EC999: Collocations

Thiemo Fetzer

University of Chicago & University of Warwick

January 17, 2017

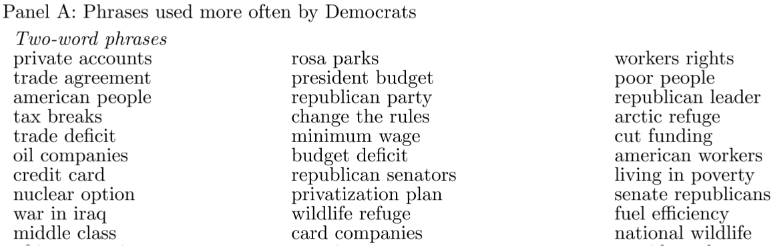
Collocations

> *Collocations of a given word are statements of the habitual or customary places of that word.*

- Noun phrases: “strong tea” and “weapons of mass destruction”
- Phrasal verbs: like “to make up”,
- Stock phrases: “the rich and powerful”

Collocations very useful for *terminology extraction*.

Measuring Political Slant



Panel A: Phrases used more often by Democrats

Two-word phrases

private accounts
trade agreement
american people
tax breaks
trade deficit
oil companies
credit card
nuclear option
war in iraq
middle class

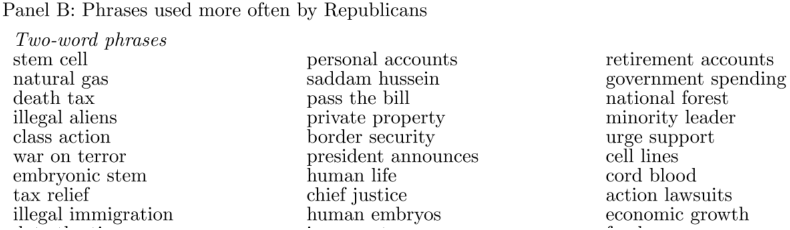
rosa parks
president budget
republican party
change the rules
minimum wage
budget deficit
republican senators
privatization plan
wildlife refuge
card companies

workers rights
poor people
republican leader
arctic refuge
cut funding
american workers
living in poverty
senate republicans
fuel efficiency
national wildlife

Identify n-grams (*collocations*) and extract those that are distinctively more likely to appear in the corpus of republican versus democratic congressional speeches.

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. Econometrica, 78(1), 3571.

Measuring Political Slant



Panel B: Phrases used more often by Republicans

Two-word phrases

stem cell
natural gas
death tax
illegal aliens
class action
war on terror
embryonic stem
tax relief
illegal immigration

personal accounts
saddam hussein
pass the bill
private property
border security
president announces
human life
chief justice
human embryos

retirement accounts
government spending
national forest
minority leader
urge support
cell lines
cord blood
action lawsuits
economic growth

Identify n-grams (*collocations*) and extract those that are distinctively more likely to appear in the corpus of republican versus democratic congressional speeches.

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. Econometrica, 78(1), 3571.

Plan

Heuristic Approaches

Statistical Tests

Application: χ^2 test for corpus similarity

A heuristic way of identifying collocations

A simple heuristic approach to identify collocations is simply counting raw occurences of word sequences, e.g. $C(w_1, w_2)$

```r
library(quanteda)
library(data.table)
data(SOTUCorpus, package = "quantedaData")
TOKENS <- unlist(tokenize(corpus_subset(SOTUCorpus, Date > "1993-01-01"), ngrams = 2,
    removeNumbers = TRUE, removePunct = TRUE, removeSymbols = TRUE, concatenator = " "))
DF <- data.table(token = TOKENS, president = names(TOKENS))
DF[, .N, by = token][order(N, decreasing = TRUE)][1:15]
```

```
##          token   N
##  1:     in the 603
##  2:     of the 577
##  3:     of our 386
##  4:     to the 324
##  5:     on the 271
##  6:    and the 264
##  7:    for the 250
##  8:  the world 245
##  9:    we must 242
## 10:    we have 203
## 11:     we can 203
## 12:    we will 185
## 13:      to do 181
## 14:  more than 170
## 15:     in our 167
```

selecting the most frequently occurring bigrams is not very interesting...

