
            A3SR.txt     AQR.txt    QMSS.txt
abil    0.0012903226 0.000000000 0.002604167
abl     0.0012903226 0.000000000 0.006510417
abroad  0.0000000000 0.001239157 0.000000000
academ  0.0025806452 0.004956629 0.011718750
acceler 0.0012903226 0.000000000 0.000000000
accept  0.0006451613 0.001239157 0.005208333
```

Now- How to do the next step

Develop relative frequencies and use those instead, a la Neal Caren

```
t.df.dtm.df new = as.data.frame(t.df.dtm.df new)
names(t.df.dtm.df new)[names(t.df.dtm.df new)=="1"] <- "A3SR"
names(t.df.dtm.df new)[names(t.df.dtm.df new)=="2"] <- "CUSP"
names(t.df.dtm.df_new)[names(t.df.dtm.df_new)=="3"] <- "QMSS"
```

How to do the next step

What words are most like QMSS and least like A3SR

```
t.df.dtm.df new$ratio = t.df.dtm.df new$QMSS - t.df.dtm.df new$A3SR
sort.QMSS <- t.df.dtm.df_new[order(-t.df.dtm.df_new$ratio) , ]
sort.QMSS[1:15, ]
```

	A3SR	CUSP	QMSS	ratio
students	0.010204082	0.000000000	0.066115702	0.055911621
qmss	0.000000000	0.000000000	0.041322314	0.041322314
program	0.023809524	0.017316017	0.049586777	0.025777253
quantitative	0.003401361	0.000000000	0.024793388	0.021392028
research	0.022108844	0.004329004	0.041322314	0.019213471
directly	0.000000000	0.000000000	0.016528926	0.016528926
incoming	0.000000000	0.000000000	0.016528926	0.016528926
year	0.000000000	0.000000000	0.016528926	0.016528926
experience	0.001700680	0.004329004	0.016528926	0.014828245
techniques	0.003401361	0.000000000	0.016528926	0.013127565
acquire	0.000000000	0.000000000	0.008264463	0.008264463
approximately	0.000000000	0.000000000	0.008264463	0.008264463
arts	0.000000000	0.000000000	0.008264463	0.008264463
basis	0.000000000	0.000000000	0.008264463	0.008264463
bring	0.000000000	0.000000000	0.008264463	0.008264463

How to do the next step

What words are most like A3SR and least like QMSS

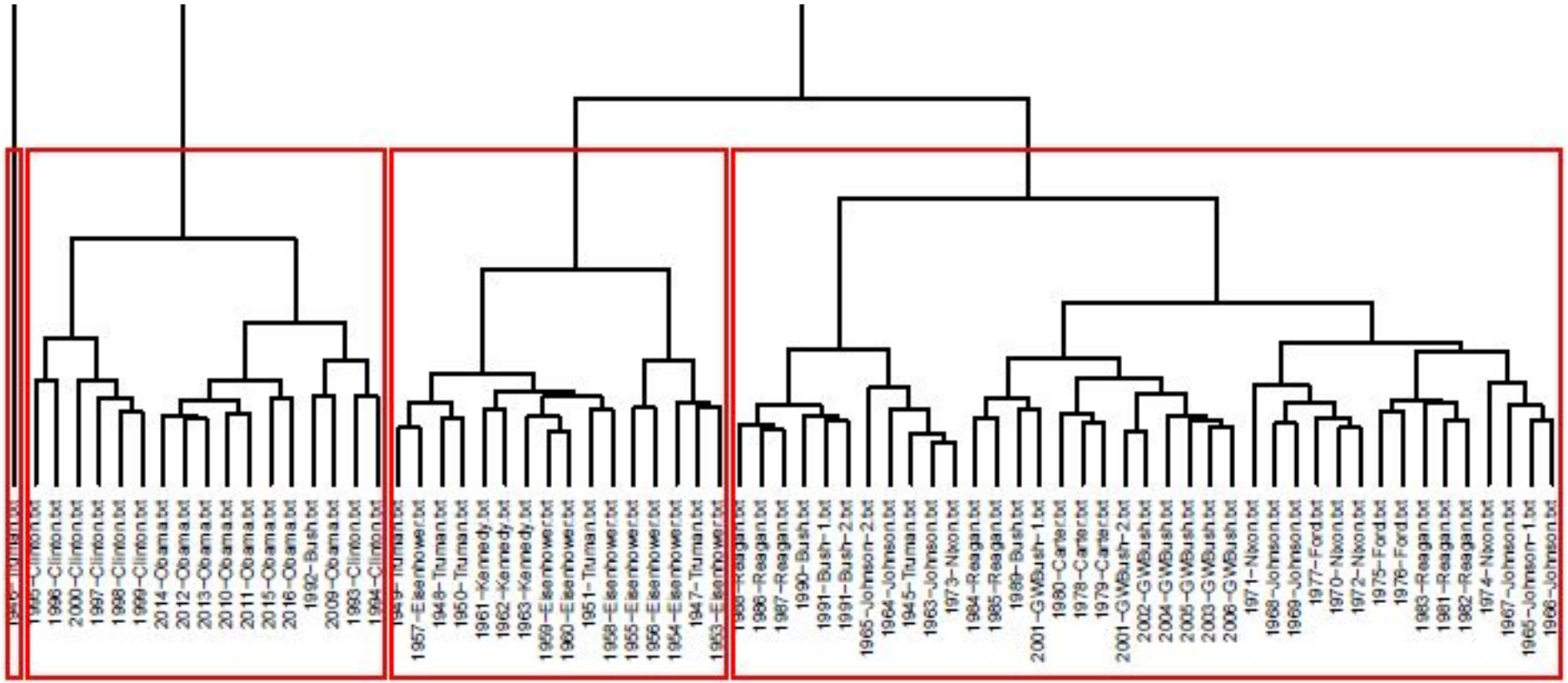
```
sort.QMSS <- t.df.dtm.df_new[order(t.df.dtm.df_new$ratio) , ]
sort.QMSS[1:15, ]
```

	A3SR	CUSP	QMSS	ratio
data	0.023809524	0.021645022	0	-0.023809524
statistical	0.020408163	0.000000000	0	-0.020408163
model	0.015306122	0.000000000	0	-0.015306122
will	0.015306122	0.021645022	0	-0.015306122
applied	0.010204082	0.008658009	0	-0.010204082
health	0.010204082	0.000000000	0	-0.010204082
develop	0.008503401	0.004329004	0	-0.008503401
nyu	0.008503401	0.000000000	0	-0.008503401
policy	0.008503401	0.000000000	0	-0.008503401
public	0.008503401	0.000000000	0	-0.008503401
analysis	0.006802721	0.000000000	0	-0.006802721
can	0.006802721	0.000000000	0	-0.006802721
faculty	0.006802721	0.000000000	0	-0.006802721
important	0.006802721	0.000000000	0	-0.006802721
time	0.006802721	0.000000000	0	-0.006802721

2-- Using clustering to determine authorship

Comparing “modern” SOUs

- Via hierarchical clustering of most commonly occurring words, excluding stop words



What did I do to make this?

Hierarchical clustering I

```
cname <- file.path("C:\\Users\\gme2101\\Documents\\Spring 2016", "sous")
cname
dir(cname)

library(tm)
docs <- Corpus(DirSource(cname))

summary(docs)

inspect(docs[2])

dtm <- DocumentTermMatrix(docs)
```

What did I do to make this?

Hierarchical clustering II

```
dtmss <- removeSparseTerms(dtm, 0.15) # This makes a matrix that is only 15% empty
maximum.
inspect(dtmss)

library(cluster)
d <- dist(dtmss, method="euclidian")
fit <- hclust(d=d, method="ward")
fit

plot(fit, hang=-1, cex=.2)

plot.new()
plot(fit, hang=-1, cex=.2)
groups <- cutree(fit, k=13)   # "k=" defines the number of clusters you are using
rect.hclust(fit, k=13, border="red") # draw dendogram with red borders around the
```

Other considerations

- What is the distance measure for hierarchical clustering?

Rank Distance as a Stylistic Similarity

Marius Popescu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
mpopescu@phobos.cs.unibuc.ro

Liviu P. Dinu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
ldinu@funinf.cs.unibuc.ro

Abstract

In this paper we propose a new distance

elty of our approach resides in the way we use information given by the function word frequen-

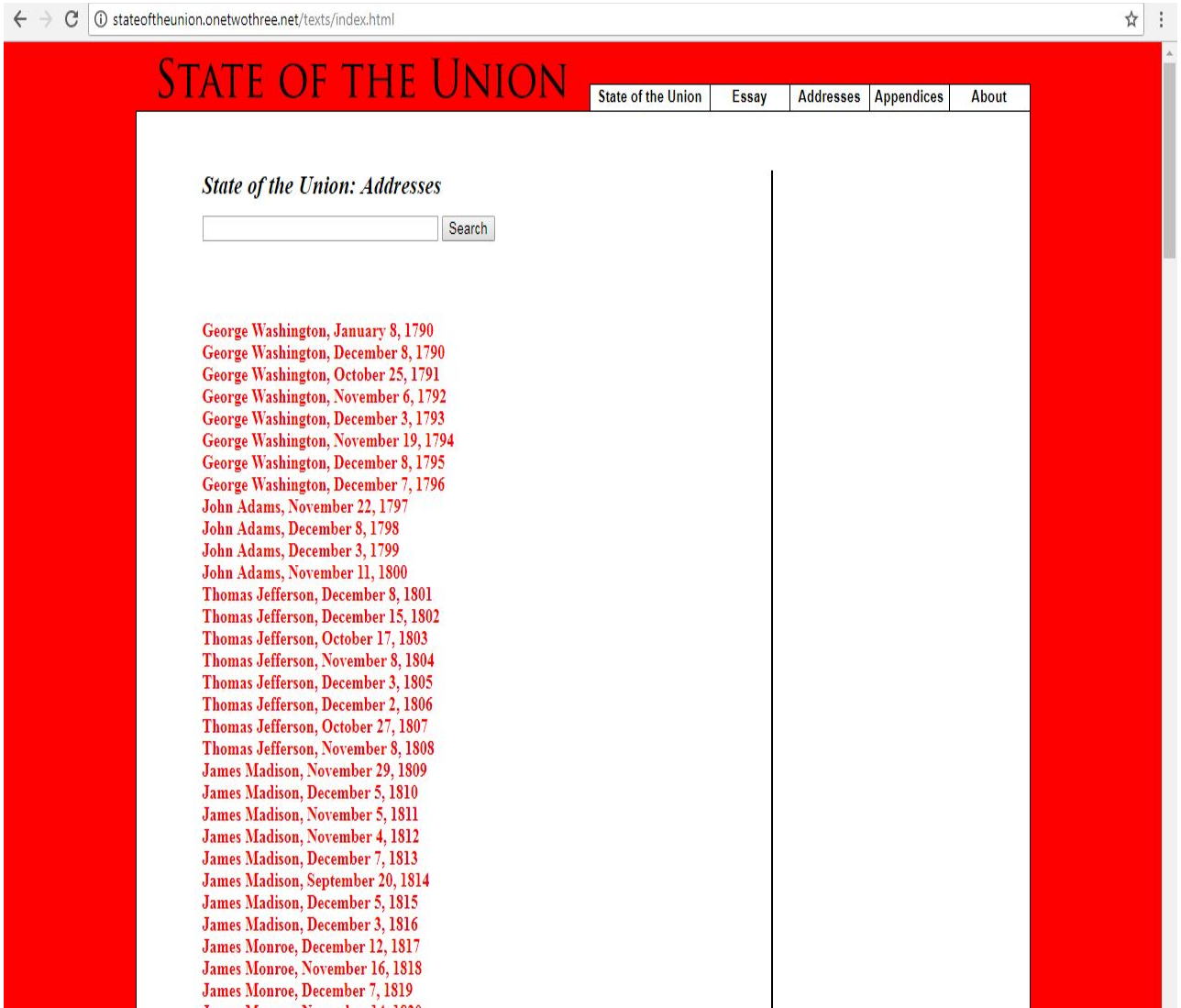
Other examples and thoughts

More R code forthcoming …

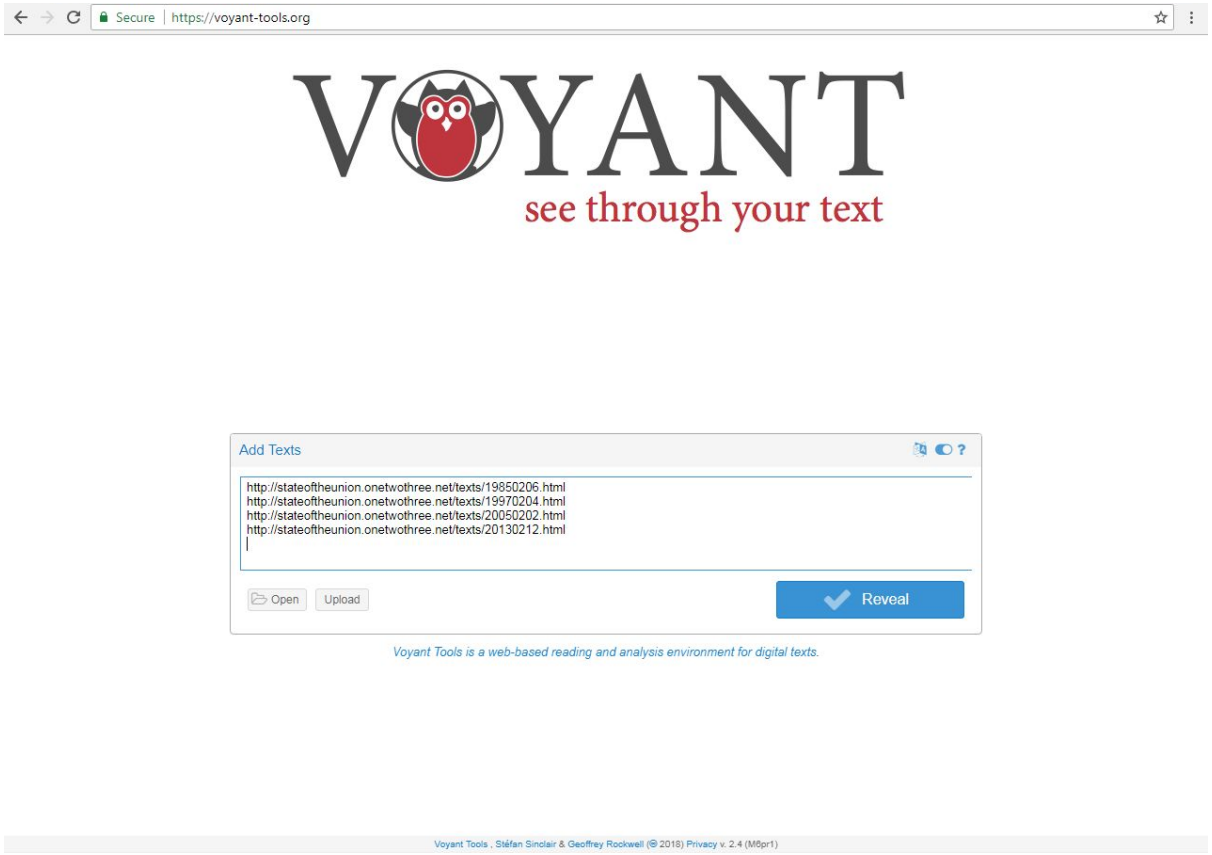
… on the hierarchical clustering

Some more on the SOU examples (in VOYANT too)

Our source material is here



Our quantitative text analysis starts here:



What did I type into that search bar?

http://stateoftheunion.onetwothree.net/texts/19850206.html
http://stateoftheunion.onetwothree.net/texts/19970204.html
http://stateoftheunion.onetwothree.net/texts/20050202.html
http://stateoftheunion.onetwothree.net/texts/20130212.html

- These are the last 4 States of the Union speeches given by a re-elected President in the year after they were re-elected:
- Reagan in 1985; Clinton in 1997; W. Bush in 2005; and Obama in 2013.

Welcome to your dashboard!

Cirrus | Terms | Links | Reader | TermsBerry | Trends | Document Terms

Scale | Terms:

State of the Union Address | Ronald Reagan | February 6, 1985

State of the Union Essay Addresses Appendices About

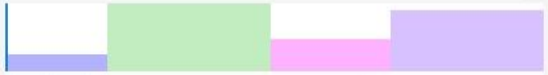
State of the Union Address

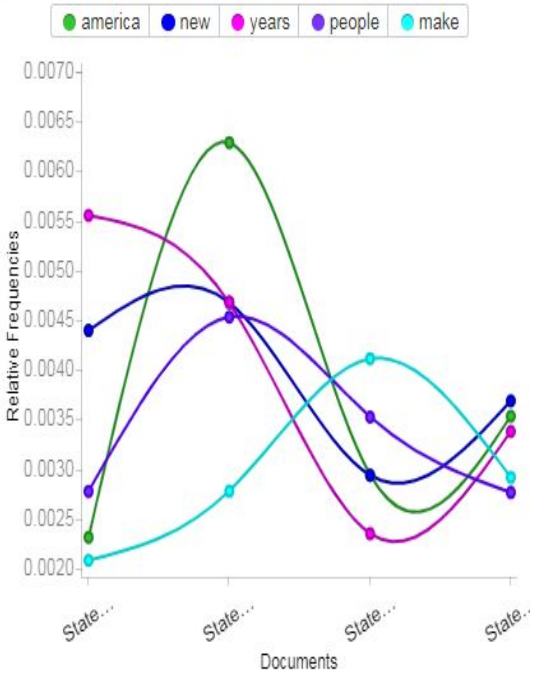
< Previous Next >

Ronald Reagan February 6, 1985

Mr. Speaker, Mr. President, distinguished Members of the Congress, honored guests, and fellow citizens:

I come before you to report on the state of our Union, and I'm pleased to report that after 4 years of united effort, the American people have brought forth a nation renewed, stronger, freer, and more secure than before.





Scale | Frequenci

Summary | Documents | Phrases

This corpus has 4 documents with 22,780 total words and 3,666 unique word forms. Created now.

