PLSC 597: Modern Measurement

Scaling Text

April 12, 2018

Scaling Text

Scaling, so far:

- UDS / MDS, FA/PCA, IRT
- Goal: Combine/aggregate information (data reduction)

Scaling text: Underlying assumptions...

- Individuals speaking/writing/etc. differ in systematic, measurable ways
- Those differences manifest themselves in text...
  - What they say
  - When they say it (topic selection)
  - How they say it (style, tone, etc.)
- The mapping from latent differences to text is *systematic* and observable, and
- Can be learned via analysis of the text itself

Scaling Text (continued)

IRT-type data:

```
> irt.df[1:3,1:7]
        Q1  Q2  Q3  Q4  Q5  Q6  Q7
R1       1   1   1   0   0   0   0
R2       0   0   1   1   1   0   1
R3       1   0   1   0   0   1   0
```

Intuition: Go from binary "correct / incorrect" responses to measures of latent phenomena.

A TDM:

```
> tdm.df[1:3,1:7]
                 ability able about above abroad absolutely abused
Debate2016-1.txt       2   13    78     1      3          5      1
Debate2016-2.txt       0   16   102     1      0          5      0
Debate2016-3.txt       0    6    75     0      0          5      0
```

Intuition: Go from word frequencies / co-occurrences to measures of latent phenomena.

Supervised Text Scoring

Basic idea:

1. We know some documents' / authors' locations

2. Assess which terms in those documents give it it's location (distinctive)

3. Use the resulting term-level scores to locate other documents

One example: "Wordscores" (originally for scoring legislative text: speeches, press releases, etc.)

Wordscores: Details

Suppose there are N legislators, $i \in \{1, 2, ...N\}$. For each legislator, we observe D_i documents. Define:

$$\begin{aligned} \boldsymbol{X}_i &= \sum_{\ell=1}^{D_i} \boldsymbol{X}_{i\ell} \\ &= \sum_{\ell=1}^{D_i} (X_{i\ell 1}, X_{i\ell 2}, \ldots, X_{i\ell J}) \end{aligned}$$

as the aggregation of words across documents for each legislator.

Set two legislators' scores:

- Legislator L is **liberal** $(Y_L = -1)$
- Legislator C is **conservative** $(Y_C = 1)$

Wordscores: Details (continued)

For each word j, define:

$$\begin{aligned} P_{jL} &= \text{The probability that a liberal uses word } j \\ P_{jC} &= \text{The probability that a conservative uses word } j \end{aligned}$$

The score for word j is:

$$\begin{aligned} S_j &= Y_C P_{jC} + Y_L P_{jL} \\ &= P_{jC} - P_{jL} \end{aligned}$$

Wordscores: Details (continued)

From there, we can scale the remaining legislators / documents. Define:

$$N_i = \sum_{j=1}^{J} \boldsymbol{X}_i$$

$\hat{\theta}_i$ is then:

$$\begin{aligned} \hat{\theta}_i &= \sum_{j=1}^{J} \left(\frac{X_{ij}}{N_i} \right) S_j \\ &= \frac{\boldsymbol{X}_i'}{N_i} \boldsymbol{S} \end{aligned}$$

Wordscores: Details (continued)

Similarly, define:

$$
\begin{aligned}
N_L &= \sum_{m=1}^{J} X_{mL} \\
N_C &= \sum_{m=1}^{J} X_{mC}
\end{aligned}
$$

And estimate P_{jL}, P_{jC}, and S_j:

$$
\begin{aligned}
P_{jL} &= \frac{\frac{X_{jL}}{N_L}}{\frac{X_{jL}}{N_L} + \frac{X_{jC}}{N_C}} \\
P_{jC} &= 1 - P_{jL} \\
&= \frac{\frac{X_{jC}}{N_C}}{\frac{X_{jL}}{N_L} + \frac{X_{jC}}{N_C}} \\
S_j &= P_{jC} - P_{jL}
\end{aligned}
$$

Wordscores: Things to Remember

- Document scores are (weighted) averages of the words in them, where
- ...the weighting is "according to the proportion of tokens of each word type in the reference document" (Lowe 2008, 357)
- So, words' importance are a function of their frequency in each document type.
- Word-level scores are similar...
- Estimated document scores have vastly underestimated variability
- Issues with rescaling original texts for comparability (Martin and Vanberg 2007; Benoit and Laver 2007)
- Lowe (2008): Wordscores $\leftrightarrow$ Correspondence Analysis $\leftrightarrow$ IRT