A refinement to heuristic

A simple method to make these more informative is to remove *stopwords*. Quanteda has a nice feature facilitating this through the `removeFeatures()` function.

```
TOKENS<-removeFeatures(tokenize(corpus_subset(SOTUCorpus, Date>'1994-01-01'), removePunct = TRUE),stopwords("
TOKENS<-unlist(tokens_ngrams(TOKENS, n=2, concatenator=" "))
DF<-data.table("token"=TOKENS, "president"=names(TOKENS))
DF<-DF[, .N, by=token][order(N, decreasing=TRUE)]
DF$N<-as.numeric(DF$N)
DF[1:10]
```

```
##              token   N
##  1:      health care 130
##  2: American people 117
##  3:   United States 106
##  4: Social Security  96
##  5:       men women  83
##  6:       make sure  73
##  7:    21st century  67
##  8:       years ago  59
##  9:       last year  53
## 10:    ask Congress  52
```

```
#DF<-DF[1:5000]
```

the resulting bigrams are much more intuitive. Though this still presents no formal statistical method.

Another refinement to heuristic

A further refinement due to Part of speech (POS) tag patterns for collocation filtering. We will talk more about POS, but here is just an illustration

```
# install.packages('pacman')
library(pacman)
# loads packages in development
pacman::p_load_gh(c("trinker/termco", "trinker/tagger", "trinker/textshape"))
# THIS TAKES A WHILE
TAGGED <- unlist(lapply(lapply(tag_pos(DF$token), names), function(x) paste(x,
    collapse = " ")))
DF <- cbind(DF, TAGGED)
head(DF)
```

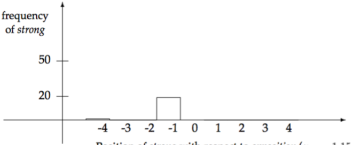
```
##                token   N   TAGGED
## 1:     health care 130    NN NN
## 2: American people 117   JJ NNS
## 3:   United States 106 NNP NNPS
## 4: Social Security  96  NNP NNP
## 5:       men women  83  NNS NNS
## 6:       make sure  73    VB JJ
```

We can now filter out individual sequences of words that are common.

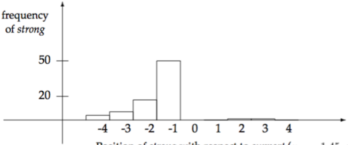
```
##                  token TAGGED         token TAGGED
## 1:       health care  NN NN 21st century  JJ NN
## 2: health insurance  NN NN    last year  JJ NN
## 3:       world power  NN NN    Last year  JJ NN
## 4:        tax relief  NN NN   first time  JJ NN
## 5:        tax credit  NN NN  high school  JJ NN
## 6:        child care  NN NN clean energy  JJ NN
```

Word location distances

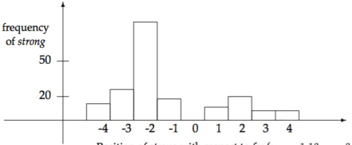
An alternative method is to look at the distribution of distances between words in a corpus of texts and pick candidate word pairs as those that are “nearby”.



Position of *strong* with respect to *opposition* ($\mu = -1.15, \sigma = 0.67$).



Position of *strong* with respect to *support* ($\mu = -1.45, \sigma = 1.07$).



Position of *strong* with respect to *for* ($\mu = -1.12, \sigma = 2.15$).

Summary Heuristic Approaches

Can be surprisingly successful in identifying collocations. In this case, we saw that

1. A simple quantitative technique - a mere frequency filter
2. Joint with the importance of parts of speech

is able to produce quite some nice results.
We next turn to more formal statistical methods to identify and differentiate collocations.

Plan

Heuristic Approaches

Statistical Tests

Application: χ^2 test for corpus similarity

Formal Statistical Tests

We now turn to formal statistical tests to identify collocations. The tests are all a variant of testing the hypothesis that the sequence of words is drawn at random, formally this hypothesis can be stated as

$$H_0 : \quad P(w_1, w_2) = P(w_1)P(w_2)$$

We present three approaches

- Simple T-tests
- χ^2 tests (also used to evaluate similarity of corpora)
- Likelihood ratio tests