Document Length:

- Longest: State of the Union Addres... (6840); State of the Union Addres... (6511)
- Shortest: State of the Union Addres... (4322); State of the Union Addres... (5107)

Vocabulary Density:

- Highest: State of the Union Addres... (0.323); State of the Union Addres... (0.292)
- Lowest: State of the Union Addres... (0.236); State of the Union Addres... (0.266)

Average Words Per Sentence:

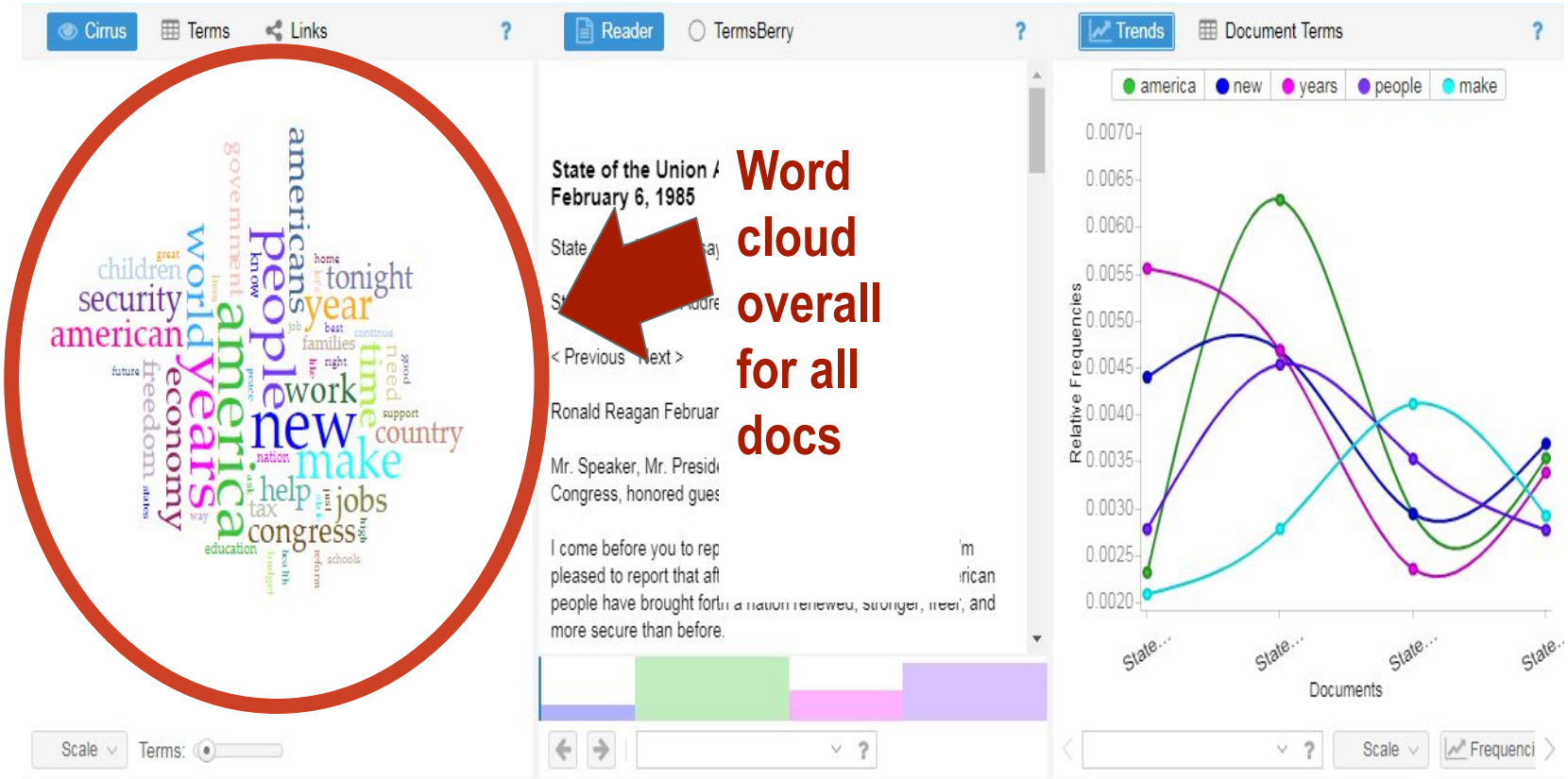
- Highest: State of the Union Addres... (21.3); State of the Union Addres... (19.9)
- Lowest: State of the Union Addres... (19.1); State of the Union Addres... (19.6)

items:

Contexts | Bubblelines | Correlations

Document	Left	Term	Right
1) State...	role in the world. Tonight	america	is stronger because of the
1) State...	for future success: For an	america	of wisdom that honors the
1) State...	goes our civilization; For an	america	of vision that sees tomorrow's
1) State...	we do today; For an	america	of courage whose service men
1) State...	frontiers of freedom; For an	america	of compassion that opens its
1) State...	that taps the soul of	america	, enabling us to summon greater
1) State...	engine of our dreams and	america	the investment capital of the
1) State...	the changes that have swept	america	the past 4 years, none
1) State...	of our assistance to Central	america	. I want to work with

91 context | expand | Scale



Summary | Documents | Phrases

This corpus has 4 documents with 22,780 total words and 3,666 unique word forms. Created now.

Document Length:

- Longest: State of the Union Addres... (6840); State of the Union Addres... (6511)
- Shortest: State of the Union Addres... (4322); State of the Union Addres... (5107)

Vocabulary Density:

- Highest: State of the Union Addres... (0.323); State of the Union Addres... (0.292)
- Lowest: State of the Union Addres... (0.236); State of the Union Addres... (0.266)

Average Words Per Sentence:

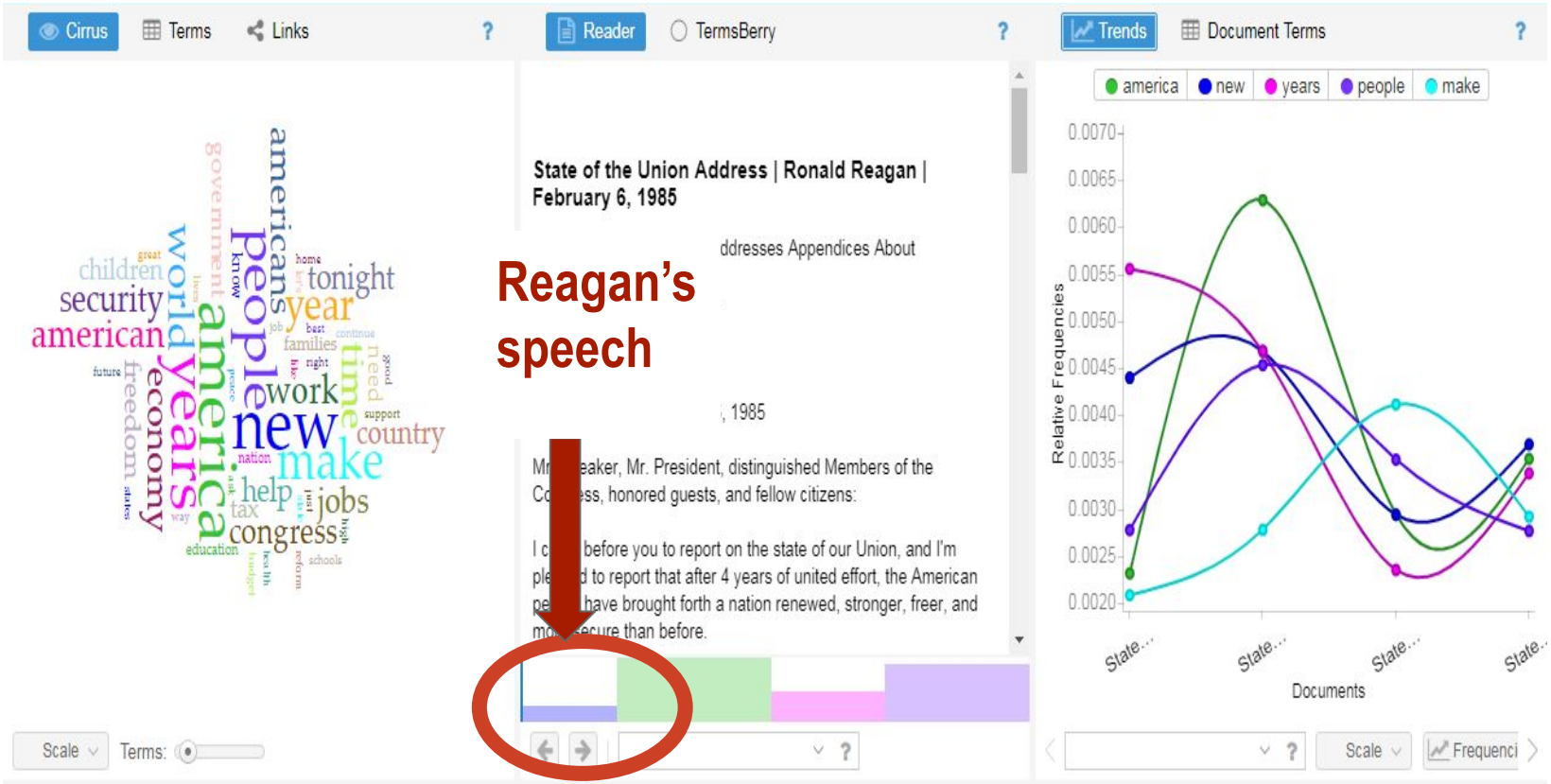
- Highest: State of the Union Addres... (21.3); State of the Union Addres... (19.9)
- Lowest: State of the Union Addres... (19.1); State of the Union Addres... (19.6)

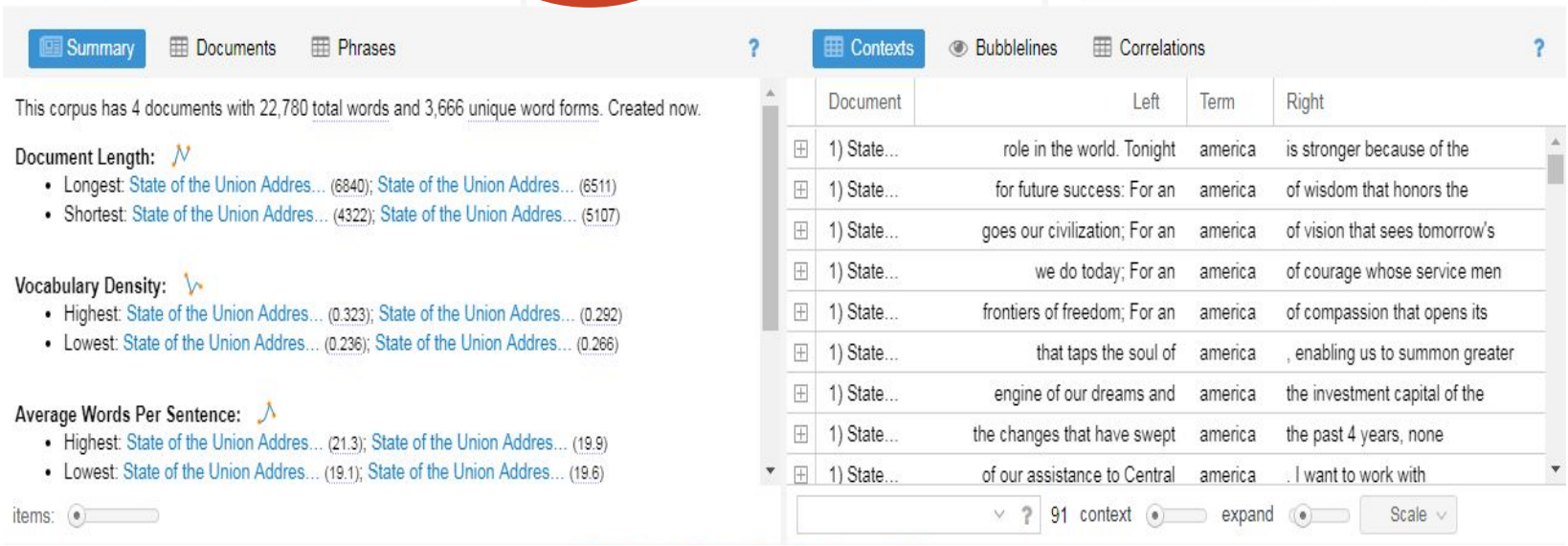
items:

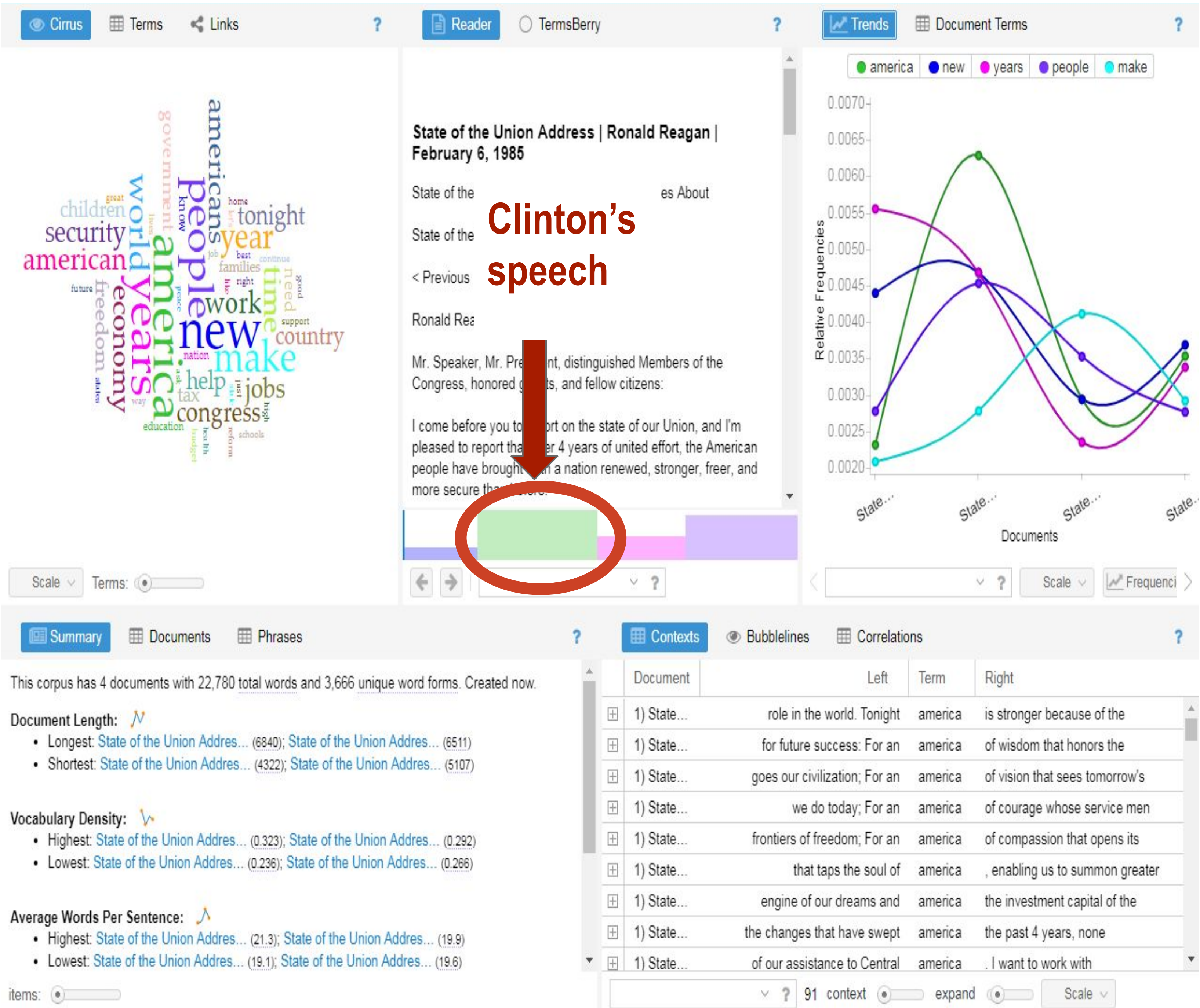
Contexts | Bubblelines | Correlations

Document	Left	Term	Right
1) State...	role in the world. Tonight	america	is stronger because of the
1) State...	for future success: For an	america	of wisdom that honors the
1) State...	goes our civilization; For an	america	of vision that sees tomorrow's
1) State...	we do today; For an	america	of courage whose service men
1) State...	frontiers of freedom; For an	america	of compassion that opens its
1) State...	that taps the soul of	america	, enabling us to summon greater
1) State...	engine of our dreams and	america	the investment capital of the
1) State...	the changes that have swept	america	the past 4 years, none
1) State...	of our assistance to Central	america	. I want to work with

91 context expand Scale


Cirrus
Terms
Links
Reader
TermsBerry
Trends
Document Terms
State of the Union Address | Ronald Reagan | February 6, 1985
Reagan's speech
, 1985
Mr. Speaker, Mr. President, distinguished Members of the Co ess, honored guests, and fellow citizens:
I c before you to report on the state of our Union, and I'm pl d to report that after 4 years of united effort, the American pe have brought forth a nation renewed, stronger, freer, and m ecure than before.
america
new
years
people
make
Relative Frequencies
0.0070
0.0065
0.0060
0.0055
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
State...
Documents
Scale
Terms:
Frequenci


Summary
Documents
Phrases
This corpus has 4 documents with 22,780 total words and 3,666 unique word forms. Created now.
Document Length:
Longest: State of the Union Addres... (6840); State of the Union Addres... (6511)
Shortest: State of the Union Addres... (4322); State of the Union Addres... (5107)
Vocabulary Density:
Highest: State of the Union Addres... (0.323); State of the Union Addres... (0.292)
Lowest: State of the Union Addres... (0.236); State of the Union Addres... (0.266)
Average Words Per Sentence:
Highest: State of the Union Addres... (21.3); State of the Union Addres... (19.9)
Lowest: State of the Union Addres... (19.1); State of the Union Addres... (19.6)
items:
Contexts
Bubblelines
Correlations
Document
Left
Term
Right
1) State... role in the world. Tonight america is stronger because of the
1) State... for future success: For an america of wisdom that honors the
1) State... goes our civilization; For an america of vision that sees tomorrow's
1) State... we do today; For an america of courage whose service men
1) State... frontiers of freedom; For an america of compassion that opens its
1) State... that taps the soul of america , enabling us to summon greater
1) State... engine of our dreams and america the investment capital of the
1) State... the changes that have swept america the past 4 years, none
1) State... of our assistance to Central america . I want to work with
91 context
expand
Scale


Cirrus
Terms
Links
Reader
TermsBerry
Trends
Document Terms
america
new
years
people
make
State of the Union Address | Ronald Reagan | February 6, 1985
State of the
es About
State of the
< Previous
Ronald Rea
Mr. Speaker, Mr. Pre nt, distinguished Members of the Congress, honored g ts, and fellow citizens:
I come before you to rt on the state of our Union, and I'm pleased to report tha r 4 years of united effort, the American people have brought n a nation renewed, stronger, freer, and more secure tha
Clinton's speech
Relative Frequencies
Documents
Scale
Terms:
Frequenci
Summary
Documents
Phrases
Contexts
Bubblelines
Correlations
This corpus has 4 documents with 22,780 total words and 3,666 unique word forms. Created now.