Unsupervised Text Scoring

Basic idea:

1. Assume that words $\mathbf{X}$ are generated according to some PDF $f(\cdot)$, with (latent) parameters θ for the units being scaled
2. Assess $\Pr(\theta|f(\cdot), \mathbf{X})$
3. Resulting posterior $\hat{\theta}$ are your scale scores

Characteristics:

- IRT-like...
- One example: "Wordfish" (also originally for scoring legislators)

Wordfish Details

Again begin with "legislators" indexed by i and words by j. Assume that words are generated according to:

$$\begin{aligned} X_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

where

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i\text{'s overall word usage} \\ \psi_j &= \text{Word } j\text{'s frequency} \\ \beta_j &= \text{Word } j\text{'s discrimination} \\ \theta_i &= \text{Legislator } i\text{'s latent position(s)} \end{aligned}$$

One interpretation: A "regression" of X_{ij} on ideal points θ_i, where we have to learn θ_i.

WordFish Details

The assumptions above imply the following posterior distribution for the parameters (Slapin and Prokschk 2008):

$$\Pr(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto \Pr(\boldsymbol{\alpha}) \Pr(\boldsymbol{\beta}) \Pr(\boldsymbol{\psi}) \Pr(\boldsymbol{\theta}) \times \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right] \left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{X_{ij}}}{X_{ij}!}$$

Estimation accomplished by (choose one):

- EM
- MCMC
- Variational Approximation

WordFish...

- Yields estimates of the parameters $(\hat{\theta}, \hat{\alpha}, \hat{\psi}, \hat{\beta})$
- Also provides estimates of variability (method varies by estimation approach)
- More recently: Also estimates ideological clarity / ambiguity (Lo, Proksch and Slapin 2014 *BJPS*)