T-Tests

For a word pair $w_1 w_2$, the hypothesis we want to test is whether:

$$H_0 : P(w_1 w_2) = P(w_1)P(w_2)$$

We can estimate the three parameters by looking at our data and estimating the number of times a word appears. For example for the word pair `health care`.

```
DF[token == "health care"]$N
## [1] 130
sum(DF[grep("^\\bhealth\\b", token)]$N)
## [1] 237
sum(DF[grep("\\bcare\\b$", token)]$N)
## [1] 239
sum(DF$N)
## [1] 78631
```

Under the Null, this is a *Bernoulli trial* whose probability of success we can estimate as:

$$\begin{aligned} P(\text{health care}) &= P(\text{health}) \times P(\text{care}) \\ &= \frac{237}{7.8631 \times 10^4} \times \frac{239}{7.8631 \times 10^4} = 9.1613326 \times 10^{-6} \end{aligned}$$

T-Tests

```
DF[token == "health care"]$N
## [1] 130
sum(DF$N)
## [1] 78631
xbar = DF[token == "health care"]$N/sum(DF$N)
```

We estimate

$$P(\text{health care}) = \frac{130}{7.8631 \times 10^4}$$

Under H_0, T-statistic follows approximately a t-distribution with $N - 1$ degrees of freedom.

$$T = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

with $\mu =$ and variance $\sigma^2 = p(1 - p)$ (Variance of Bernoulli distribution).
So compute

$$T = \frac{0.0016533 - 9.1613326 \times 10^{-6}}{\sqrt{\frac{9.1612486 \times 10^{-6}}{7.8631 \times 10^4}}} = 152.3196326$$

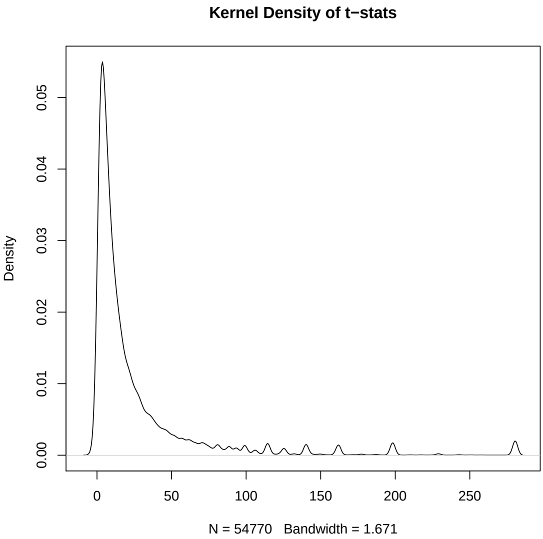
T-Tests for whole data frame

```
DF[, `:=`(ttest, (N/total - c_1/total * c_2/total)/(((c_1/total * c_2/total *
    (1 - c_1/total * c_2/total))/total)^0.5))]
DF[order(ttest, decreasing = TRUE)][c_1 + c_2 > 20][1:20]
```

```
##                     token   N   TAGGED total            w1          w2 c_1
##  1:    vacant storefronts  19   JJ NNS 78631        vacant storefronts  20
##  2:           Left Behind  11   VBN IN 78631          Left      Behind  11
##  3:       Social Security  96  NNP NNP 78631        Social    Security  96
##  4:            Child Left  11   NN VBN 78631         Child        Left  13
##  5:           Middle East  43  NNP NNP 78631        Middle        East  48
##  6:        Saddam Hussein  25  NNP NNP 78631        Saddam     Hussein  31
##  7:          Armed Forces  12 NNP NNPS 78631         Armed      Forces  12
##  8:         United States 106 NNP NNPS 78631        United      States 107
##  9:              al Qaeda  11   JJ NNP 78631            al       Qaeda  17
## 10:          minimum wage  24    NN NN 78631       minimum        wage  26
## 11:             men women  83  NNS NNS 78631           men       women 106
## 12:             God bless  28   NNP VB 78631           God       bless  47
## 13:          New Covenant  12  NNP NNP 78631           New    Covenant  23
## 14:       Tucson reminded  10  NNP VBD 78631        Tucson    reminded  14
## 15: distinguished guests  10  VBN NNS 78631 distinguished      guests  14
## 16:      mass destruction  17    NN NN 78631          mass destruction  24
## 17:       activist judges   8   NN NNS 78631      activist      judges   9
## 18:        teen pregnancy   7    JJ NN 78631          teen   pregnancy  13
## 19:        General ferret   7   NNP NN 78631       General      ferret  15
## 20:         Laden Zarqawi   7  NNP NNP 78631         Laden     Zarqawi  13
##     c_2    ttest
##  1:  19 273.2425
##  2:  12 268.4333
##  3: 108 264.0123
##  4:  11 257.8991
##  5:  46 256.4379
##  6:  25 251.7183
##  7:  16 242.7947
##  8: 141 241.5544
##  9:  11 225.5147
## 10:  36 219.8643
```