Document Length:
Longest: State of the Union Addres... (6840); State of the Union Addres... (6511)
Shortest: State of the Union Addres... (4322); State of the Union Addres... (5107)
Vocabulary Density:
Highest: State of the Union Addres... (0.323); State of the Union Addres... (0.292)
Lowest: State of the Union Addres... (0.236); State of the Union Addres... (0.266)
Average Words Per Sentence:
Highest: State of the Union Addres... (21.3); State of the Union Addres... (19.9)
Lowest: State of the Union Addres... (19.1); State of the Union Addres... (19.6)
items:
Document
Left
Term
Right
1) State... role in the world. Tonight america is stronger because of the
1) State... for future success: For an america of wisdom that honors the
1) State... goes our civilization; For an america of vision that sees tomorrow's
1) State... we do today; For an america of courage whose service men
1) State... frontiers of freedom; For an america of compassion that opens its
1) State... that taps the soul of america , enabling us to summon greater
1) State... engine of our dreams and america the investment capital of the
1) State... the changes that have swept america the past 4 years, none
1) State... of our assistance to Central america . I want to work with
91 context
expand
Scale

Lots more to look at here

The (relatively) most used words

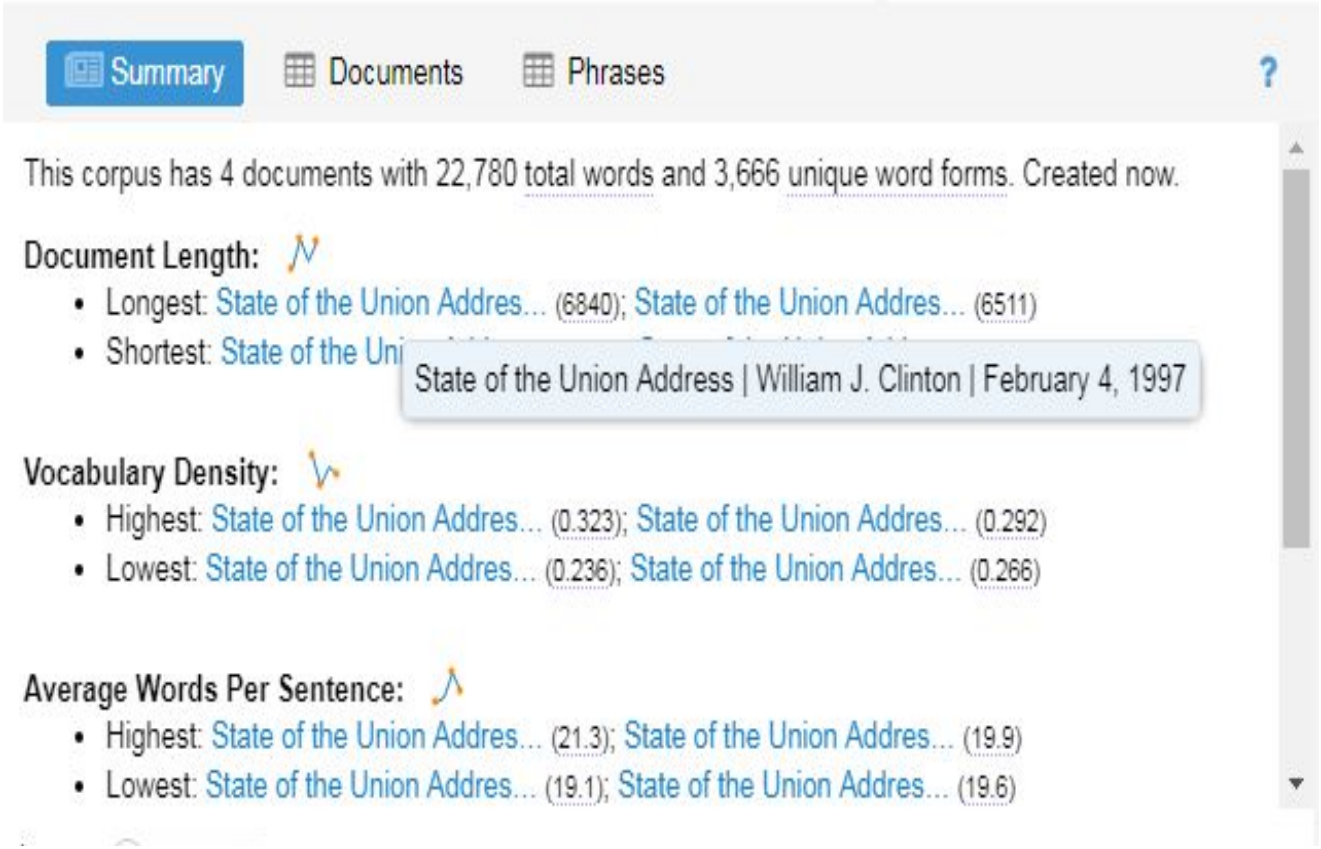
- Clinton (Speech #2) used *america* and *new* the most of any president
- W. Bush (Speech #3) used *security* the next most
- Reagan (Speech #1) used the word *freedom* (and *tax*) a whole lot
- Obama (Speech #4) used the word *jobs* a whole bunch too

#	Term	Count	Relative	Trend
2	america	43	6,287	
3	security	29	5,678	
1	years	24	5,553	
4	jobs	31	4,761	
2	new	32	4,678	
2	years	32	4,678	
2	people	31	4,532	
1	freedom	19	4,396	
1	new	19	4,396	
3	make	21	4,112	
2	world	27	3,947	
1	american	16	3,702	
1	tax	16	3,702	
4	new	24	3,686	
4	america	23	3,532	

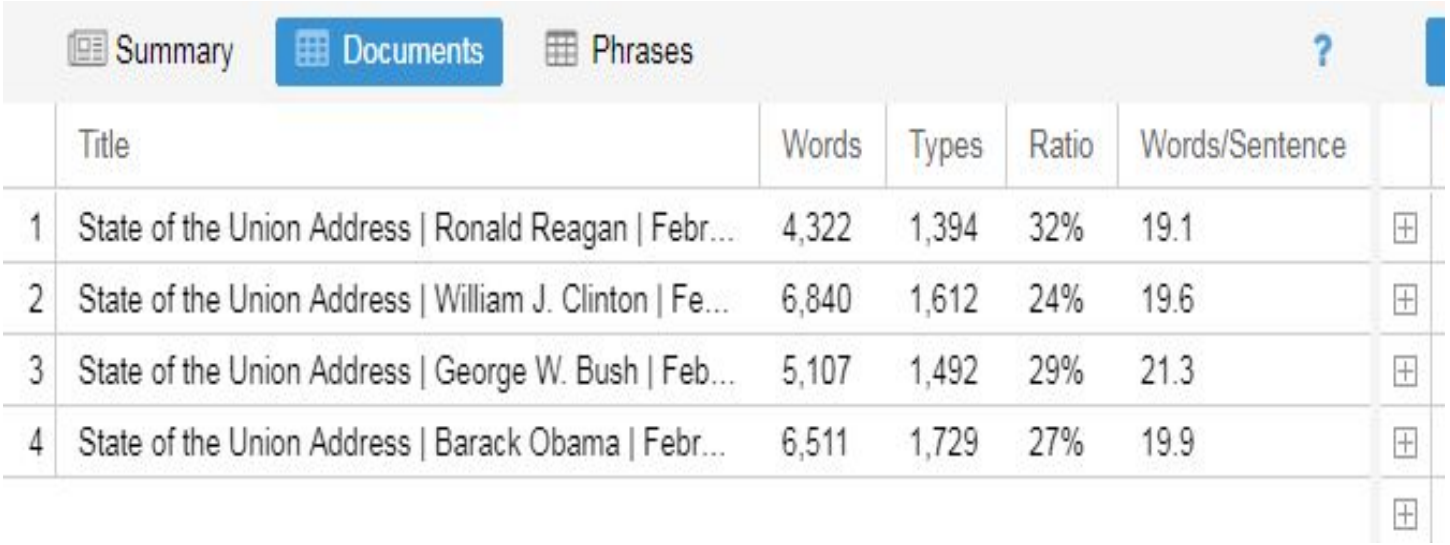
Who had the longest speech?

The longest speech

- No surprise:



Summarizing the docs a bit more

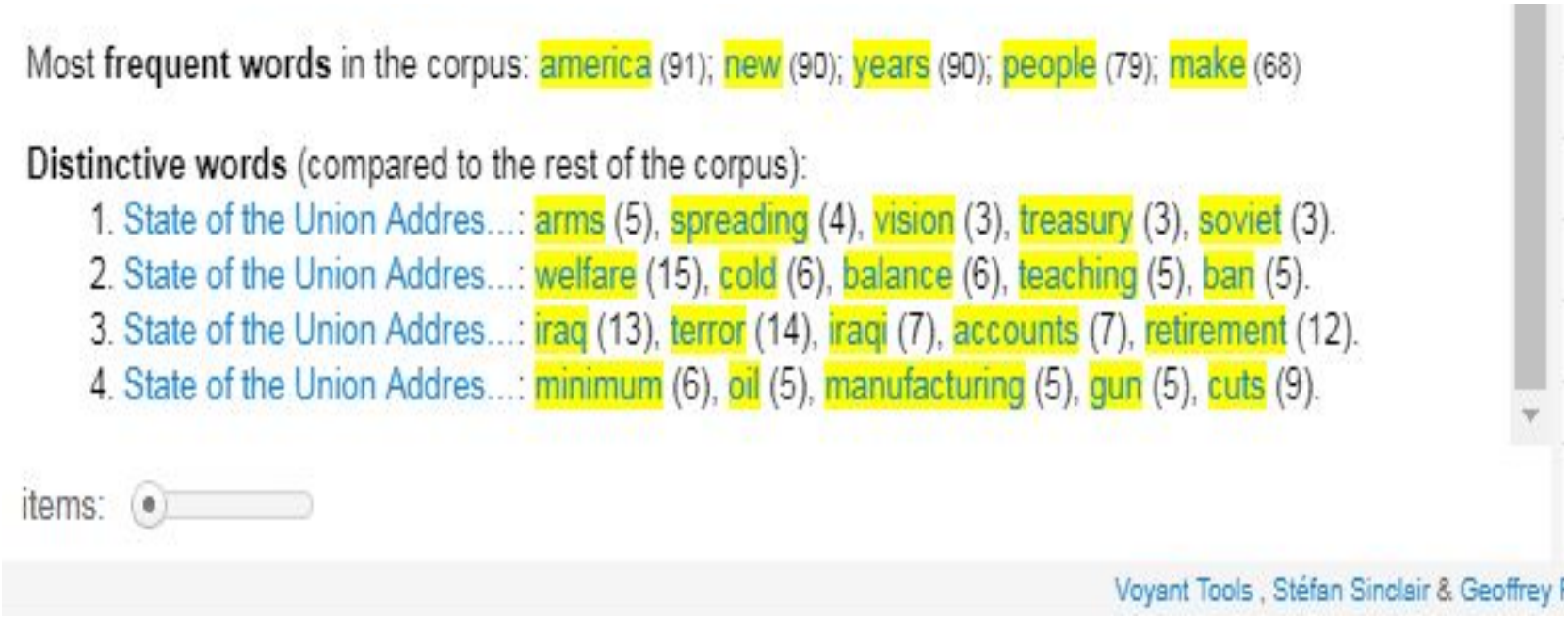

Summary Documents Phrases ?

	Title	Words	Types	Ratio	Words/Sentence
1	State of the Union Address | Ronald Reagan | Febr...	4,322	1,394	32%	19.1
2	State of the Union Address | William J. Clinton | Fe...	6,840	1,612	24%	19.6
3	State of the Union Address | George W. Bush | Feb...	5,107	1,492	29%	21.3
4	State of the Union Address | Barack Obama | Febr...	6,511	1,729	27%	19.9

- *Types* are are the number of word forms found in the document (e.g. all occurrences of "the" are counted as one word form)
- *Ratio* is "types over tokens," expressed as a percentage – where higher numbers generally mean greater vocabulary diversity and less repetition

More stuff on the bottom left

- What you get:



Word cloud overall

- Any patterns?

The (relatively) most used words

- Clinton (Speech #2) used *new* the most of any president
- W. Bush (Speech #3) used *security* the next most
- Reagan (Speech #1) used the word *freedom* (and *tax*) a whole lot
- Obama (Speech #4) used the word *jobs* a whole bunch too

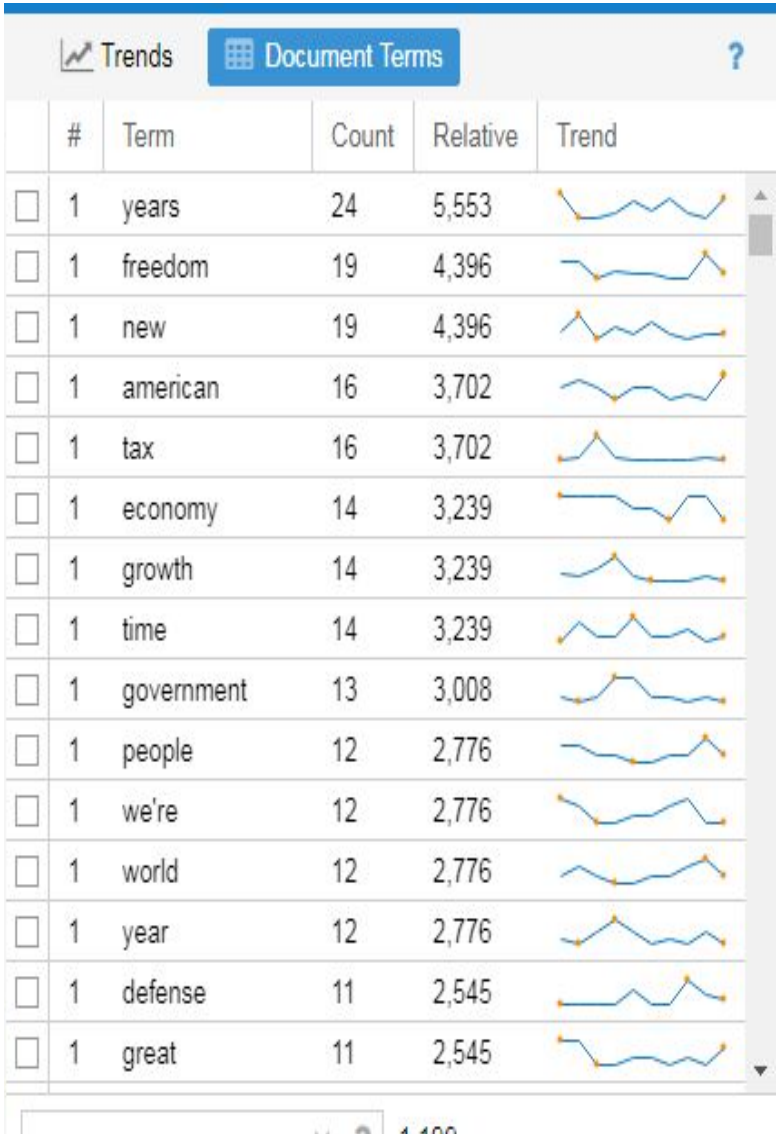
#	Term	Count	Relative	Trend
2	america	43	6,287	
3	security	29	5,678	
1	years	24	5,553	
4	jobs	31	4,761	
2	new	32	4,678	
2	years	32	4,678	
2	people	31	4,532	
1	freedom	19	4,396	
1	new	19	4,396	
3	make	21	4,112	
2	world	27	3,947	
1	american	16	3,702	
1	tax	16	3,702	
4	new	24	3,686	
4	america	23	3,532	

Word cloud - Reagan in 1985

- What do you notice?

Top terms - Reagan in 1985

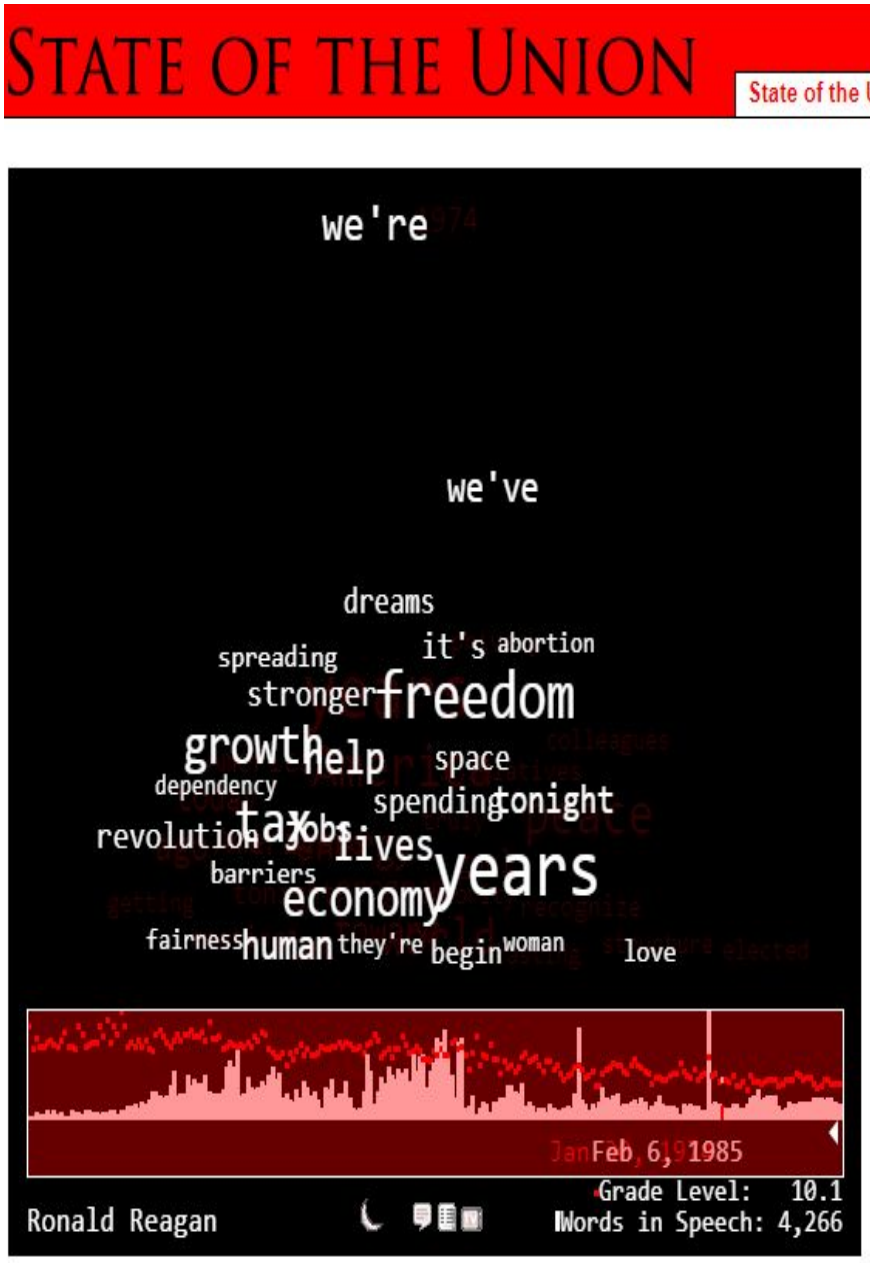
- What do you notice?



#	Term	Count	Relative	Trend
1	years	24	5,553	
1	freedom	19	4,396	
1	new	19	4,396	
1	american	16	3,702	
1	tax	16	3,702	
1	economy	14	3,239	
1	growth	14	3,239	
1	time	14	3,239	
1	government	13	3,008	
1	people	12	2,776	
1	we're	12	2,776	
1	world	12	2,776	
1	year	12	2,776	
1	defense	11	2,545	
1	great	11	2,545	

Reagan in 1985

- The picture from the State of the Union website too
- Does it comport with what we have just seen?

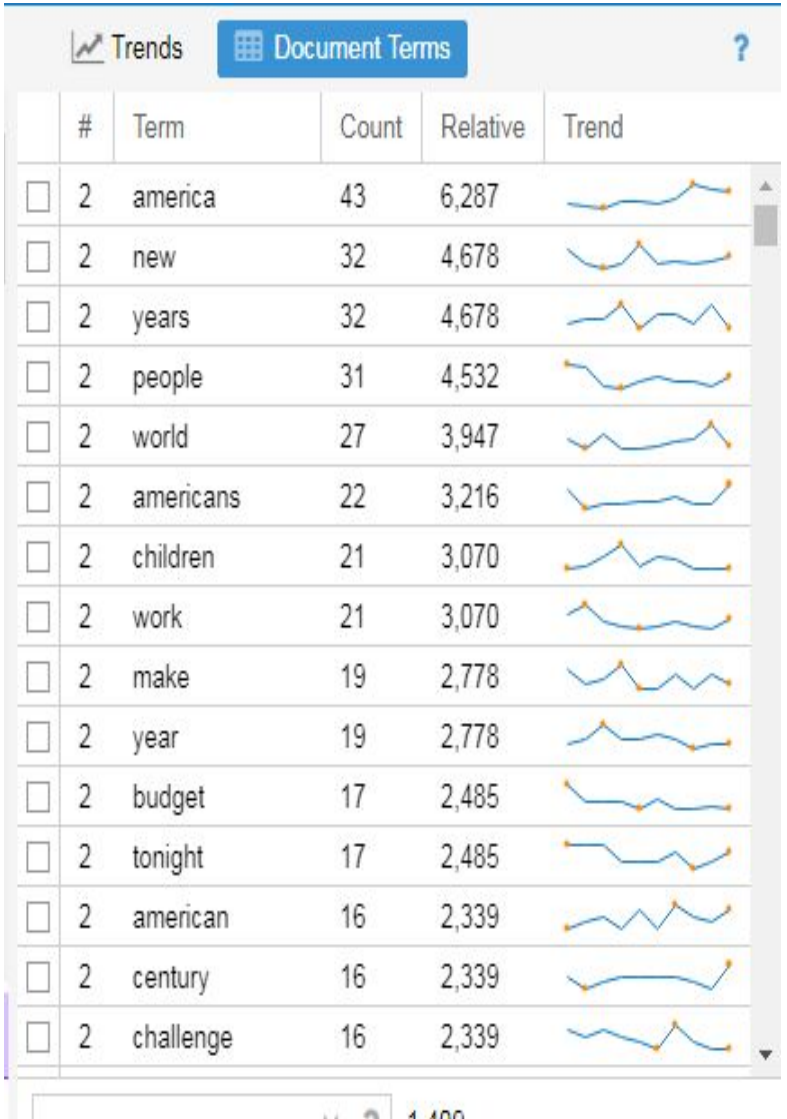


Word cloud - Clinton in 1997

- What do you see?