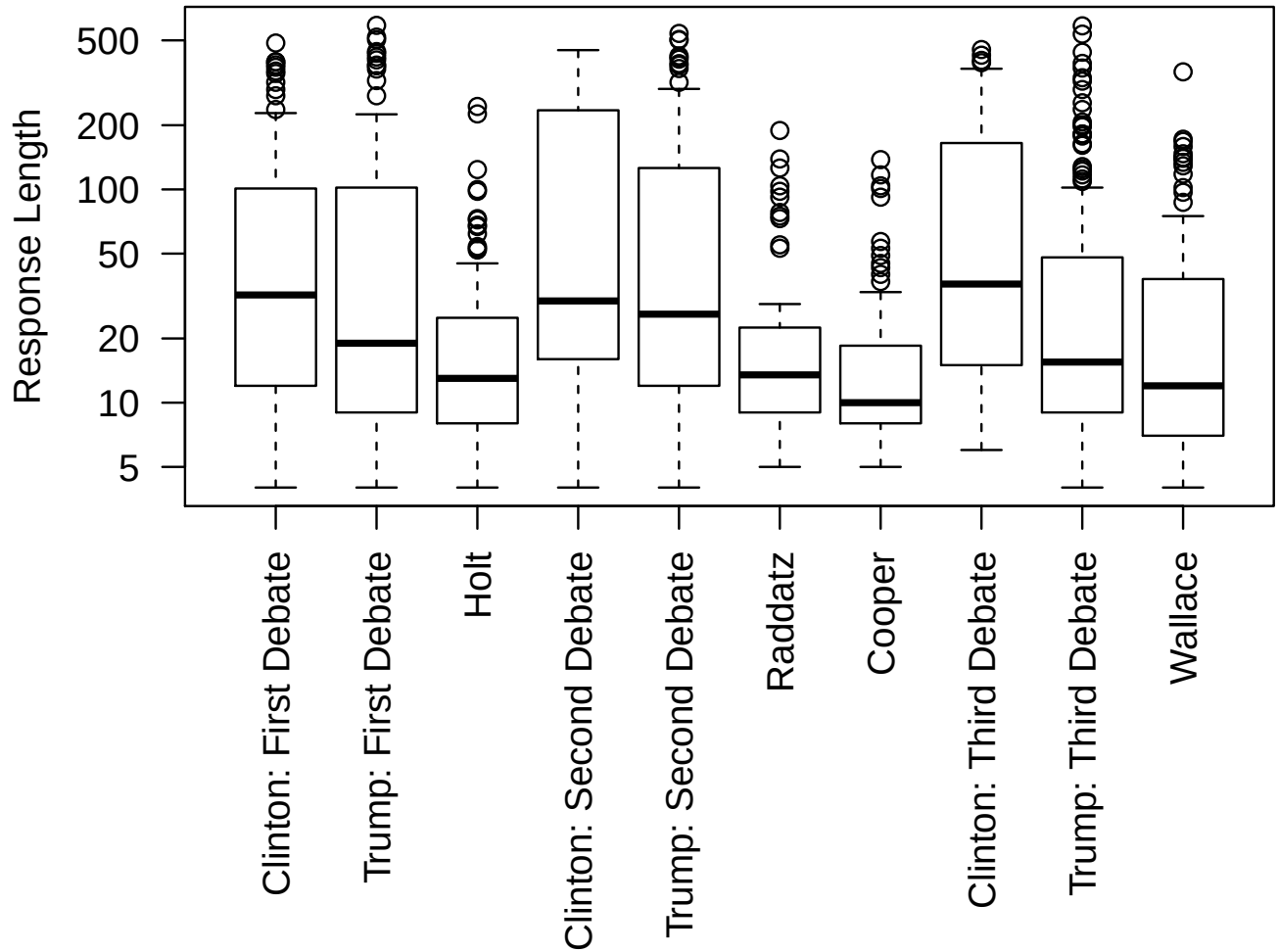
Text Scaling: Options in R

- `quanteda` (Benoit et al.)
- `austin` (Lowe)
- Various others (e.g., Slapin's `wordfish` code)

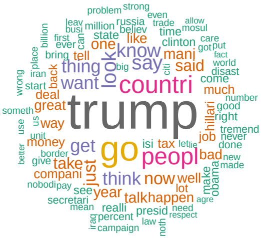
Example: The 2016 Presidential Debates

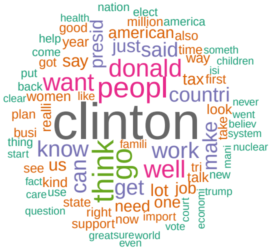
- Transcripts from all three general election (Clinton/Trump) debates
  - First Debate: 9/26/16, Hofstra University (Lester Holt moderating)
  - Second Debate: 10/9/16, Washington University (Martha Raddatz and Anderson Cooper moderating, town hall format)
  - Third Debate: 10/19/16, UNLV (Chris Wallace moderating)
- $N = 922$ "documents" (instances of one person speaking), 3986 sentences, 59256 tokens (34943 unique terms)
- Goals:
  - Scale Clinton, Trump, perhaps the moderators
  - Assess change from one debate to the next
  - ???