T-Tests for whole data frame

```
plot(density(DF$ttest), main = "Kernel Density of t-stats")
```



Pearson's χ^2 tests

It turns out that t-tests are extremely optimistic (lots of false positives), but also that the underlying assumption of approximate normality is often invalid due to low counts. The χ^2 test we discuss next is more useful and allows for meaningfull cross corpora analysis. The χ^2 test

- Compares the observed frequencies in the table with the frequencies expected for independence.
- If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

We can think of as word counts for a specific bigram to be arranged in tabular format

	$w_1 =$ health	$w_1 \neq$ health
$w_2 =$ care	130	109
$w_2 \neq$ care	107	7.8394×10^4

Note that $\sum_j C(w_1 = \text{health}, w_j) = C(w_1 = \text{health})$.

Pearson's χ^2 tests

We can think of as word counts for a specific bigram to be arranged in tabular format

	$w_1 =$ health	$w_1 \neq$ health
$w_2 =$ care	130	109
$w_2 \neq$ care	107	7.8394×10^4

The χ^2 test statistic sums the differences between observed and expected values in all cells of the table, scaled by the magnitude of the expected values, as follows

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} measures the observed count, while E_{ij} measures the expected count. It is usually thought of as a measure of *goodness of fit* to evaluate the fit of empirical against. χ^2 tests are commonly implemented for collocation detected, e.g. in the `quanteda` R-package.

Computing χ^2 tests statistic

	w_1 = health	$w_1 \neq$ health	
w_2 = care	130	109	239
	0.7203647	238.2796353	
$w_2 \neq$ care	107	7.8285×10^4	7.8392×10^4
	236.2796353	7.815572×10^4	
	237	7.8394×10^4	7.8631×10^4

- Expected frequencies E_{ij} computed from the marginal probabilities; compute totals of rows and columns and convert to proportions.
- Example: expected frequency ("health care") is marginal probability of "health" occurring as the first part of a bigram times the marginal probability of "care" occurring as the second.

$$X^2 = \frac{(130 - 0.7203647)^2}{0.7203647} + \frac{(109 - 238.2796353)^2}{238.2796353}$$
$$+ \frac{(107 - 236.2796353)^2}{236.2796353} + \frac{(7.8285 \times 10^4 - 7.815572 \times 10^4)^2}{7.815572 \times 10^4}$$

Special formula for 2x2 tables

	$w_1 =$ health	$w_1 \neq$ health	
$w_2 =$ care	130	109	239
$w_2 \neq$ care	107	7.8285×10^4	7.8392×10^4
	237	7.8394×10^4	7.8631×10^4

For 2 x 2 tables, there is a condensed formula [Can you show this?]

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

For a 2 x 2 design, this statistic has a χ^2 distribution with one degree of freedom. Can you show that this statistic reaches maximal value in case off diagonals are zero (words exclusively appear together)?