Top terms - Clinton in 1997

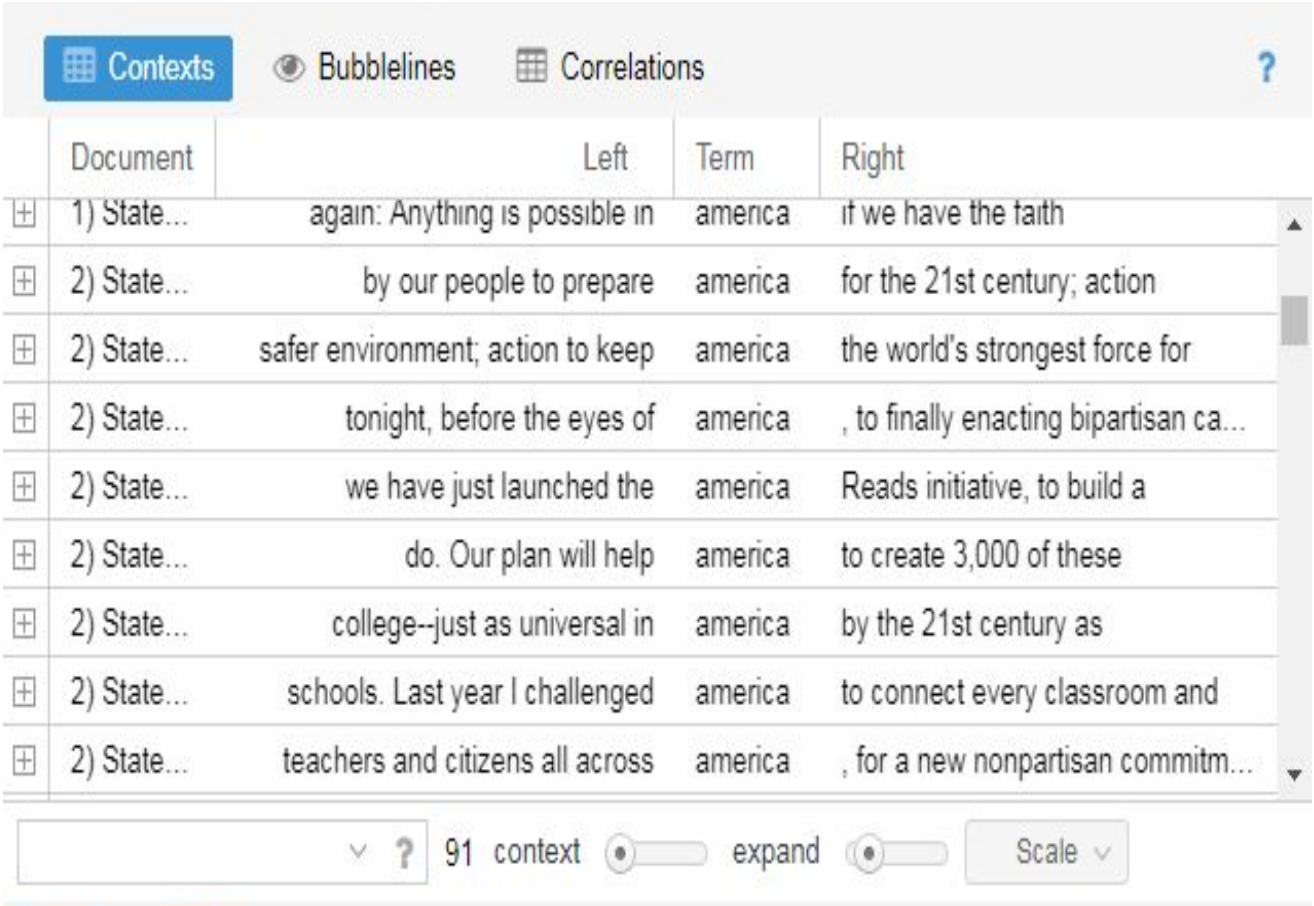
- What do you see now?



#	Term	Count	Relative	Trend
2	america	43	6,287	
2	new	32	4,678	
2	years	32	4,678	
2	people	31	4,532	
2	world	27	3,947	
2	americans	22	3,216	
2	children	21	3,070	
2	work	21	3,070	
2	make	19	2,778	
2	year	19	2,778	
2	budget	17	2,485	
2	tonight	17	2,485	
2	american	16	2,339	
2	century	16	2,339	
2	challenge	16	2,339	

“America” in context for Clinton

- What you get:



Clinton in 1997

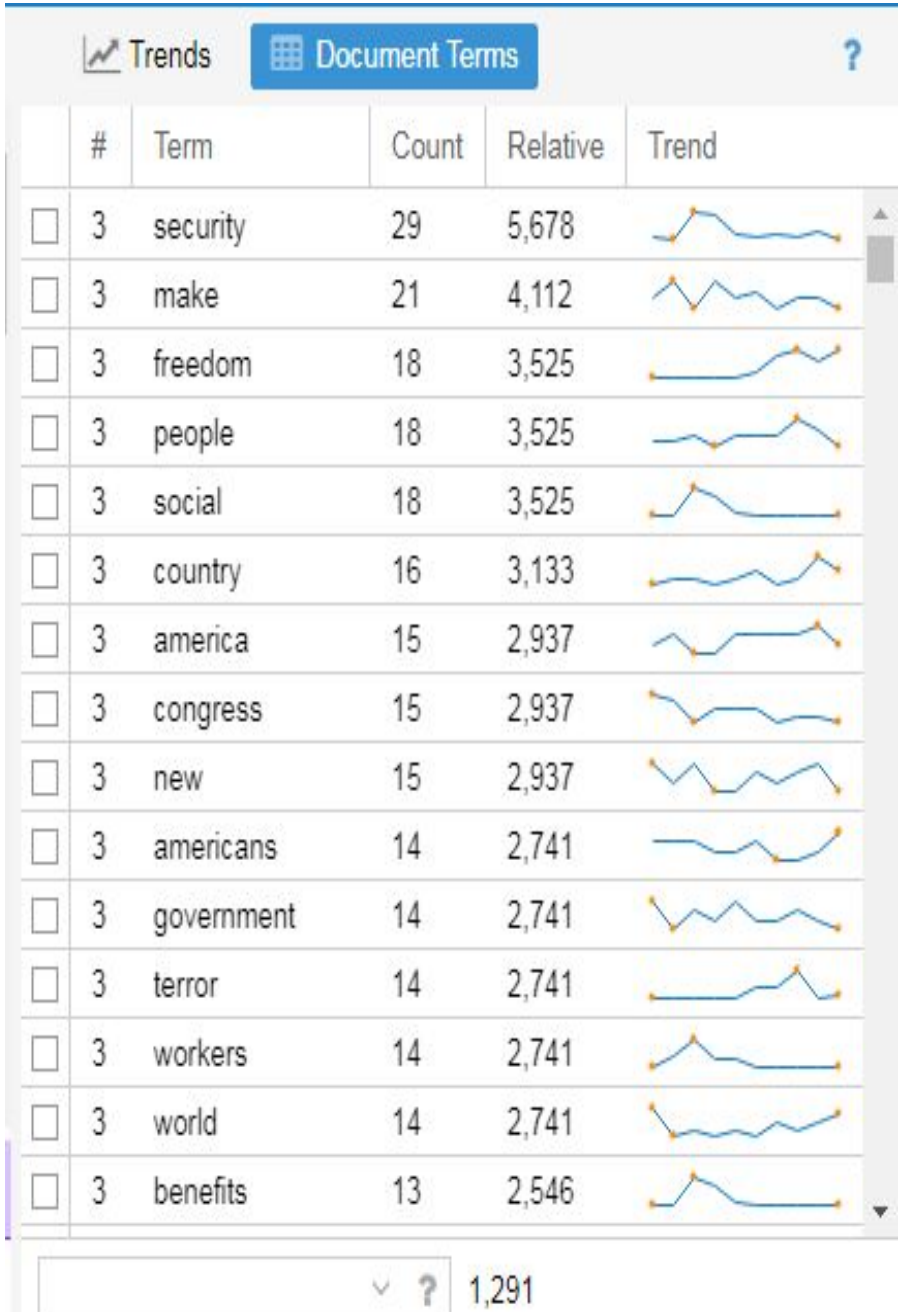
- Their take:

Word cloud - Bush in 2005

- What do you see?

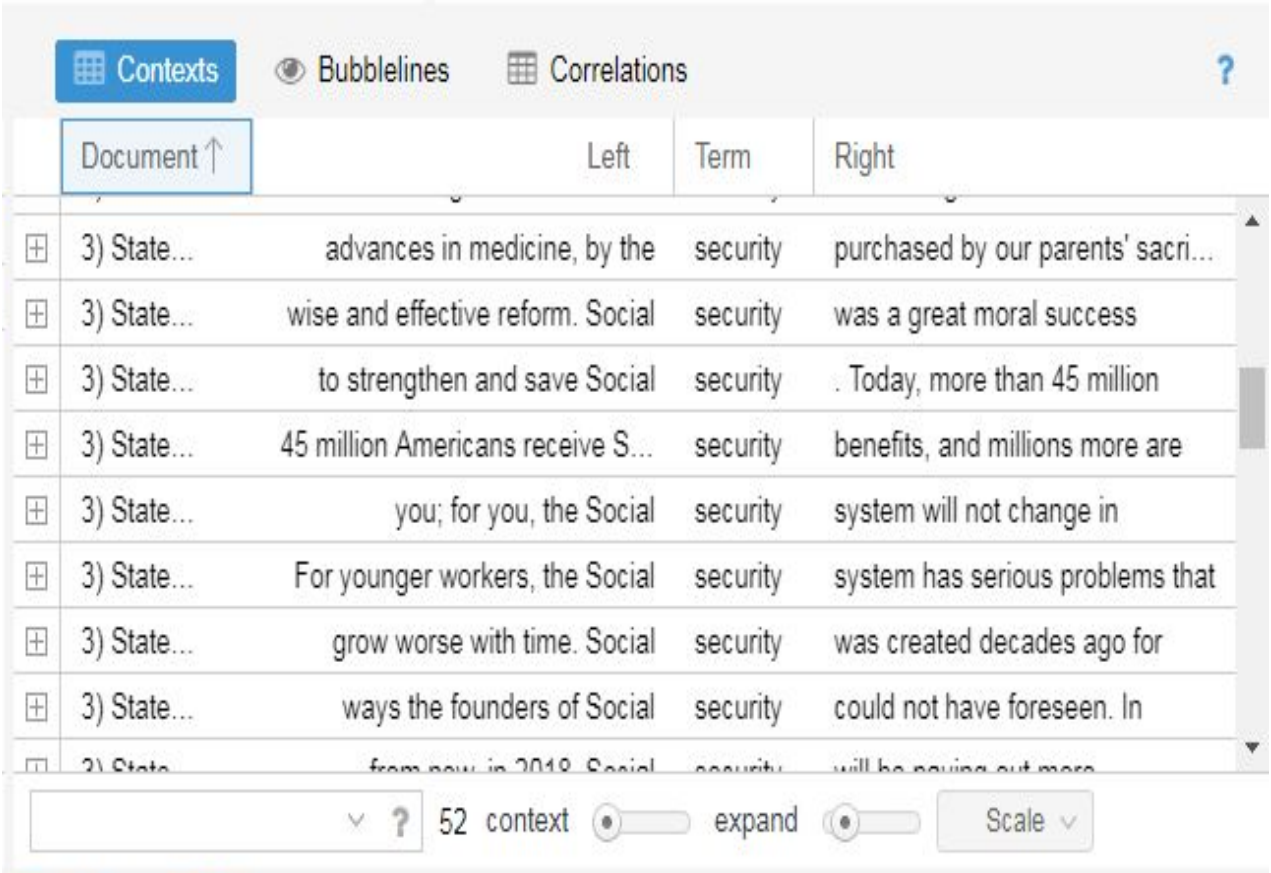
Top terms - Bush

- What do you see?



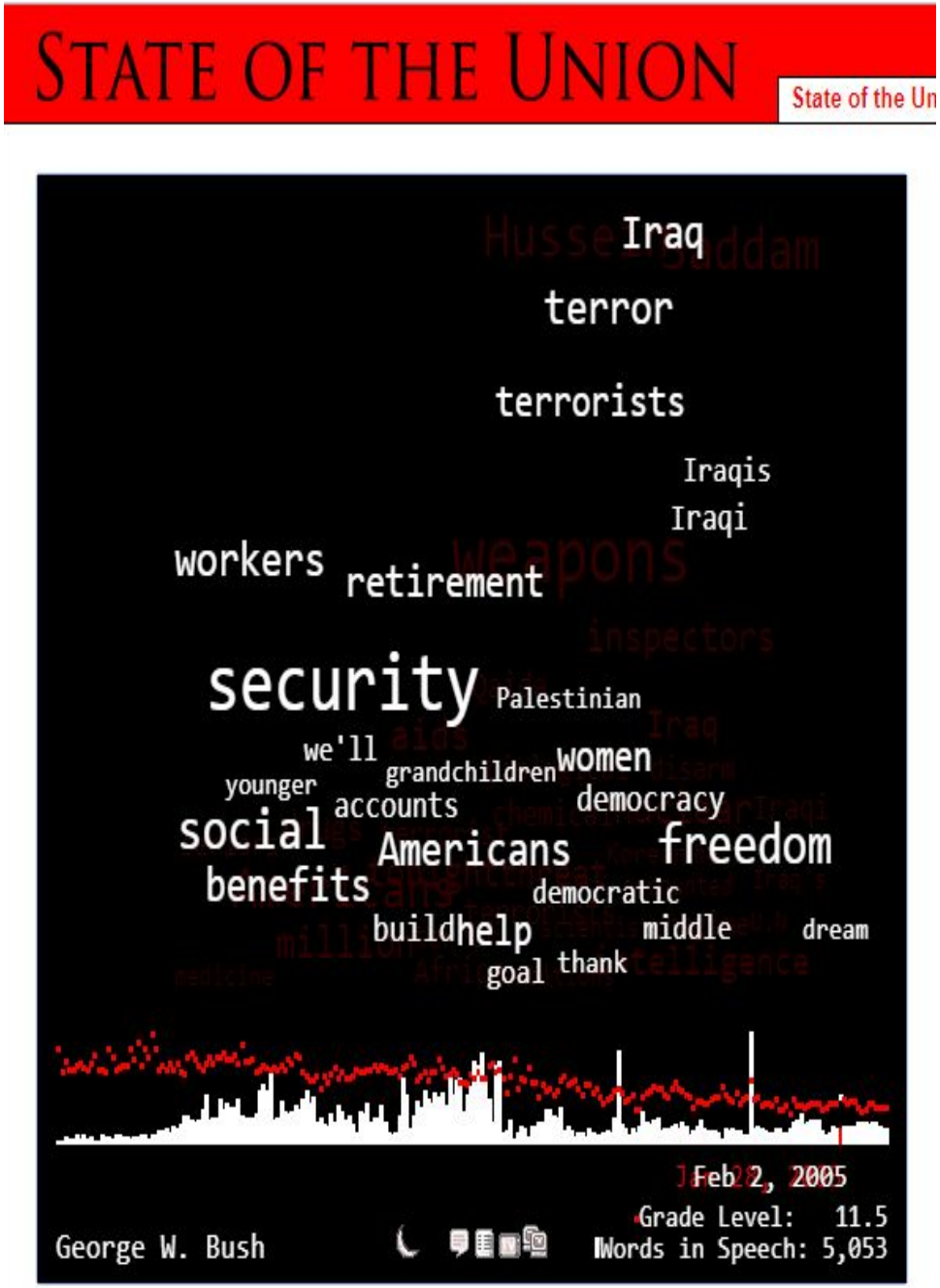
“Security” in context for Bush

- What you get:



Bush in 2005

- What they display:

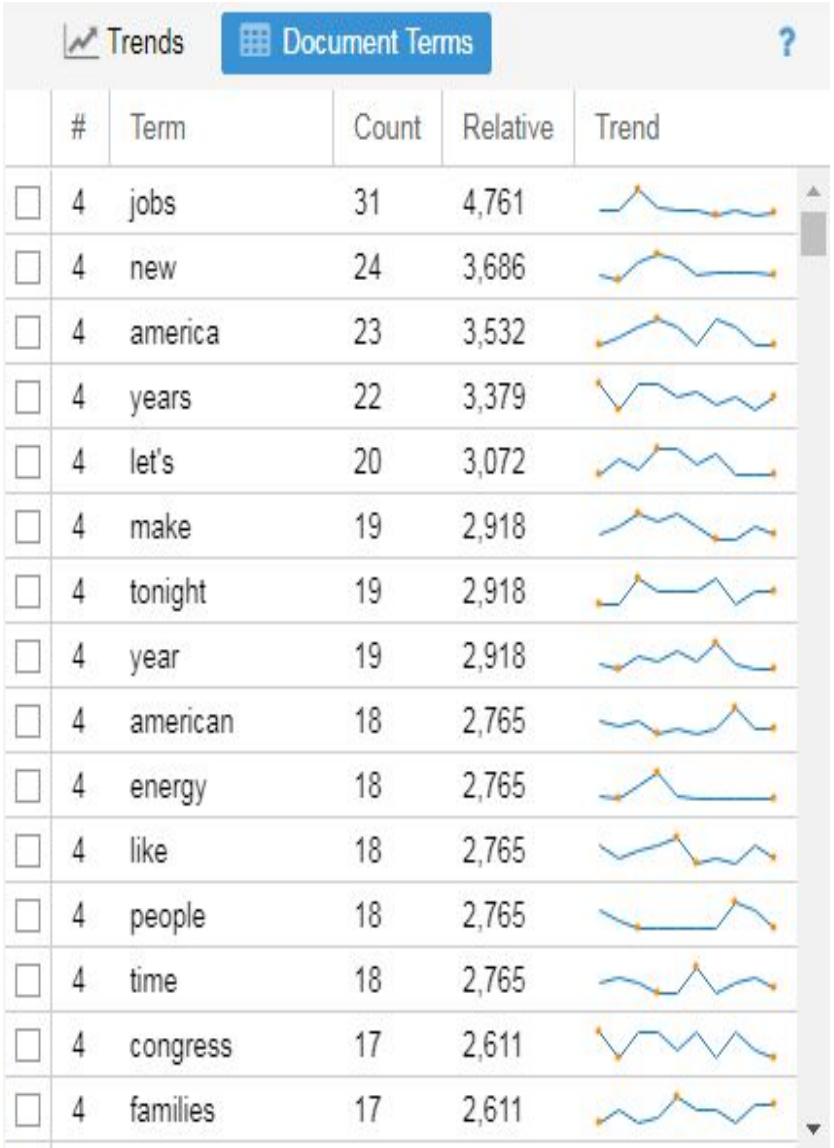


Word cloud - Obama in 2013

- What you get:

Top terms - Obama in 2013

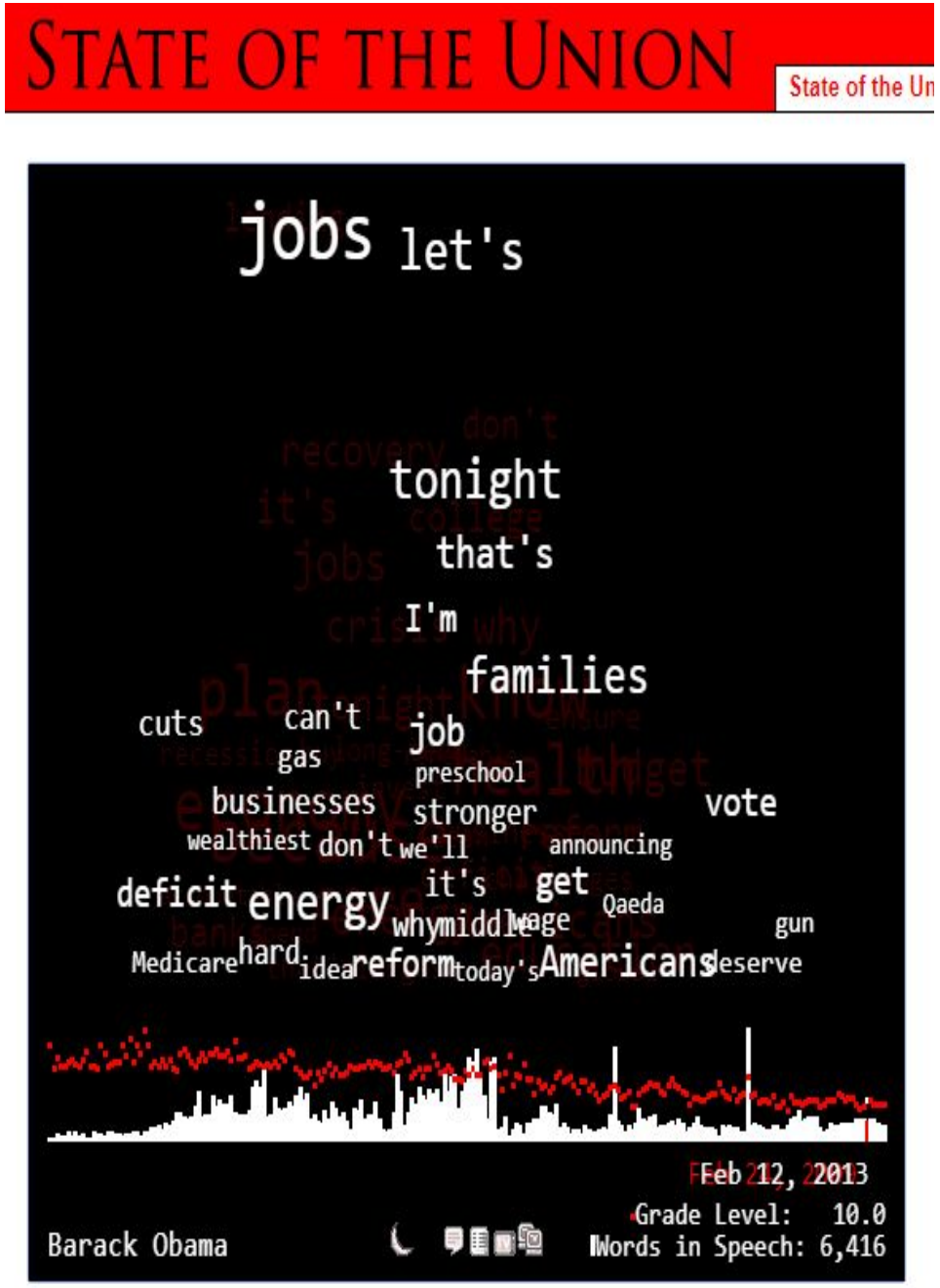
- What do you see?