Length of Responses

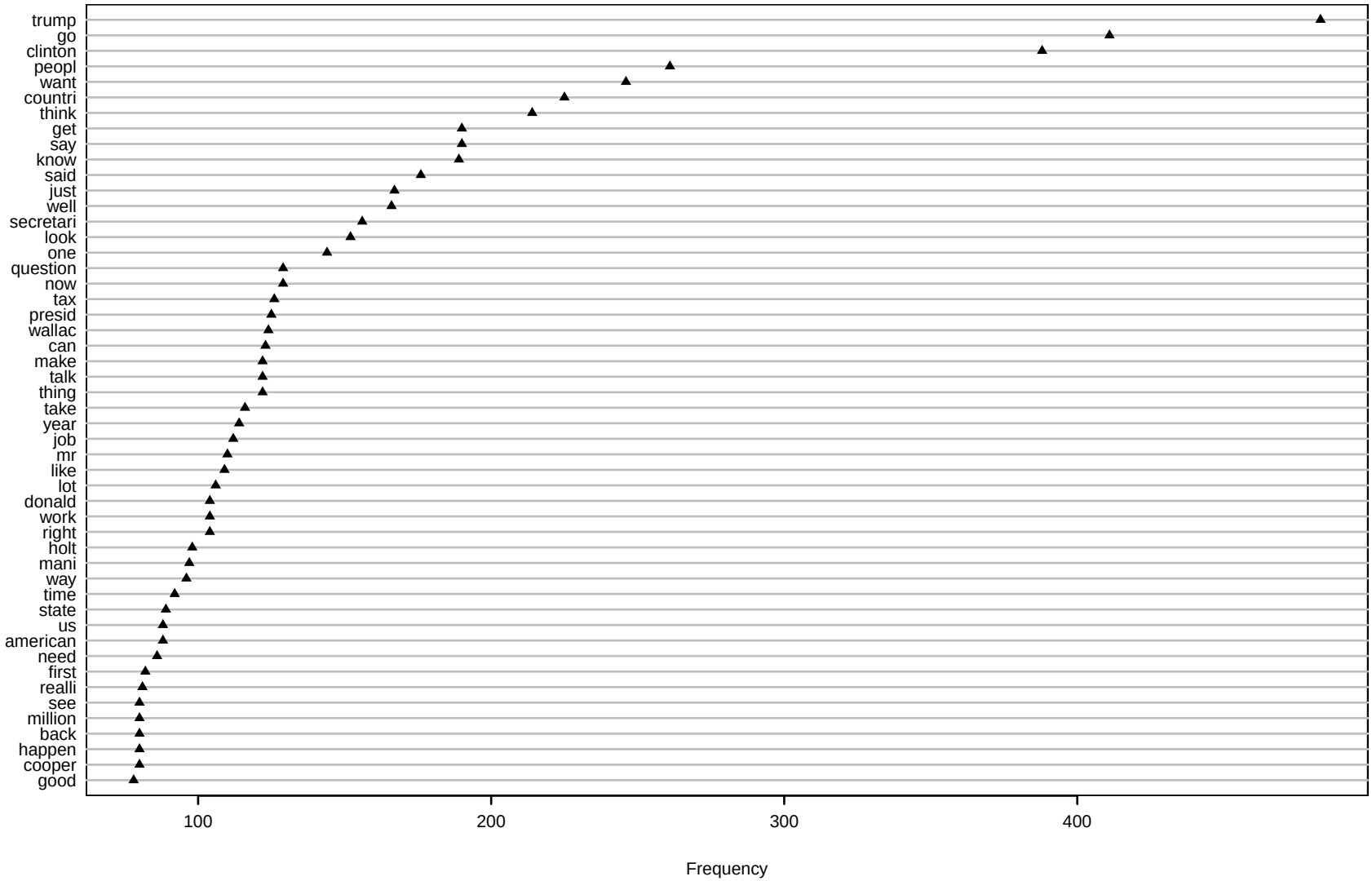


Word Clouds!


trump
go
countri
peopl
know
say
look
thing
want
get
think
now
just
take
one
mani
said
like
tell
clinton
time
year
talk
happen
well
job
bad
tax
isi
deal
great
way
money
compani
hillari
make
obama
much
problem
strong
even
trade
allow
mosul
russia
believ
million
leav
busi
state
first
ever
can
billion
bring
place
iran
back
wrong
start
someth
us
use
unit
better
border
give
nobodi
pay
see
secretari
mean
realli
percent
campaign
iraq
presid
law
noth
need
respect
agre
lot
new
made
done
never
tremend
right
good
number
come
disast
world
fact
put
got
care
big
citi
let
lie


clinton
peopl
donald
want
think
go
well
know
work
countri
said
just
presid
american
make
can
get
lot
job
one
need
say
year
good
health
nation
elect
million
america
also
time
way
someth
children
isi
tax
first
help
come
got
put
back
clear
women
like
plan
realli
busi
thing
start
see
us
fact
kind
care
use
question
state
right
support
now
import
vote
court
economi
trump
new
talk
tri
mani
nuclear
system
believ
went
never
look
take
famili
great
sure
world
even

Top 50 Words


trump
go
clinton
peopl
want
countri
think
get
say
know
said
just
well
secretari
look
one
question
now
tax
presid
wallac
can
make
talk
thing
take
year
job
mr
like
lot
donald
work
right
holt
mani
way
time
state
us
american
need
first
realli
see
million
back
happen
cooper
good
100
200
300
400
Frequency

Diversion: "Keyness"

Q: How good is a word (say, "terrorist") at *discriminating* among documents?

- Equally common (or rare) in both = not very
- Common in one, rare in the other = *very*

Intuition: a χ^2 statistic from a 2×2 frequency table:

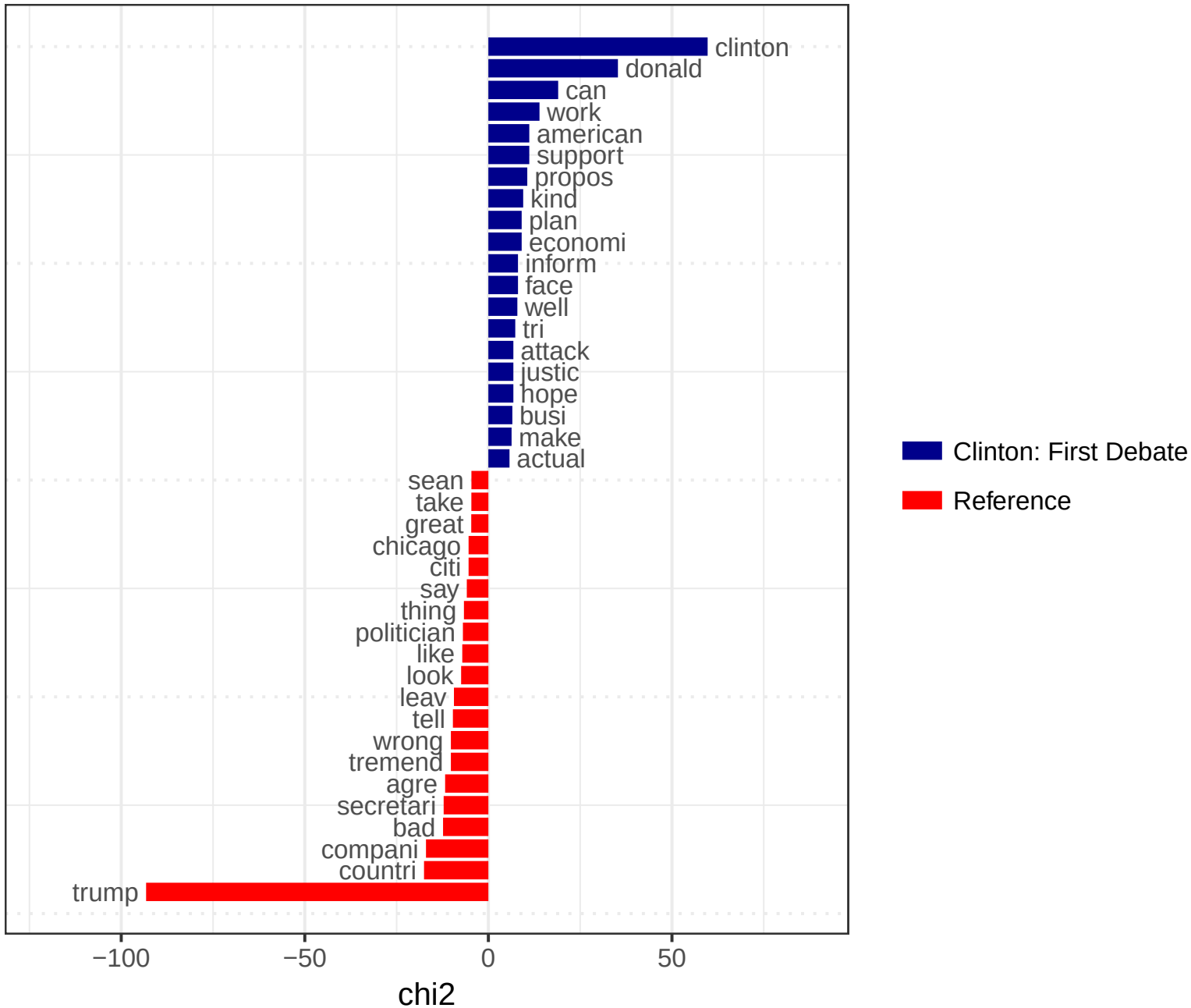
	"terrorist"	All other words	Total
Document A	N_{TA}	N_{OA}	N_A
Document B	N_{TB}	N_{OB}	N_B
Total	N_T	N_A	N

Larger values of $\chi^2 \rightarrow$ higher "keyness"

Word "Keyness": First Debate

```
> D1C2<-corpus_subset(D1C,Speaker %in% c("Clinton: First Debate",
+                                        "Trump: First Debate"))
> D1C2DFM <- dfm(D1C2,remove=stopwords("english"),stem=TRUE,
+            remove_punct=TRUE,groups="Speaker")
>
> D1Key <- textstat_keyness(D1C2DFM, target = "Clinton: First Debate")
>
> head(D1Key,12)
    feature   chi2         p n_target n_reference
1   clinton 59.722 1.088e-14       87          21
2    donald 35.291 2.840e-09       30           1
3       can 19.012 1.299e-05       30           8
4      work 13.924 1.903e-04       31          12
5   support 11.142 8.440e-04       13           2
6  american 11.142 8.440e-04       13           2
7    propos 10.600 1.131e-03       10           0
8      kind  9.480 2.078e-03       15           4
9   economi  9.064 2.607e-03       13           3
10     plan  9.064 2.607e-03       13           3
11     face  8.068 4.506e-03        8           0
12   inform  8.068 4.506e-03        8           0
```