Identified Bigrams in State of the Union Speeches

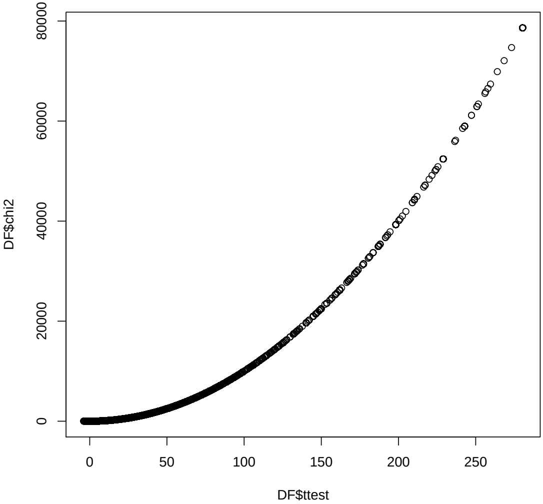
```
DF[, `:=`(O11, N)]
DF[, `:=`(O12, (c_2 - N))]
DF[, `:=`(O21, (c_1 - N))]
DF[, `:=`(O22, (total - c_1 - c_2 + N))]
DF[, `:=`(chi2, (total * (O11 * O22 - O12 * O21)^2)/((O11 + O12) * (O11 + O21) *
    (O12 + O22) * (O21 + O22)))]

DF[c_1 + c_2 > 20][order(chi2, decreasing = TRUE)][1:30]

##                    token   N   TAGGED total            w1          w2 c_1
##  1:   vacant storefronts  19   JJ NNS 78631        vacant storefronts  20
##  2:          Left Behind  11   VBN IN 78631          Left      Behind  11
##  3:      Social Security  96  NNP NNP 78631        Social    Security  96
##  4:           Child Left  11   NN VBN 78631         Child        Left  13
##  5:          Middle East  43  NNP NNP 78631        Middle        East  48
##  6:       Saddam Hussein  25  NNP NNP 78631        Saddam     Hussein  31
##  7:         Armed Forces  12 NNP NNPS 78631         Armed      Forces  12
##  8:        United States 106 NNP NNPS 78631        United      States 107
##  9:             al Qaeda  11   JJ NNP 78631            al       Qaeda  17
## 10:         minimum wage  24    NN NN 78631       minimum        wage  26
## 11:            men women  83  NNS NNS 78631           men       women 106
## 12:            God bless  28   NNP VB 78631           God       bless  47
## 13:         New Covenant  12  NNP NNP 78631           New    Covenant  23
## 14:      Tucson reminded  10  NNP VBD 78631        Tucson    reminded  14
## 15: distinguished guests  10  VBN NNS 78631 distinguished      guests  14
## 16:     mass destruction  17    NN NN 78631          mass destruction  24
## 17:      activist judges   8   NN NNS 78631      activist      judges   9
## 18:       teen pregnancy   7    JJ NN 78631          teen   pregnancy  13
## 19:       General ferret   7   NNP NN 78631       General      ferret  15
## 20:        Laden Zarqawi   7  NNP NNP 78631         Laden     Zarqawi  13
## 21:          White House  23  NNP NNP 78631         White       House  29
## 22:    preparing abandon  11   VBG NN 78631     preparing     abandon  15
## 23:          face bigger  49   NN RBR 78631          face      bigger 107
## 24:        playing field   9   VBG NN 78631       playing       field  15
## 25:        Madam Speaker  18  NNP NNP 78631         Madam     Speaker  18
## 26:          North Korea  12  NNP NNP 78631         North       Korea  25
```

Identified Bigrams in State of the Union Speeches

```
plot(DF$ttest, DF$chi2)
```



Problems for χ^2 tests

- T-test and χ^2 test statistic provide almost identical ordering
- χ^2 test is also appropriate for large probabilities, for which the normality assumption of the t-test fails.
- But approximation to the chi-squared distribution breaks down if expected frequencies are too low. It will normally be acceptable so long as no more than 20% of the events have expected frequencies below 5 (Read and Cressie 1988) $\rightarrow$ this is violated here.
- Rule of thumb: advise against using χ^2 if the expected value in any of the cells is 5 or less, use likelihood ratio test presented next.
- In case of low expected counts, perform *Yates correction*, modifying $X^2 = \sum_{ij} \frac{|(O_{ij}-E_{ij}|-0.5)^2}{E_{ij}}$

Likelihood Ratio Tests

Likelihood ratios are another approach to hypothesis testing. Developed in Dunning (1993), they are most appropriate for working with sparse data (few cell counts).

Two alternative hypothesis:

- Hypothesis 1: $P(w_2|w_1) = p = P(w_2|\neg w_1)$
- Hypothesis 2: $P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$

Hypothesis is just another way of stating the independence assumption (a draw of word w_2 is independent of any information regarding the occurence or non-occurence of word w_1). Hypothesis 2 says that the probability of w_2 following w_1 is different from probability of w_2 not following w_1.

It is clear that H_1 is *nested* into H_2.

Likelihood ratio test

Denote c_1, c_2 and c_{12} , for the number of occurences of word w_1, w_2 and the pair $w_1 w_2$.