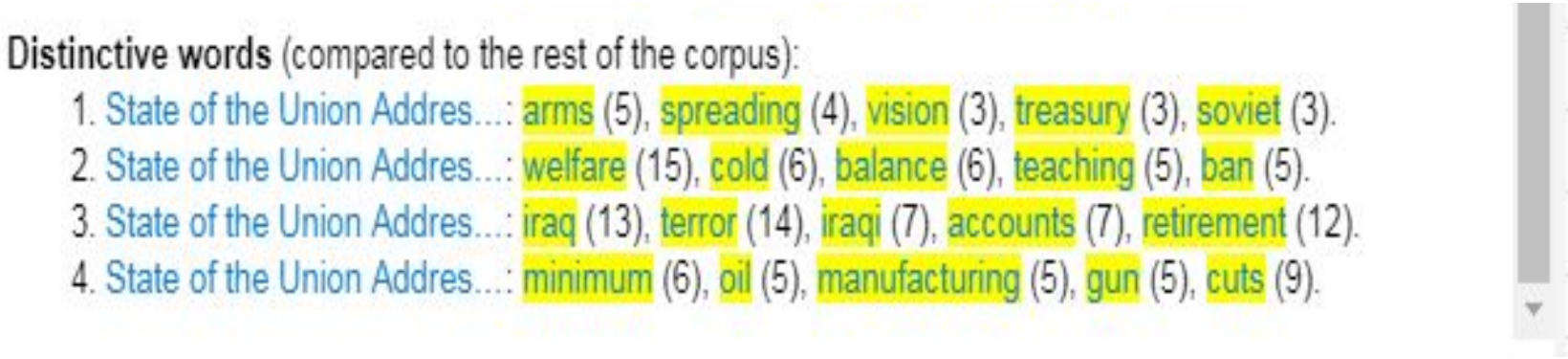
	#	Term	Count	Relative	Trend
☐	4	jobs	31	4,761	
☐	4	new	24	3,686	
☐	4	america	23	3,532	
☐	4	years	22	3,379	
☐	4	let's	20	3,072	
☐	4	make	19	2,918	
☐	4	tonight	19	2,918	
☐	4	year	19	2,918	
☐	4	american	18	2,765	
☐	4	energy	18	2,765	
☐	4	like	18	2,765	
☐	4	people	18	2,765	
☐	4	time	18	2,765	
☐	4	congress	17	2,611	
☐	4	families	17	2,611	

Obama in 2013

- What they draw:



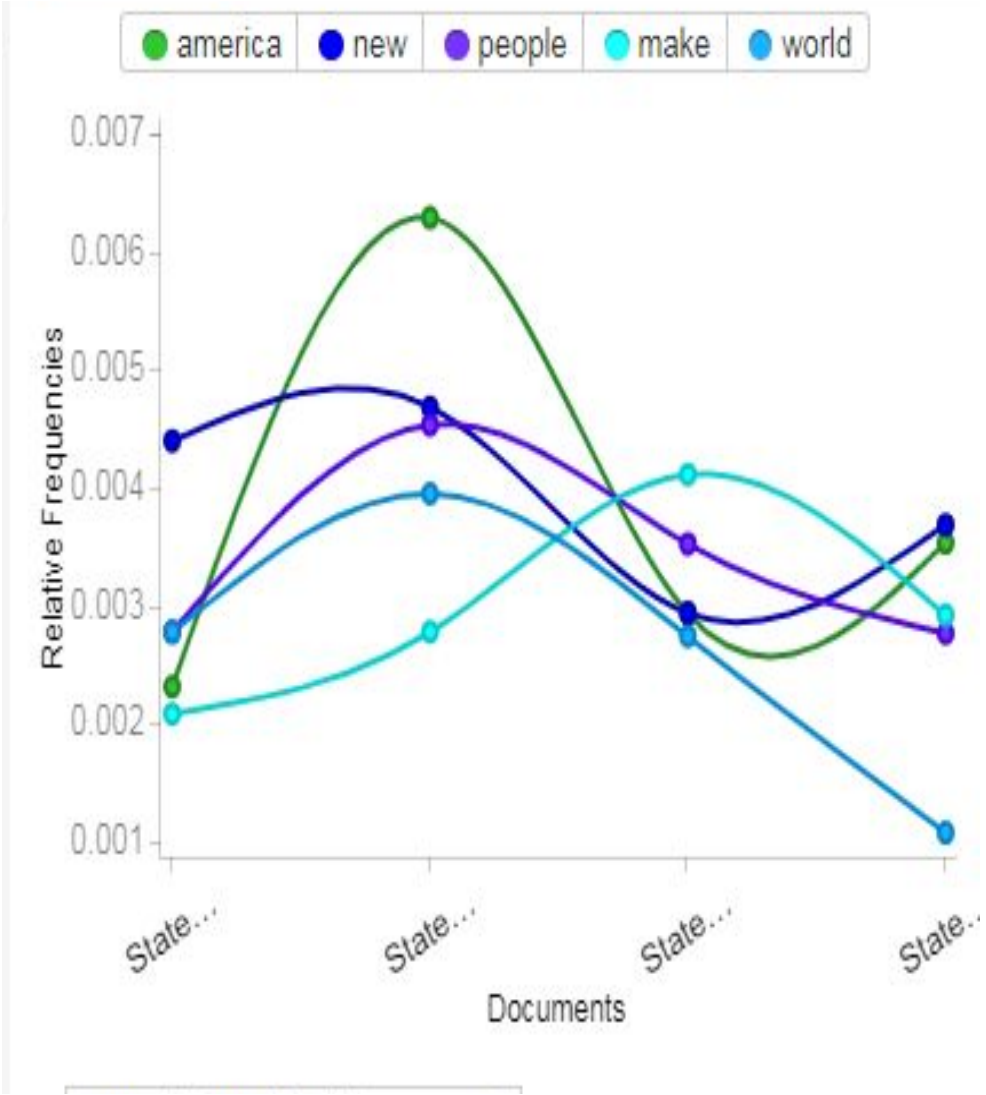
Most unique words by each president



- Some of these words are good “give-aways” of who might be speaking …
- Based on TF-IDF

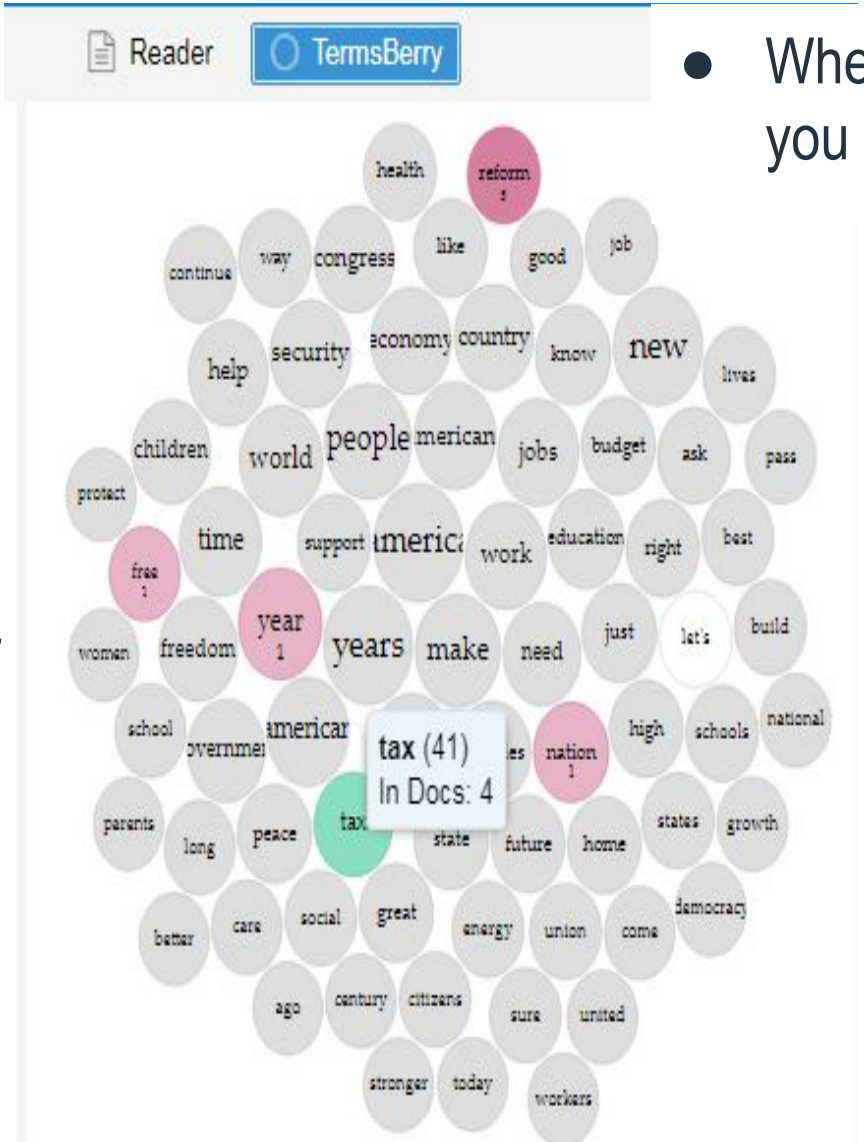
Trends

- Any partisan patterns?
- Other patterns?



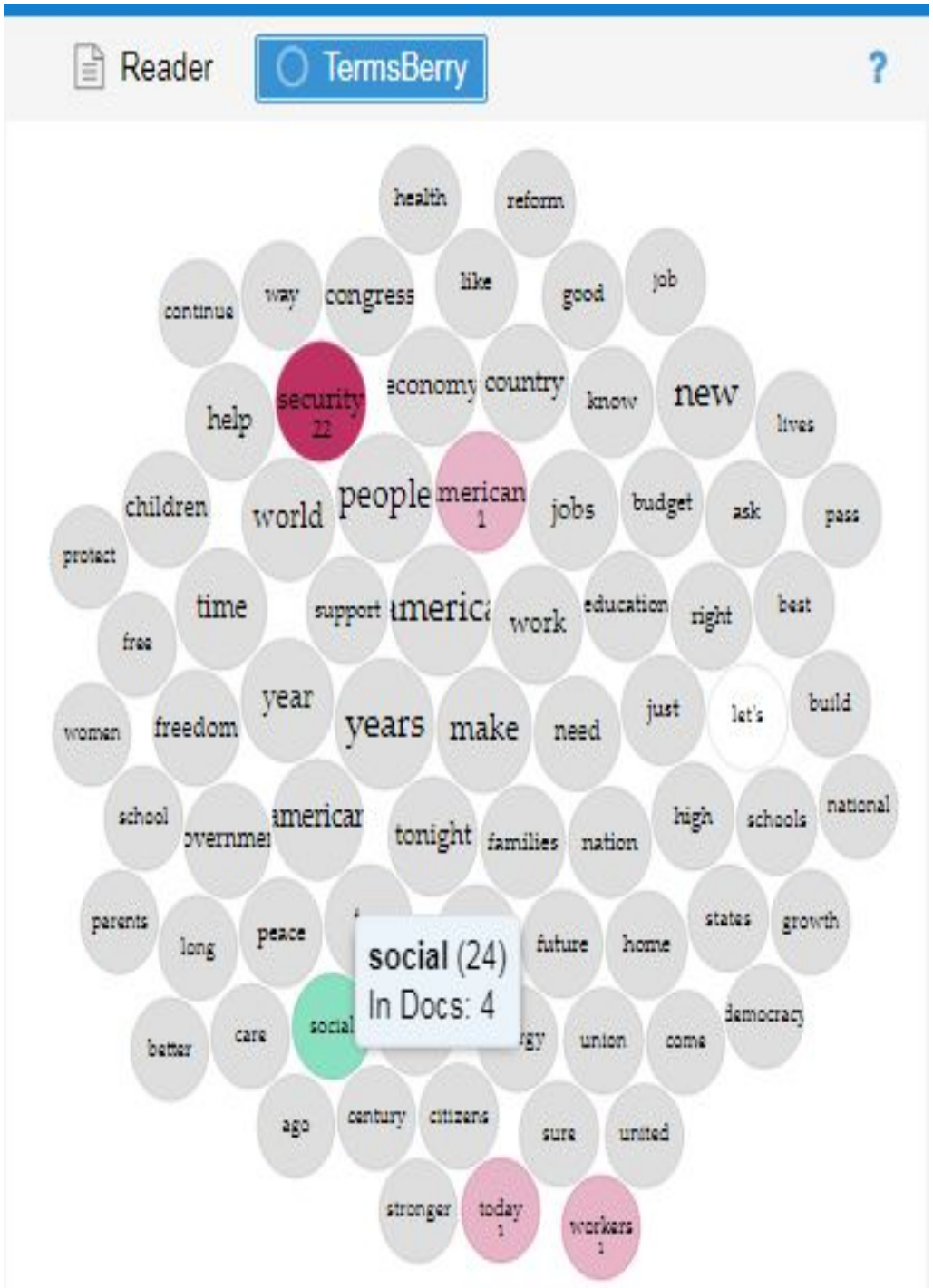
Seeing how words connect

- The TermsBerry tool explores high frequency terms and their collocates (words that occur in proximity).
- The darkness of the terms represents the proportion of the documents where the term appears.

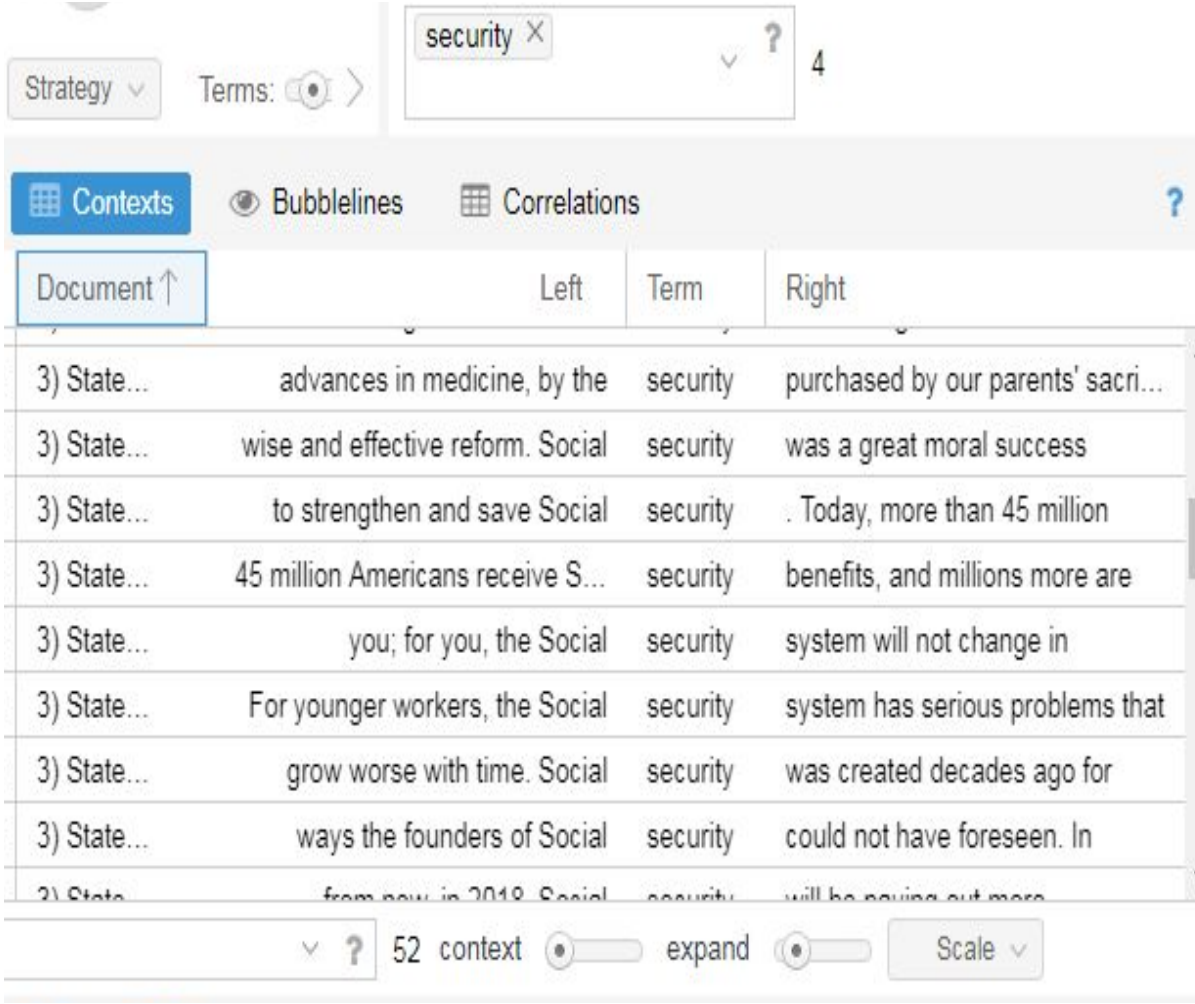


- When I say tax, you say ...

When I say “social,” you say what?