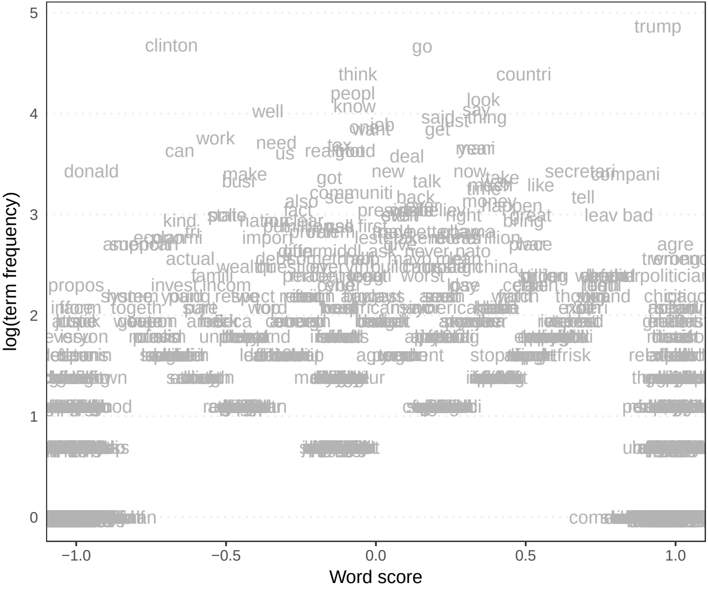
First Debate Keyness Differentials



Wordscores: Code

```
> # Work with the "candidates only" corpuses.
> # Set the statements made by Clinton in the first
> # debate equal to -1, and those made by Trump
> # to -1:
>
> TScores <- c(-1,1,NA)
>
> WS.train <- textmodel_wordscores(D1C2DFM,TScores,
+                                  scale="linear")
> summary(WS.train)

Call:
textmodel_wordscores.dfm(x = D1C2DFM, y = TScores, scale = "linear")

Reference Document Statistics:
                      score total min max mean median
Clinton: First Debate    -1  3030   0  87 1.89      1
Trump: First Debate       1  3866   0 127 2.42      1
Holt                     NA     0   0   0 0.00      0

Wordscores:
(showing first 30 elements)
    clinton      donald     applaus        well       thank      lester
   -0.68183    -0.94908    -0.01026    -0.36031    -0.43691     0.00383
    hofstra        host          us     central    question       elect
   -1.00000    -1.00000    -0.30346    -1.00000    -0.28219    -0.12123
     realli        kind     countri        want       futur       build
   -0.17276    -0.65426     0.49375    -0.01504    -1.00000     0.04637
     togeth       today granddaught      second    birthday       think
   -0.79862    -0.12123    -1.00000    -0.67233    -1.00000    -0.05995
         lot       first     economi        work     everyon        just
   -0.06904    -0.01026    -0.69367    -0.53446    -1.00000     0.26321
```