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

We assume that word counts are binomially distributed

$$B(k; n, p) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Binomial distribution gives the probability of observing k heads in a sequence of n coin tosses, with success probability p.

What is the probability of observing counts c_{12} in c_1 trials?

$$B(c_{12}; c_1, p) = \binom{c_1}{c_{12}} p^{c_{12}} (1-p)^{(c_1 - c_{12})}$$

What is the probability of observing counts $c_2 - c_{12}$ in $N - c_{12}$ trials? I.e. probability of seeing c_2 by itself?

$$B(c_2 - c_{12}; N - c_{12}, p) = \binom{N - c_{12}}{c_2 - c_{12}} p^{c_2 - c_{12}} (1-p)^{(N - c_{12})}$$

Likelihood ratio test

Under Hypothesis 1 & 2, the likelihood of observing counts are given as

$$L(H_1) = B(c_{12}; c_1, p)B(c_2 - c_{12}; N - c_1, p)$$
$$L(H_2) = B(c_{12}; c_1, p_1)B(c_2 - c_{12}; N - c_1, p_2)$$

Likelihood ratio

$$\begin{aligned}\log(\lambda) &= \log \frac{L(H_1)}{L(H_2)} \\ &= \log(p^{(c_1 - c_{12})}(1-p)^{c_1}) + \log(p^{(c_2 - c_{12})}(1-p)^{(N-c_1)}) \\ &\quad - \log(p_1^{(c_1 - c_{12})}(1-p_1)^{c_1}) - \log(p_2^{(c_2 - c_{12})}(1-p_2)^{(N-c_1)})\end{aligned}$$

Advantage of LR test

- One advantage of likelihood ratios is that they have a clear intuitive interpretation: the exp of the LR provides a number that tells us how much more likely one hypothesis is than the other.
- So numbers are easier to interpret than the scores of the χ^2 test
- If λ is a likelihood ratio of a particular form, then the quantity $-2\log\lambda$ is asymptotically χ^2 distributed
- Dunning (1993) shows they are more appropriate for sparse data.

Plan

Heuristic Approaches

Statistical Tests

Application: χ^2 test for corpus similarity

χ^2 test for corpus similarity

- So far, we have used various statistical tests to study whether words appearing together appear so in a non-random fashion.
- We were comparing observed frequencies of pairs appearing with some notion of expected frequency under a null-hypothesis of independence.
- We can apply the same test to distinguish word use *between* texts.
- The null-hypothesis here is that the probability of observing a word or a word pair is independent across speakers.

χ^2 test for corpus similarity

We can use the χ^2 statistic to differentiate two corpora from one another or to identify distinctive word features characteristic of a corpus. Below is an example of Bush versus Obama state of the union speeches.

```
TOK <- data.table(sotu = names(TOKENS), token = TOKENS)
TOK[, `:=`(president, str_extract(sotu, "([A-z]*)"))]
TOK <- TOK[, .N, by = c("president", "token")][president %in% c("Bush", "Obama")]
TOK[order(N, decreasing = TRUE)][1:20]
```

```
##     president            token  N
##  1:      Bush  Social Security 45
##  2:      Bush    United States 44
##  3:     Obama      health care 44
##  4:     Obama  American people 44
##  5:      Bush        men women 41
##  6:     Obama    United States 37
##  7:      Bush      health care 34
##  8:      Bush    ground United 33
##  9:     Obama        make sure 31
## 10:     Obama      face bigger 28
## 11:     Obama     clean energy 28
## 12:     Obama        right now 27
## 13:      Bush  American people 26
## 14:      Bush      Middle East 26
## 15:      Bush     America will 25
## 16:     Obama        years ago 24
## 17:      Bush Members Congress 23
## 18:     Obama   States America 23
## 19:      Bush   Saddam Hussein 22
## 20:      Bush       tax relief 21
```

χ^2 test to identify distinct words across corpora

We can coonvert this into wide format and compute the χ^2 test statistic for each word feature, i.e. computing