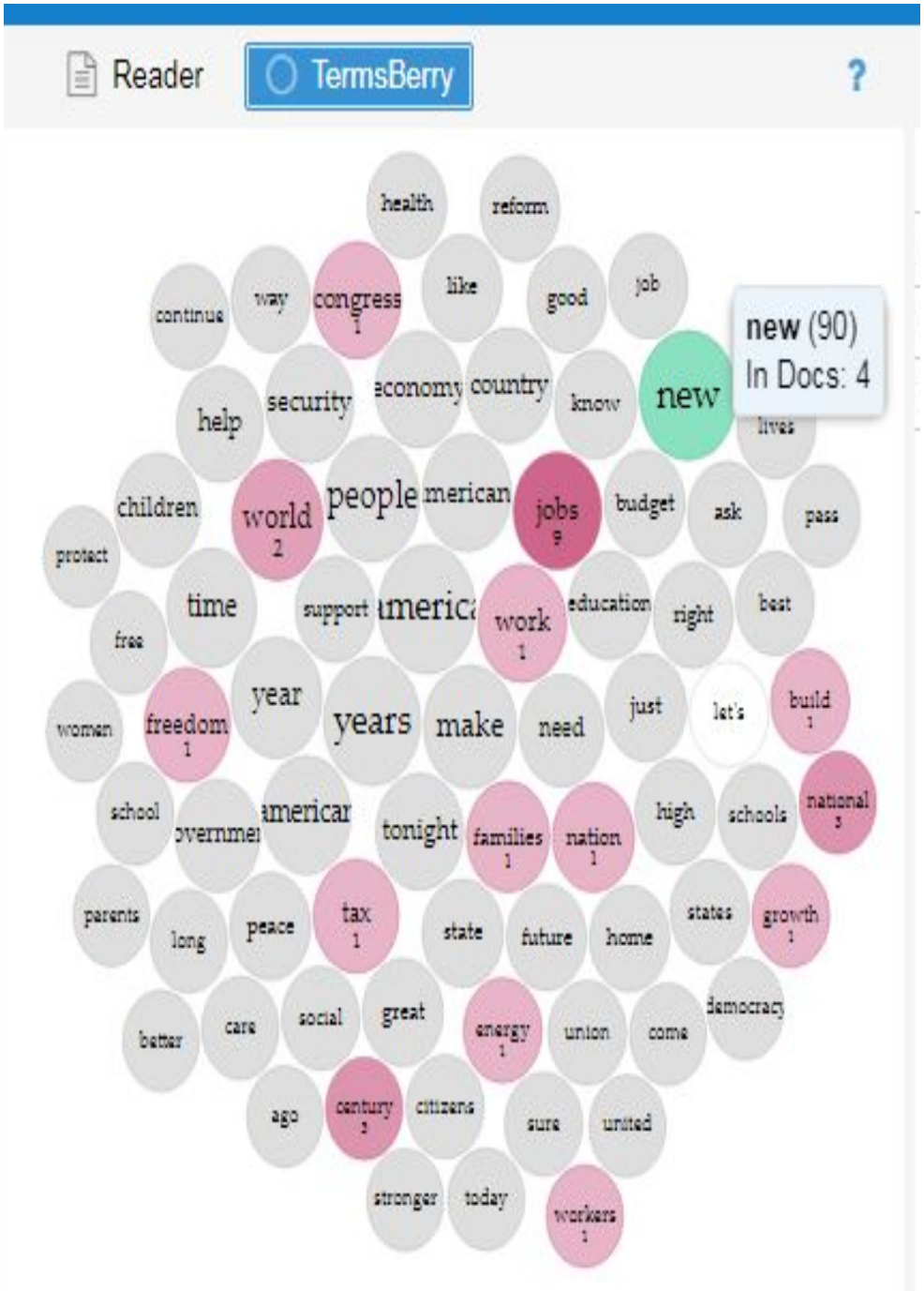
Finding “security” in context



- It’s almost always preceded by “social”

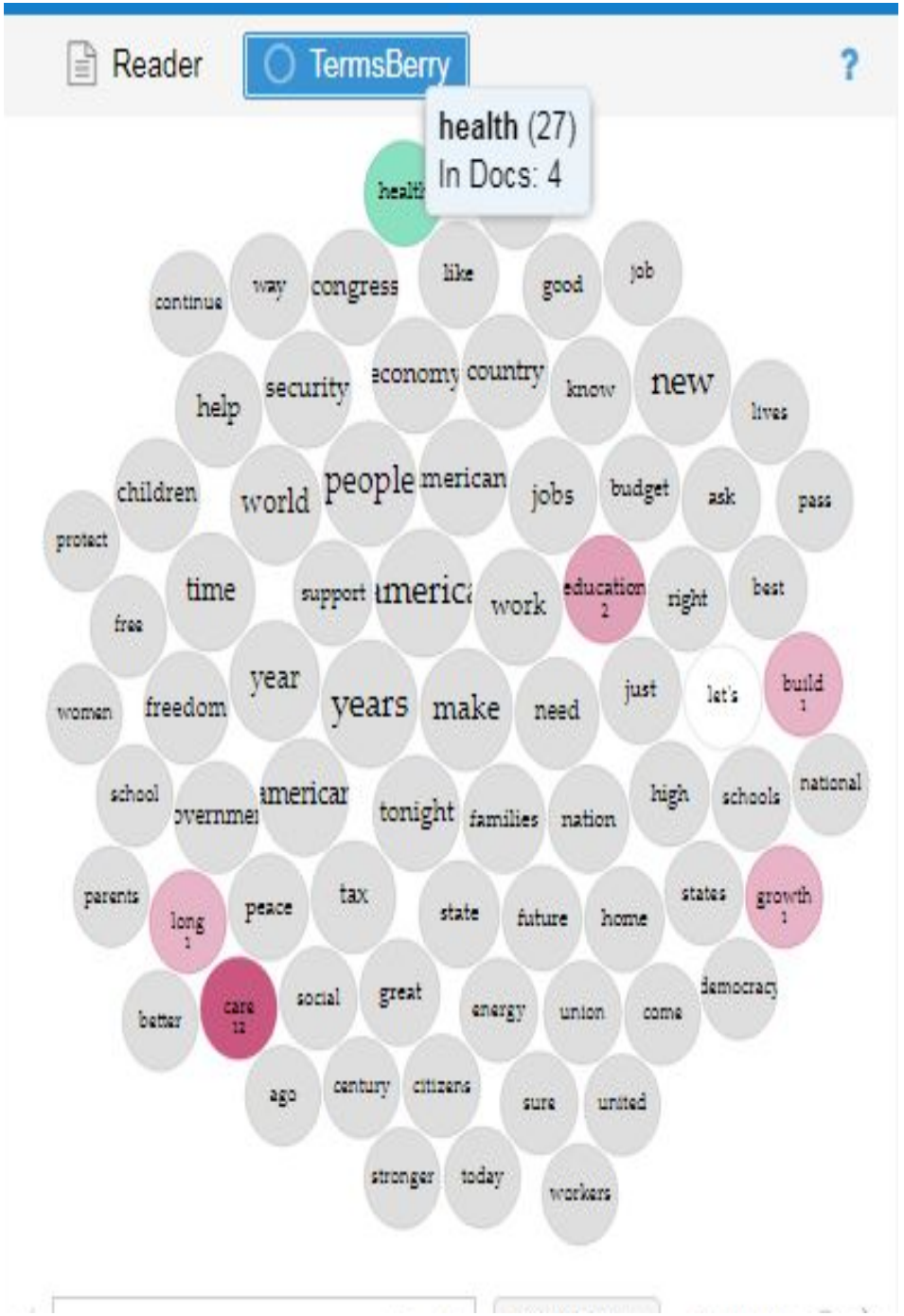
New what?

- … jobs



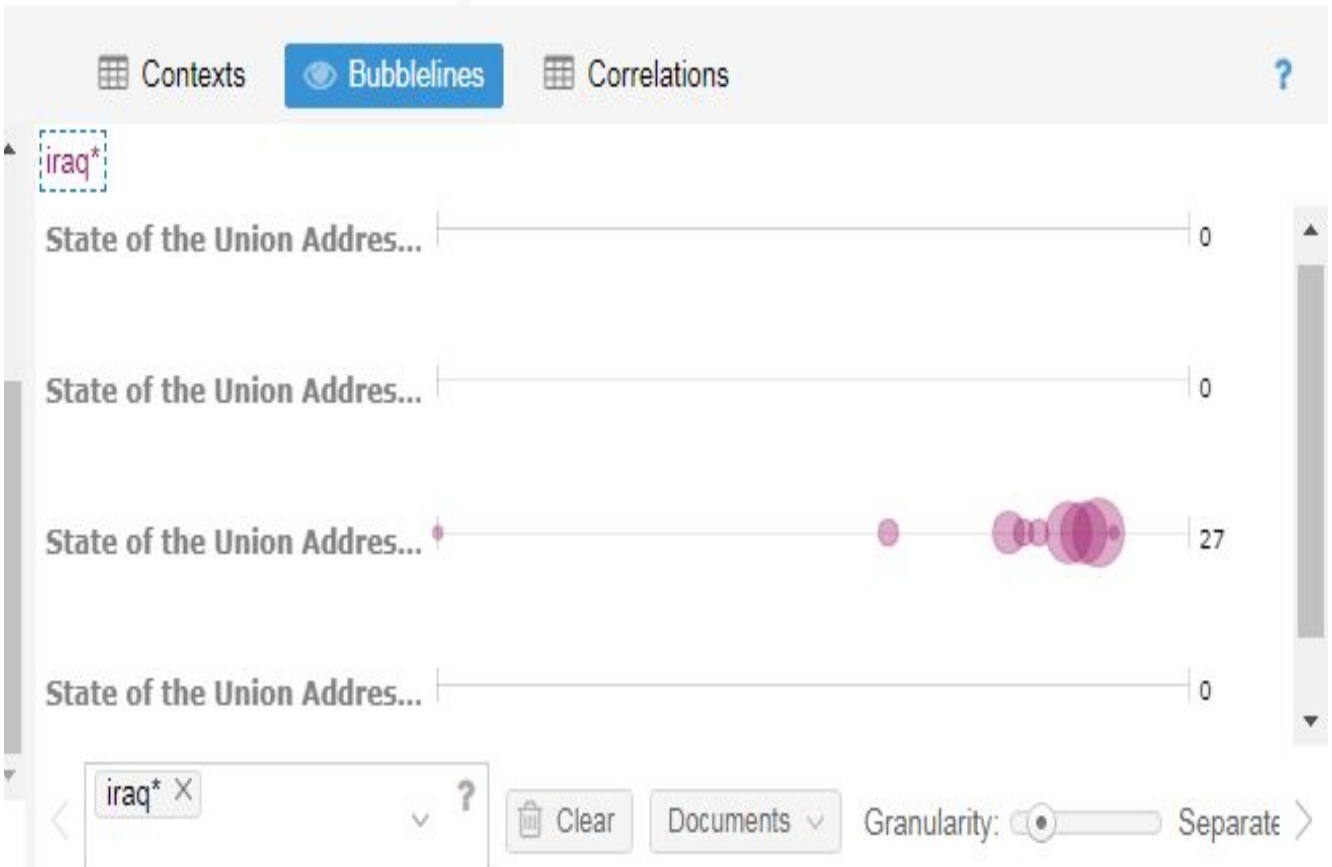
When I say "health," you say ...

- … care



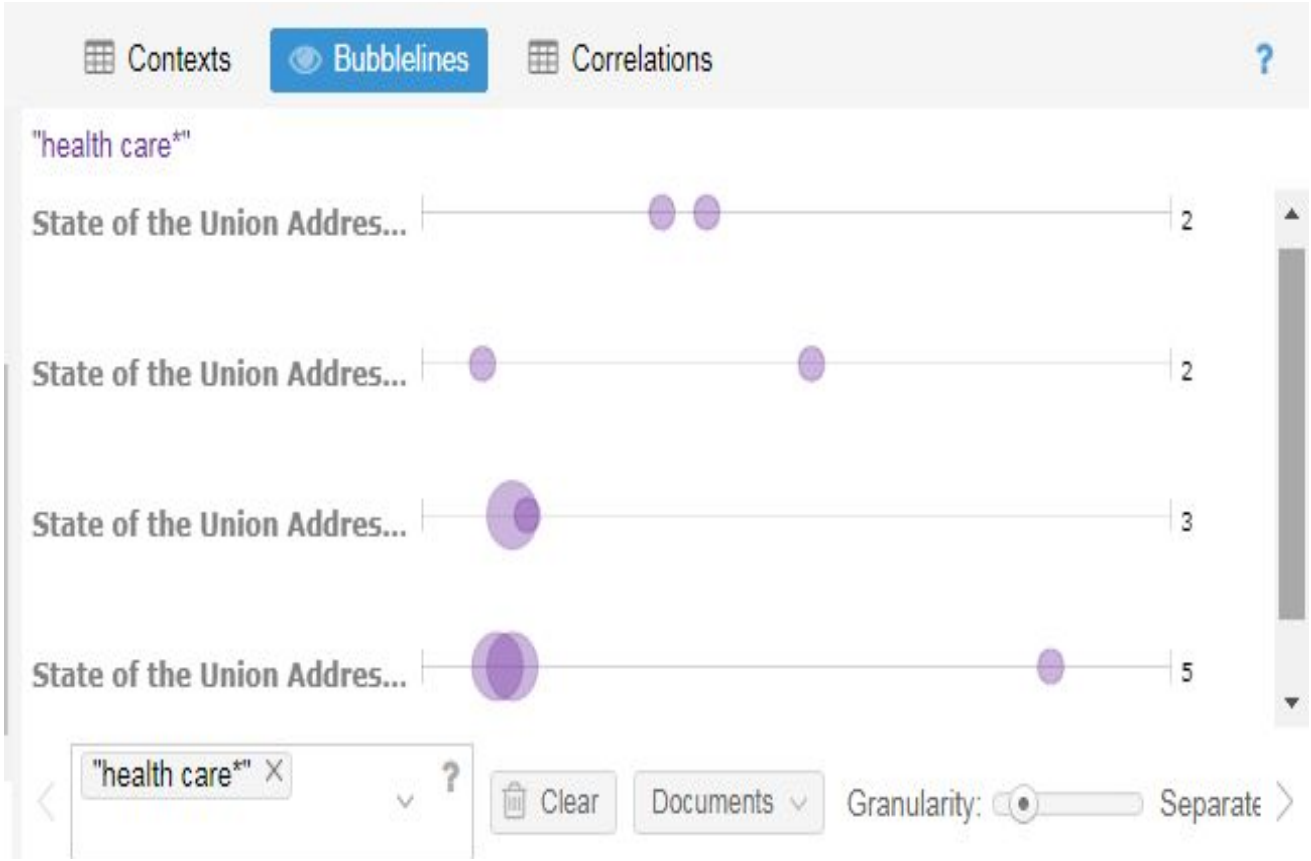
When is Iraq* mentioned?

- Bush mentioned it 27 times, mostly in the closing of his speech



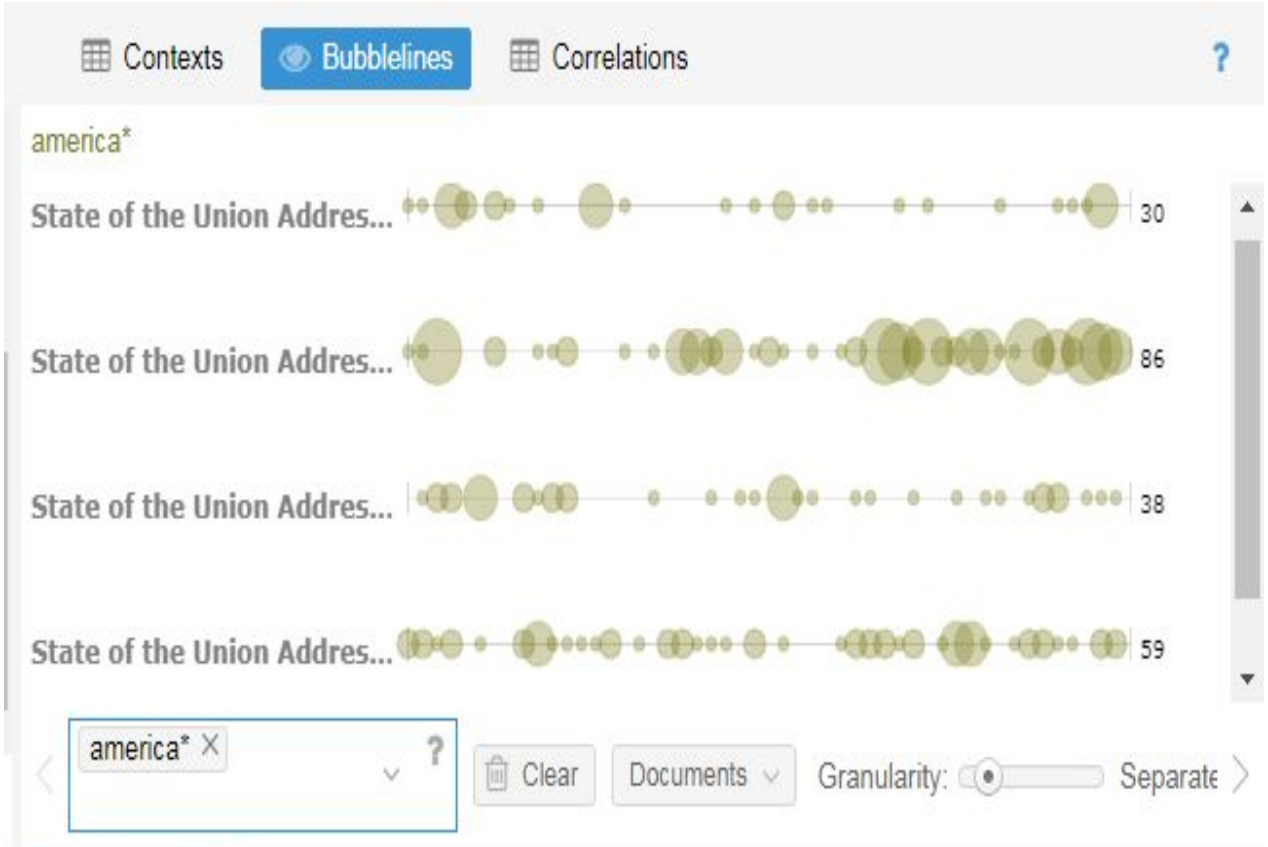
“Health care” smattered in

- Every president has to mention health care...



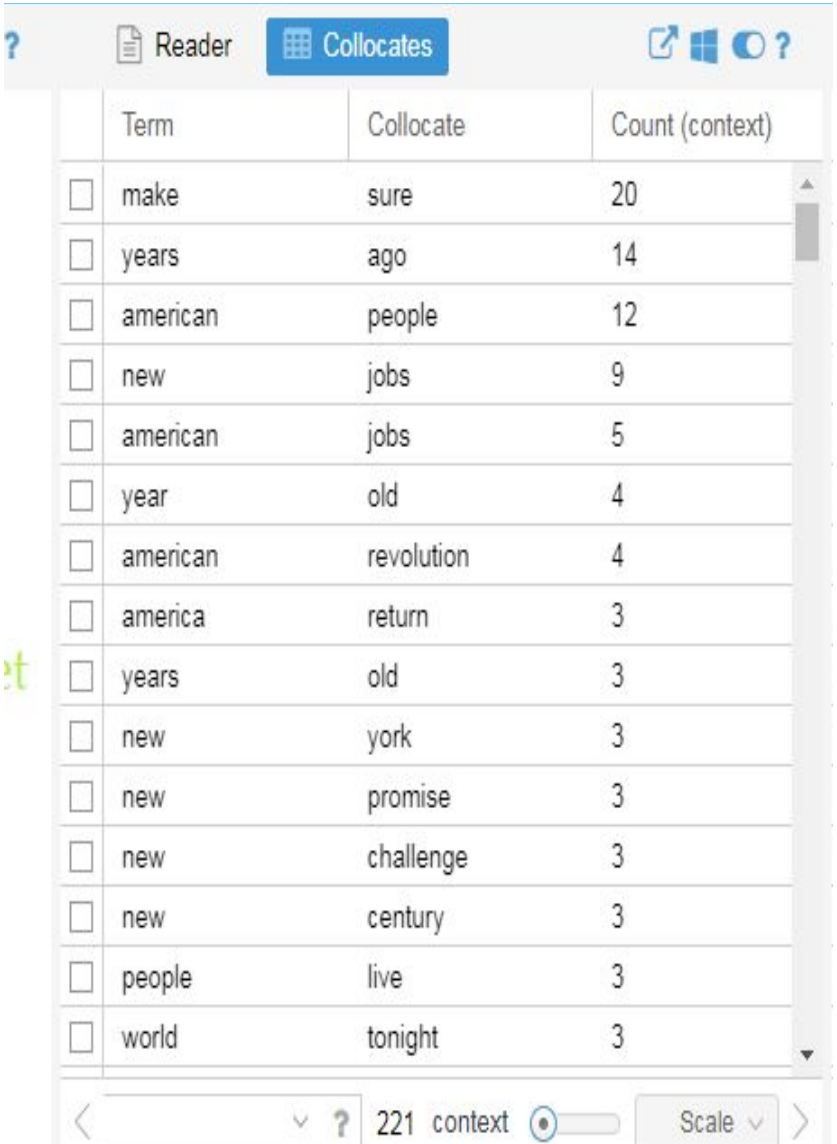
“America*” is everywhere...