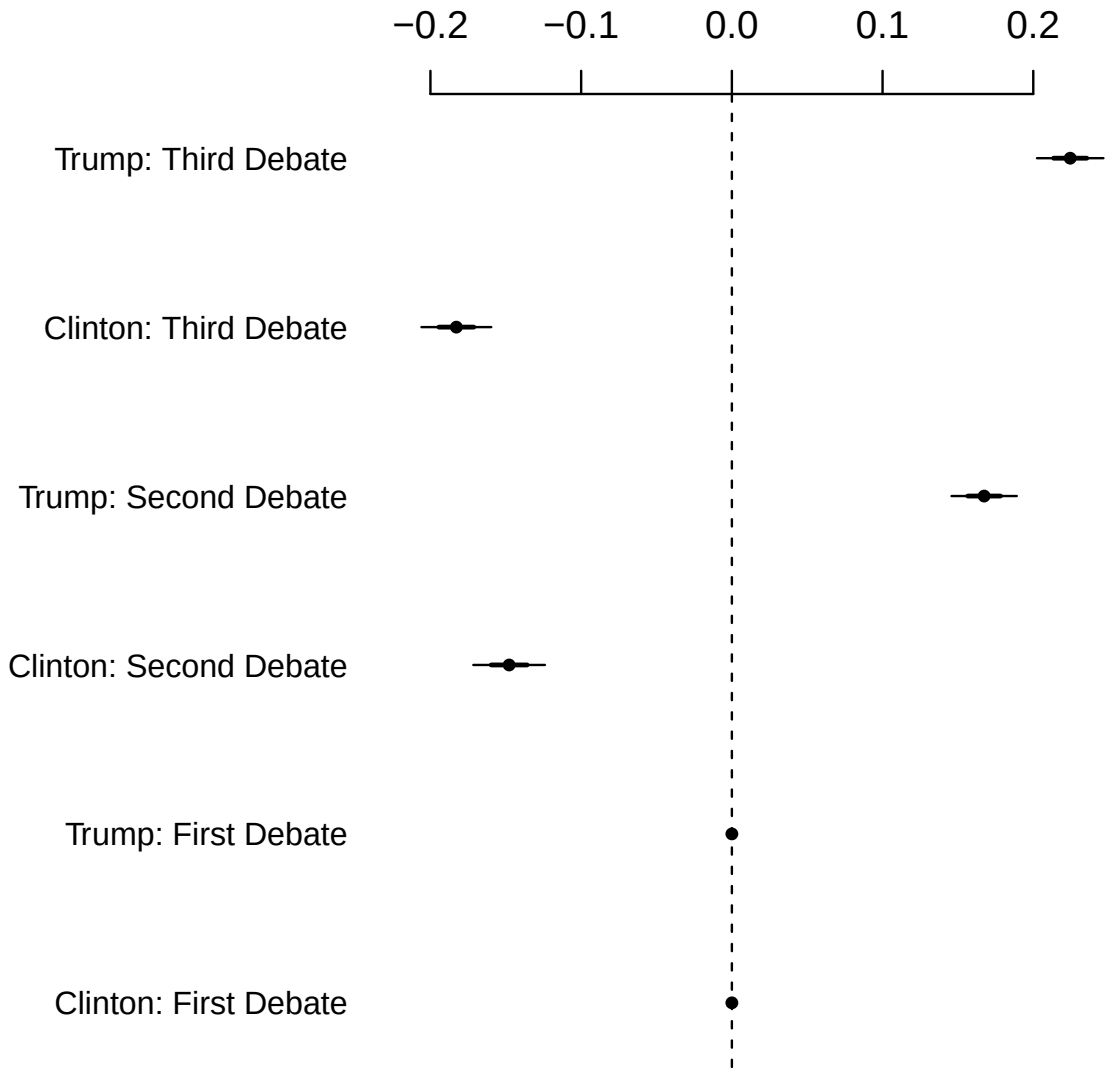
Word Scores...


trump
clinton
go
think
countri
peopl
know
look
say
well
thing
said
just
want
get
work
need
can
us
tax
real
good
year
mean
deal
donald
make
busi
new
now
take
secretari
compani
got
talk
like
time
communiti
see
back
money
tell
also
fact
happen
president
right
great
kind
state
nuclear
leav
bad
import
obama
place
agre
actual
nato
china
famili
wealth
politician
propos
invest
incom
cyber
worst
lose
inform
togeth
america
american
stop
frisk
agreement
log(term frequency)
Word score
5
4
3
2
1
0
−1.0
−0.5
0.0
0.5
1.0

Wordscores: Predicting

```
> # Predict to the second and third debates:
>
> D23Corpus <- corpus_subset(CTonly,Speaker=="Clinton: Second Debate" |
+                            Speaker=="Clinton: Third Debate" |
+                            Speaker=="Trump: Second Debate" |
+                            Speaker=="Trump: Third Debate",
+                            select=Speaker)
> D23DFM <- dfm(D23Corpus,remove=stopwords("english"),stem=TRUE,
+               remove_punct=TRUE,groups="Speaker")
>
> WS.test <- predict(WS.train,D23DFM,se.fit=TRUE,interval="confidence")
> WS.test
$fit
                           fit   lwr   upr
Clinton: First Debate     0.00   NaN   NaN
Trump: First Debate       0.00   NaN   NaN
Holt                      0.00   NaN   NaN
Clinton: Second Debate   -0.15 -0.17 -0.12
Trump: Second Debate      0.17  0.15  0.19
Raddatz                   0.00   NaN   NaN
Cooper                    0.00   NaN   NaN
Clinton: Third Debate    -0.18 -0.21 -0.16
Trump: Third Debate       0.22  0.20  0.25
Wallace                   0.00   NaN   NaN

$se.fit
 [1]   NaN   NaN   NaN 0.012 0.011   NaN   NaN 0.012 0.011   NaN
```