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

```r
WIDE <- data.table(join(TOK[president == "Bush"][, list(token, bushcount = as.numeric(N))],
    TOK[president == "Obama"][, list(token, obamacount = as.numeric(N))], type = "full"))
WIDE[is.na(obamacount)]$obamacount <- 0
WIDE[is.na(bushcount)]$bushcount <- 0
WIDE[, `:=`(totalcount, obamacount + bushcount)]
WIDE <- WIDE[order(totalcount, decreasing = TRUE)][totalcount > 5]
WIDE[, `:=`(totalobama, sum(obamacount))]
WIDE[, `:=`(totalbush, sum(bushcount))]
WIDE[, `:=`(chi2, (totalobama + totalbush) * (bushcount * (totalobama - obamacount) -
    obamacount * (totalbush - bushcount))^2/((bushcount + obamacount) * (bushcount +
    (totalbush - bushcount)) * (obamacount + (totalobama - obamacount)) * ((totalbush -
    bushcount) + (totalobama - obamacount))))]
```

χ^2 test to identify distinct words across corpora

Present list sorted by X^2 test statistic score

```
WIDE[1:20][order(chi2, decreasing = TRUE)]
```

```
##                token bushcount obamacount totalcount totalobama totalbush
##  1:  Social Security        45         11         56       2225      1776
##  2:    ground United        33          5         38       2225      1776
##  3:        right now         1         27         28       2225      1776
##  4:     clean energy         2         28         30       2225      1776
##  5: Members Congress        23          5         28       2225      1776
##  6:      Middle East        26          7         33       2225      1776
##  7:        men women        41         21         62       2225      1776
##  8:      face bigger         8         28         36       2225      1776
##  9:     America will        25         14         39       2225      1776
## 10:   States America         7         23         30       2225      1776
## 11:        years ago         8         24         32       2225      1776
## 12:   every American         7         21         28       2225      1776
## 13:     ask Congress        18         11         29       2225      1776
## 14:    United States        44         37         81       2225      1776
## 15:        will help        16         10         26       2225      1776
## 16:        make sure        16         31         47       2225      1776
## 17: health insurance        16         12         28       2225      1776
## 18:  American people        26         44         70       2225      1776
## 19:    will continue        18         16         34       2225      1776
## 20:      health care        34         44         78       2225      1776
##            chi2
##  1: 29.76542402
##  2: 28.01000046
##  3: 19.03112218
##  4: 17.42404175
##  5: 16.28159834
##  6: 15.95019450
##  7: 12.05764904
##  8:  7.23091592
##  9:  6.20036699
## 10:  5.42860232
```

Combine this with POS Tag patterns

```
WIDE <- join(WIDE, DF[, list(token, TAGGED)])[grep("NN.? NN.?|JJ.? NN.?", TAGGED)]
WIDE[1:20][order(chi2, decreasing = TRUE)]
```

```
##                token bushcount obamacount totalcount totalobama totalbush
##  1:  Social Security        45         11         56       2225      1776
##  2:    ground United        33          5         38       2225      1776
##  3:   Saddam Hussein        22          0         22       2225      1776
##  4:       tax relief        21          2         23       2225      1776
##  5:     clean energy         2         28         30       2225      1776
##  6: Members Congress        23          5         28       2225      1776
##  7:      Middle East        26          7         33       2225      1776
##  8:  fellow citizens        20          4         24       2225      1776
##  9:        men women        41         21         62       2225      1776
## 10:       first time         2         20         22       2225      1776
## 11:        hard work         3         15         18       2225      1776
## 12:   States America         7         23         30       2225      1776
## 13:    United States        44         37         81       2225      1776
## 14:      high school         7         19         26       2225      1776
## 15:        last year         7         17         24       2225      1776
## 16: health insurance        16         12         28       2225      1776
## 17:  American people        26         44         70       2225      1776
## 18: small businesses        10         14         24       2225      1776
## 19:      health care        34         44         78       2225      1776
## 20:   Vice President        10         12         22       2225      1776
##            chi2   TAGGED
##  1: 29.76542402  NNP NNP
##  2: 28.01000046   NN NNP
##  3: 27.71432764  NNP NNP
##  4: 20.62658903    NN NN
##  5: 17.42404175    JJ NN
##  6: 16.28159834  NNS NNP
##  7: 15.95019450  NNP NNP
##  8: 14.83470943   NN NNS
##  9: 12.05764904  NNS NNS
## 10: 11.16558446    JJ NN
## 11:  5.62926109    JJ NN
```