- Clinton’s bubbles are the biggest on this though (n=86)

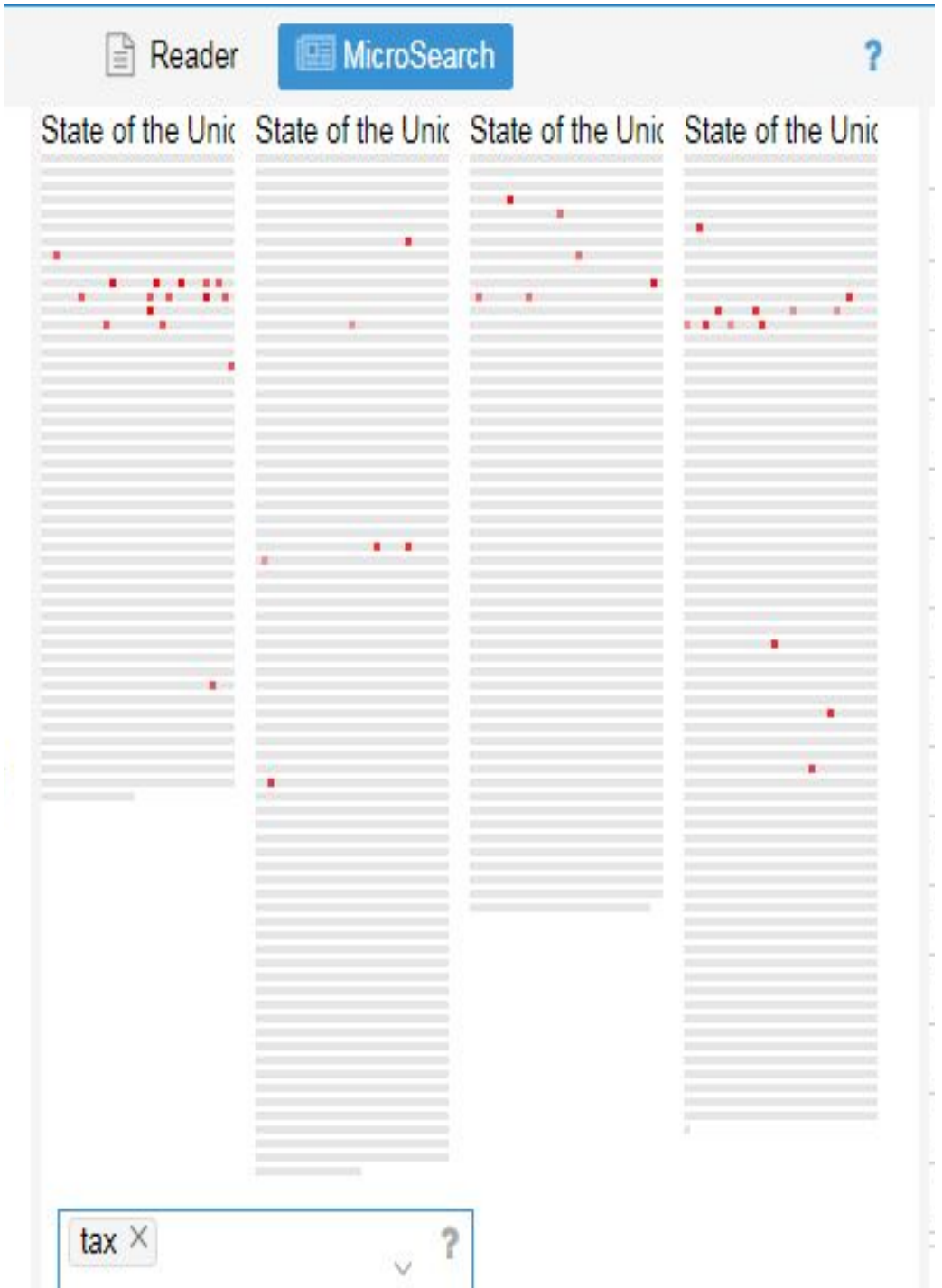


The most common collated tokens ...

- These words are often collocated

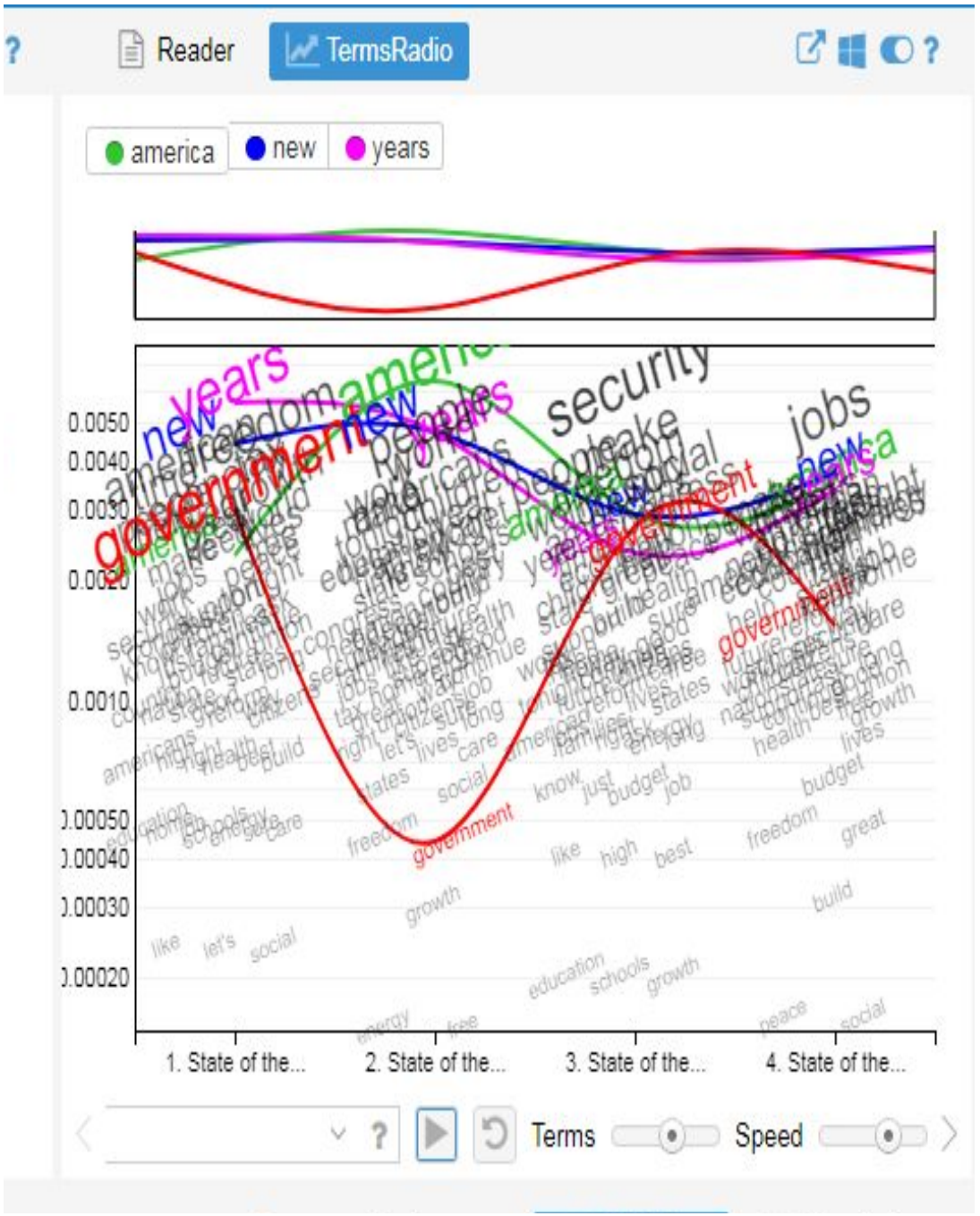


Where is tax mentioned in each speech?



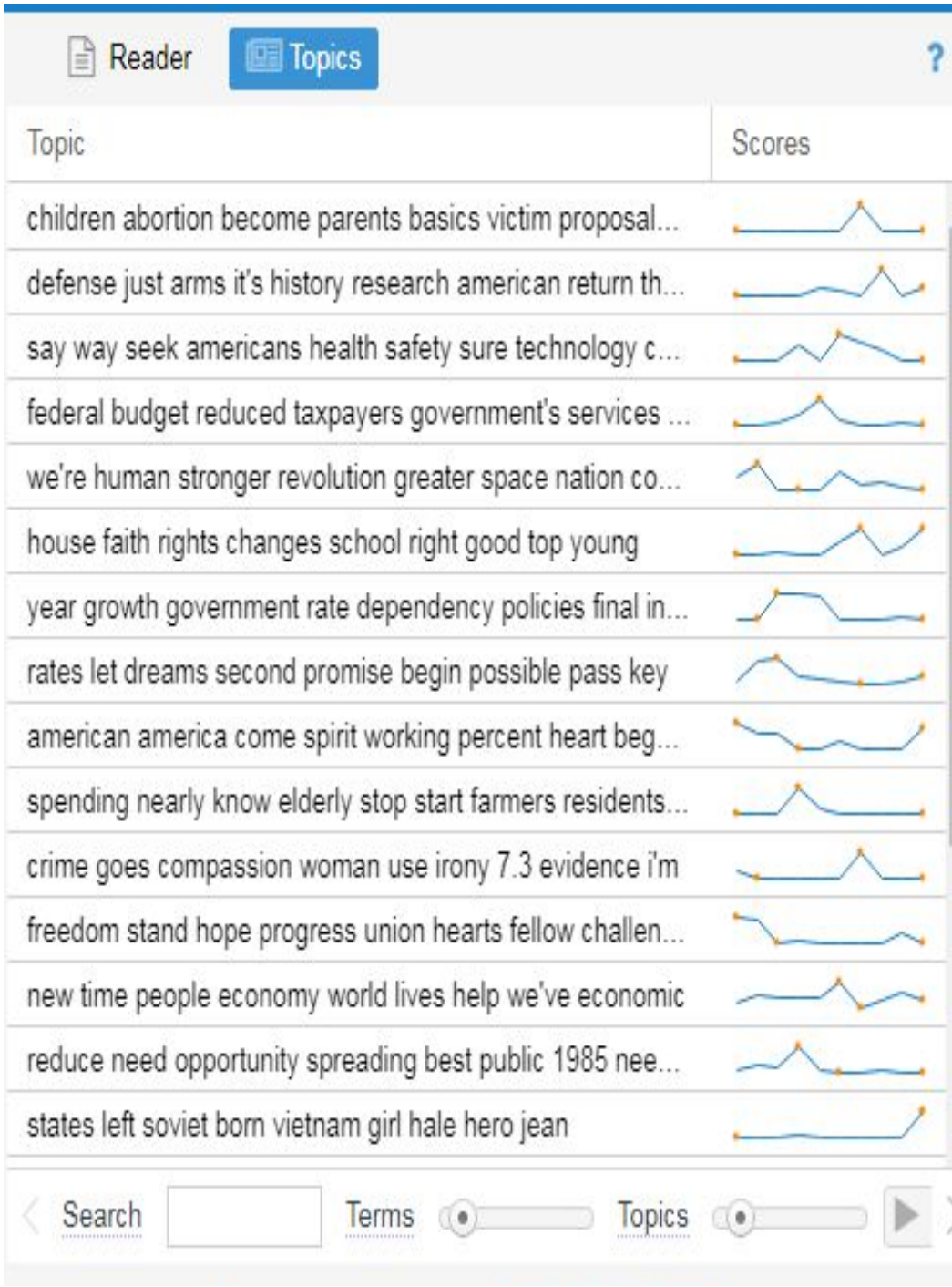
Terms radio-

- Tracking use of words over speeches
- “Government” is talked about more by Republicans than Democrats

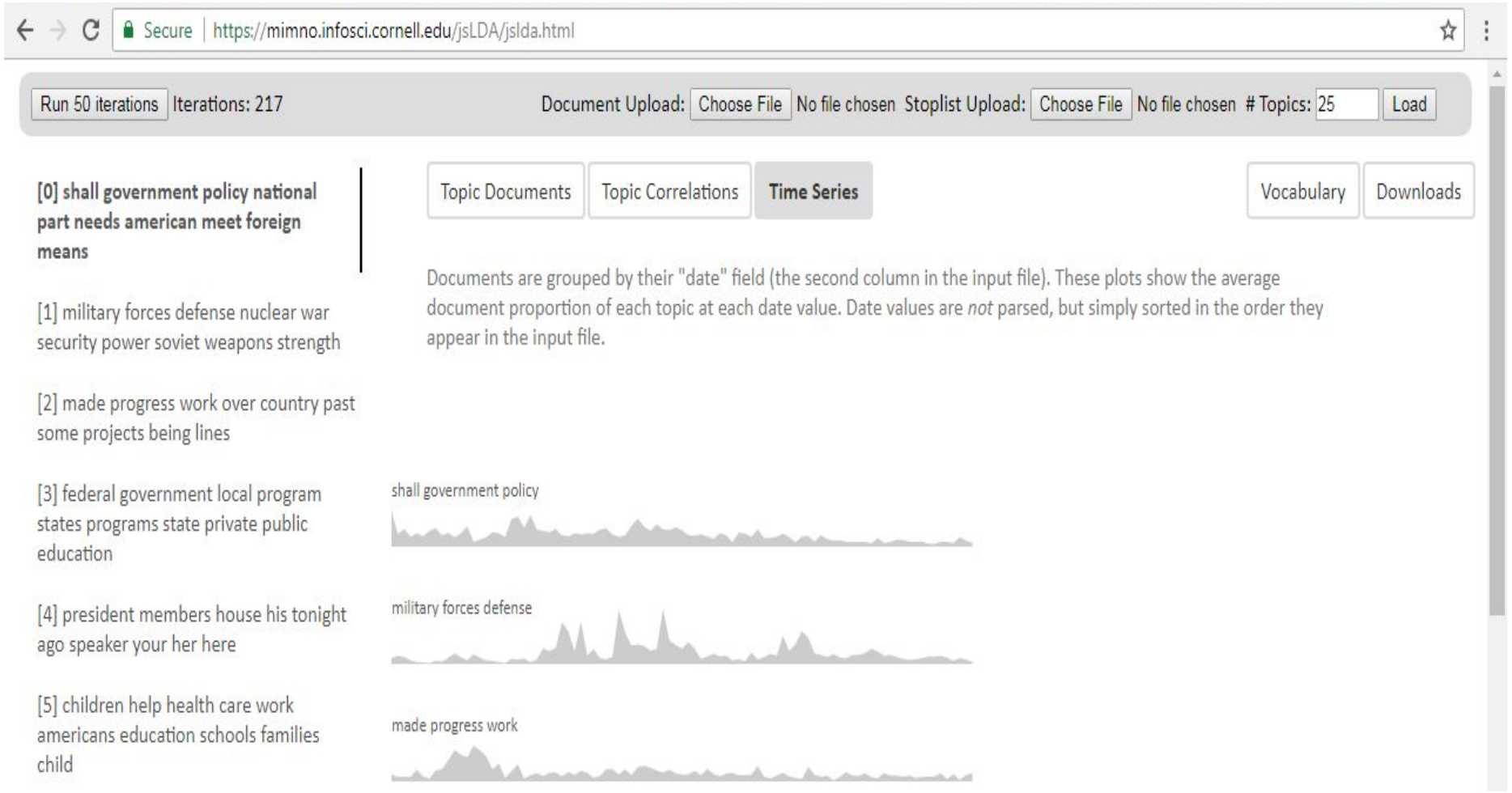


Topic modeling

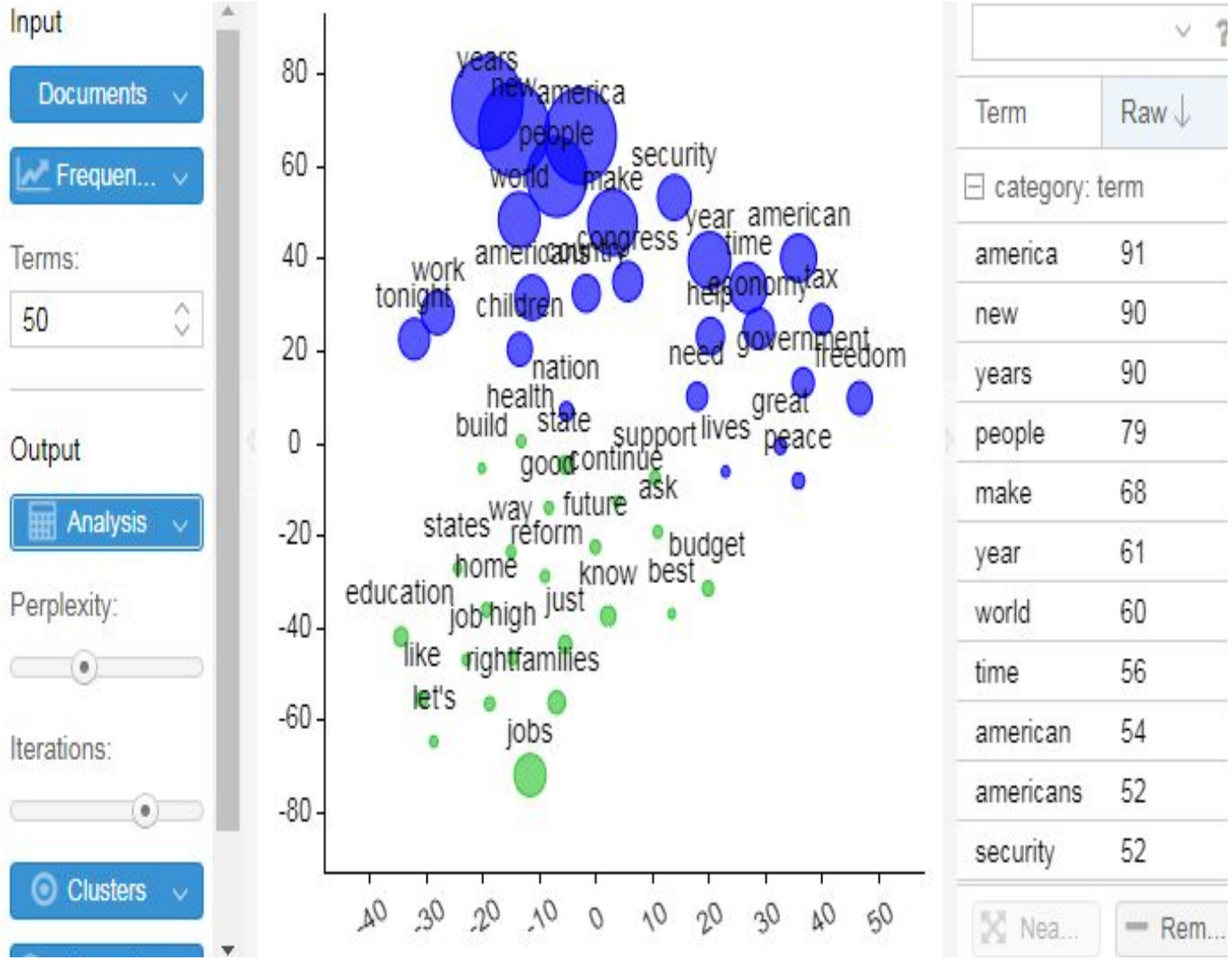
- 22 topics chosen.
- If you have multiple documents then the topic modelling is performed on each document.
- By default Topics only uses the first 1,000 words in a document, which you can increase, but be wary.



The Topics algorithm comes from here:



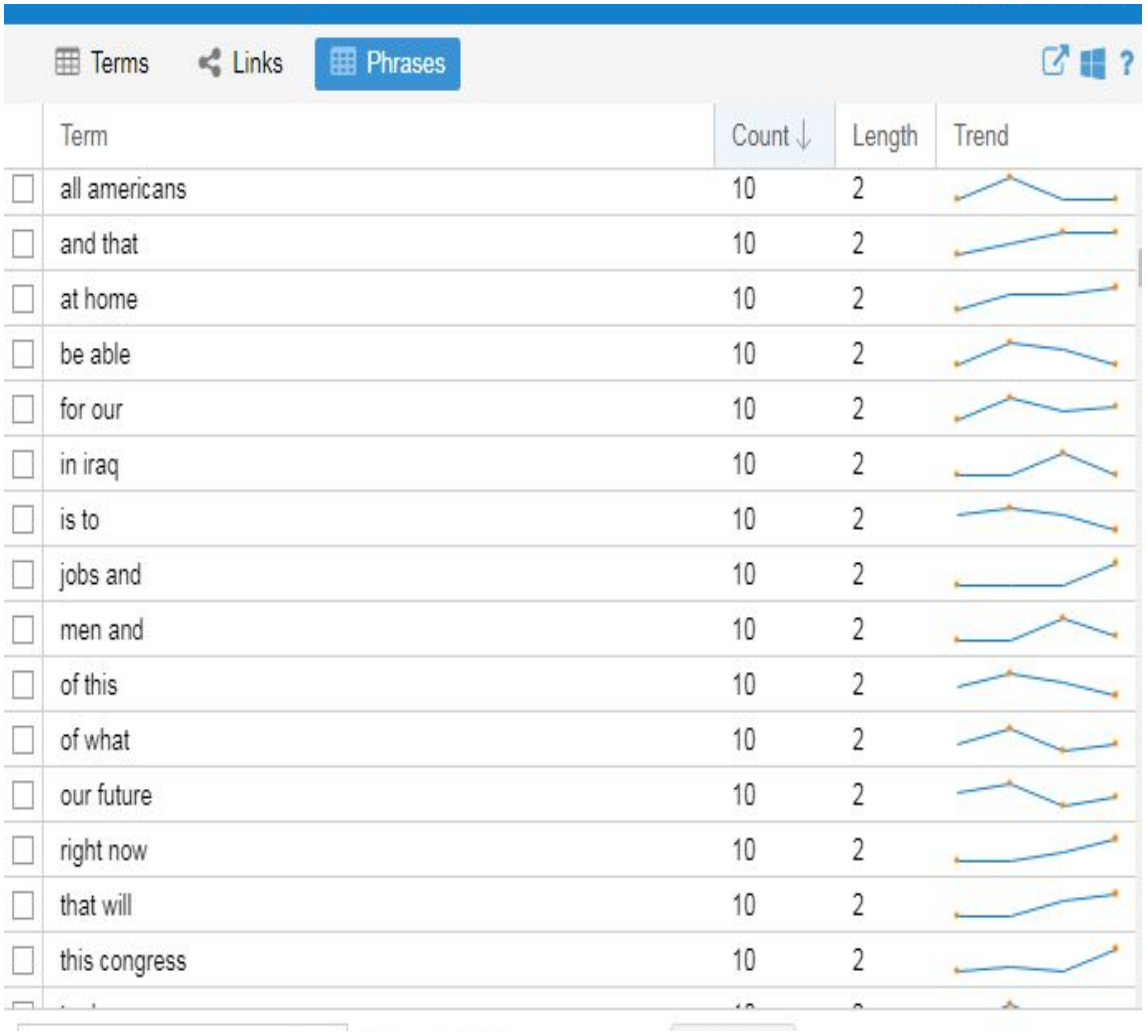
t-SNE



- t-SNE is a form of similarity analysis

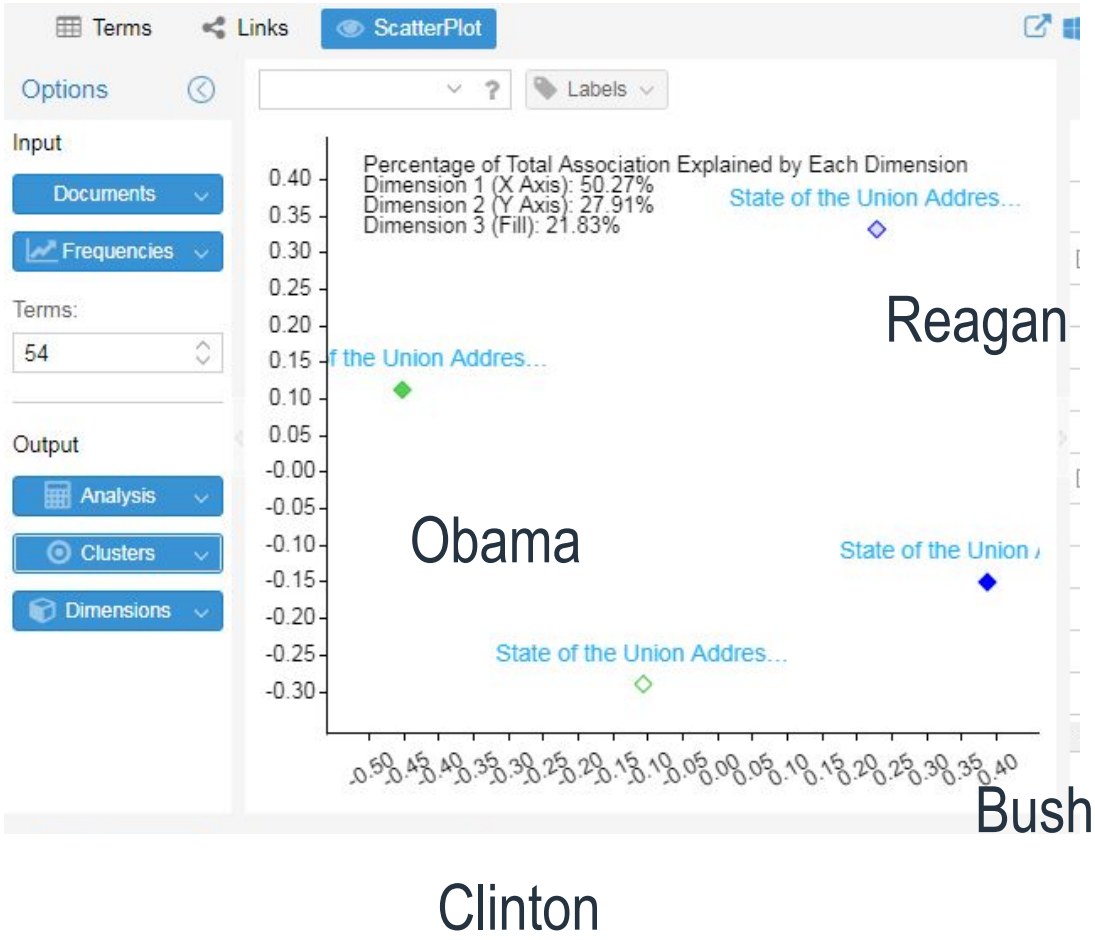
Bigrams

- Common bigrams...



Term	Count	Length	Trend
all americans	10	2	
and that	10	2	
at home	10	2	
be able	10	2	
for our	10	2	
in iraq	10	2	
is to	10	2	
jobs and	10	2	
men and	10	2	
of this	10	2	
of what	10	2	
our future	10	2	
right now	10	2	
that will	10	2	
this congress	10	2	

Document similarity



- Seems to go from liberal to conservative along the X-axis (which captures most of the variance) - and I don't know what is going on with the Y-axis. Thoughts?

Big conclusions from all these tools

- All SOUs talk a lot about America, Americans, American people, things that have happened or we would like to happen, as well as things happening in the world

- Sometimes, the most frequently used words are pretty good summary of the messages, and sometimes maybe not so much.

Can we summarize the SOUs?

- Reagan's speech was about *freedom* (and *taxes*)? - Speech #1
- Clinton's speech was about *new* stuff? - Speech #2
- Bush's speech was about *Social Security*? - Speech #3
- Obama's speech was about *jobs*? - Speech #4

#	Term	Count	Relative	Trend
2	america	43	6,287	
3	security	29	5,678	
1	years	24	5,553	
4	jobs	31	4,761	
2	new	32	4,678	
2	years	32	4,678	
2	people	31	4,532	
1	freedom	19	4,396	
1	new	19	4,396	
3	make	21	4,112	
2	world	27	3,947	
1	american	16	3,702	
1	tax	16	3,702	
4	new	24	3,686	
4	america	23	3,532	

How'd we do?

- Reagan 1985
- "Opportunity"?
- Hmmm?

REAGAN SKETCHES LEGISLATIVE GOALS FOR NEXT 4 YEARS

'2D AMERICAN REVOLUTION'

Agenda Includes Tax Revision, Economic Growth and Cut in Nuclear War Threat

WASHINGTON, Feb. 6 — President Reagan urged the nation today to forge "a second American revolution of hope and opportunity" with his agenda for the next four years of tax revision and economic growth and the elimination of the threat of nuclear war.

In a State of the Union Message to Congress designed to set the legislative

sounded the same buoyant note that marked his triumphal Presidential campaign last year. He said:

"The time has come to proceed toward a great new challenge, a second American revolution of hope and opportunity; a revolution carrying us to new heights of progress by pushing back frontiers of knowledge and space; a revolution of spirit that taps the soul of America, enabling us to summon greater strength that we've ever known, and a revolution that carries beyond our shores the gold promise of human freedom in a world at peace."

'Opportunity' Is Theme

Dominating his speech was the

Did Reagan say “opportunity” a lot?

#	Term	Count	Relative	Trend
1	opport*	7	1,620	
4	opport*	7	1,075	
2	opport*	4	585	
3	opport*	2	392	

- He did say opportunity the most of any president in our sample, but ...

- … he said a lot of other words a lot more often, like freedom and tax

#	Term	Count	Relative	Trend
1	years	24	5,553	
1	freedom	19	4,396	
1	new	19	4,396	
1	american	16	3,702	
1	tax	16	3,702	

How was he talking about it?

- “Opportunity” in context

	D…	Left	Term	Right
⊞	1) …	American Revolution of hope and	opportunity	; a revolution carrying us to
⊞	1) …	for capital formation. To encourage	opportunity	and jobs rather than dependency
⊞	1) …	must carry the promise of	opportunity	for all. It is time
⊞	1) …	to pass our youth employment	opportunity	wage proposal. We can help
⊞	1) …	public housing residents have that	opportunity	of ownership. The Federal Government
⊞	1) …	spreading dependency and start sprea…	opportunity	; that we will stop spreading
⊞	1) …	manned space station and new	opportunit…	for free enterprise, because in

Can we summarize the SOUs?

- Reagan's speech was about *freedom* (and *taxes*)? - Speech #1

Sort of. Maybe.

How'd we do?

- **Clinton in 1997**
- Feels like a laundry list...

U.S. | THE SECOND TERM: THE SCENE

On a Split-Screen Night, Clinton Gets the Full Attention of Congress

By FRANCIS X. CLINES FEB. 5, 1997

The Capitol scene was a study in acute distractedness tonight as the television monitors of the nation's lawmakers flashed away from California pictures of the climax of another O. J. Simpson trial even as President Clinton sought to pronounce on the State of the Union.

In the traditional Capitol gantlet of television interviews, one black lawmaker was asked about the Simpson trial even before the Clinton agenda. It appeared to take extra devotion to public service to sit still in the chamber for Mr. Clinton's speech on a split-screen night.

But the President soon took hold with his speech. And as he touched on the subject of campaign financing, the chamber took on the aspect of a surgery theater packed with tiers of ranking specialists warily attending the latest radical procedure being prescribed for their common ailment. If there was a certain air of wincing in the audience, it was for the talk of "tough" self-surgery, the prescription offered by the nation's ranking political practitioner.

President Clinton was delivering a warning to himself as much as to the

gletagmanager.com...

Can we summarize the SOUs?

- Clinton's speech was about *new* stuff? - Speech #2

Maybe, probably. Laundry list?

How'd we do?

- **Bush in 2005**
- "Social Security"

POLITICS | STATE OF THE UNION: THE OVERVIEW

In Speech, Bush Sketches a Bold Domestic and Foreign Agenda

By RICHARD W. STEVENSON and DAVID E. SANGER FEB. 3, 2005

WASHINGTON, Feb. 2 - President Bush challenged a wary Congress on Wednesday night to join him in reinventing Social Security for the 21st century, saying his generation had a duty to preserve the retirement system for those who follow and for the first time setting out details of the individual investment accounts at the heart of his proposal.

"Social Security was a great moral success of the 20th century, and we must honor its great purposes in this new century," he said. "The system, however, on its current path, is headed toward bankruptcy. And so we must join together to strengthen and save Social Security." [Transcript, Page A22.]

Delivering his State of the Union address three days after Iraqis went to the polls in their first free election in half a century, Mr. Bush promised not to end the American mission there before the Iraqis are capable of providing their own security against the bloody insurgency.

Can we summarize the SOUs?

- Bush's speech was about *Social Security*? - Speech #3

  Boom! Check!

How'd we do?

- Obama 2013
- Jobs, jobs, jobs

Obama Pledges Push to Lift Economy for Middle Class

By MARK LANDLER FEB. 12, 2013

WASHINGTON — President Obama, seeking to put the prosperity and promise of the middle class at the heart of his second-term agenda, called on Congress on Tuesday night to raise the federal minimum wage to $9 an hour, saying that would lift millions out of poverty and energize the economy.

In an assertive State of the Union address that fleshed out the populist themes of his inauguration speech, Mr. Obama declared it was "our generation's task" to "reignite the true engine of America's economic growth — a rising, thriving middle class."

"Every day," he said, "we should ask ourselves three questions as a nation: How do we attract more jobs to our shores? How do we equip our people with the skills to get those jobs? And how do we make sure that hard work leads to a decent living?"

The increase in the minimum wage, from $7.25 an hour now, was the most tangible of a raft of initiatives laid out by the president, from education and energy to public works projects. Taken together, Mr. Obama said, these

Can we summarize the SOUs?

- Obama's speech was about *jobs*? - Speech #4

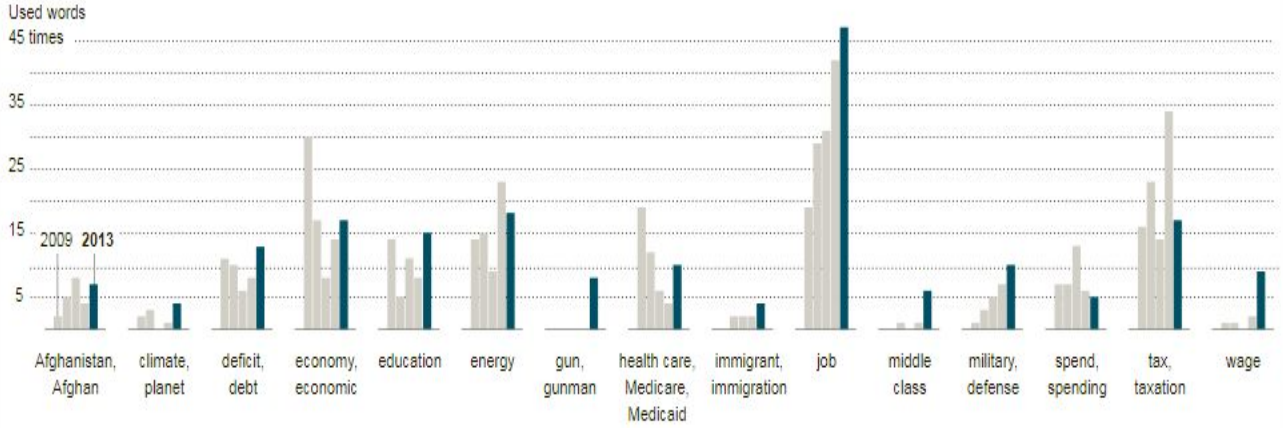
Boom! Check!

They have their own tools now ...

Published: February 12, 2013

Obama's State of the Union Themes

Selected words used by President Obama on Tuesday night and in his previous annual addresses. Related Article »



Note: Mr. Obama's 2009 speech was analogous to a State of the Union address, but without that title. Plural words are included in counts.

By ALICIA PARLAPIANO

Sources: The White House; Federal News Service

The good news

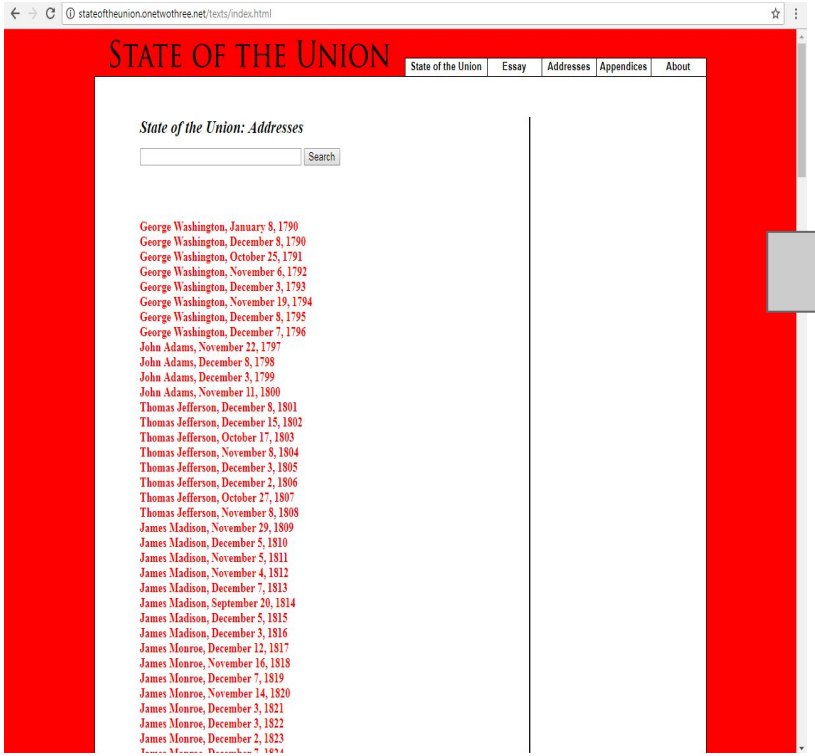
- You can do a lot of cool stuff

The bad news

- It’s hard to do all the cool stuff with one straightforward package

Doing text analysis in R

Getting the speeches in R




```
> reagan85 =
readLines("http://stateoftheunion.onetwothree.net/
texts/19850206.html")

> n.reagan85 = reagan85[59:517]

> head(n.reagan85)
[1] "Mr. Speaker, Mr. President, distinguished
Members of the Congress, honored"
[2] "guests, and fellow citizens:</p>"
[3] ""
[4] "<p>I come before you to report on the state
of our Union, and I'm pleased to"
[5] "report that after 4 years of united effort,
the American people have"
[6] "brought forth a nation renewed, stronger,
freer, and more secure than"
```

R tools

- See earlier...