Wordscores: Ladder Plot



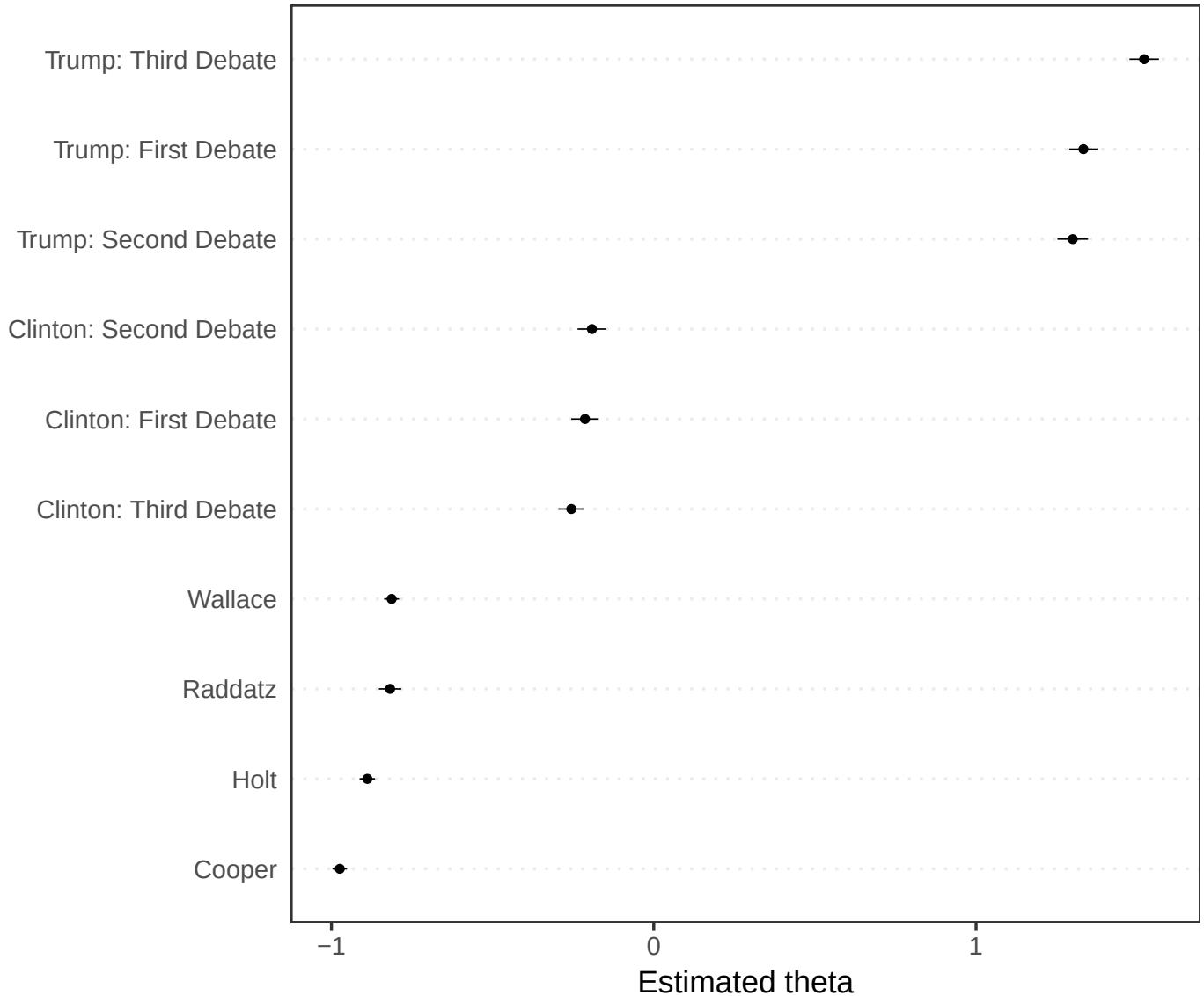
Rescaled Wordscores

```
> # Rescaled wordscores:
>
> WS.test2 <- predict(WS.train,D23DFM,se.fit=TRUE,interval="confidence",
+                     rescaling="lbg")
> WS.test2
$fit
                          fit  lwr  upr
Clinton: First Debate  -0.065  NaN  NaN
Trump: First Debate    -0.065  NaN  NaN
Holt                   -0.065  NaN  NaN
Clinton: Second Debate -1.782 -2.1 -1.5
Trump: Second Debate    1.880  1.6  2.1
Raddatz                -0.065  NaN  NaN
Cooper                 -0.065  NaN  NaN
Clinton: Third Debate  -2.189 -2.5 -1.9
Trump: Third Debate     2.544  2.3  2.8
Wallace                -0.065  NaN  NaN

$se.fit
 [1]   NaN   NaN   NaN 0.073 0.061   NaN   NaN 0.070 0.063   NaN
```

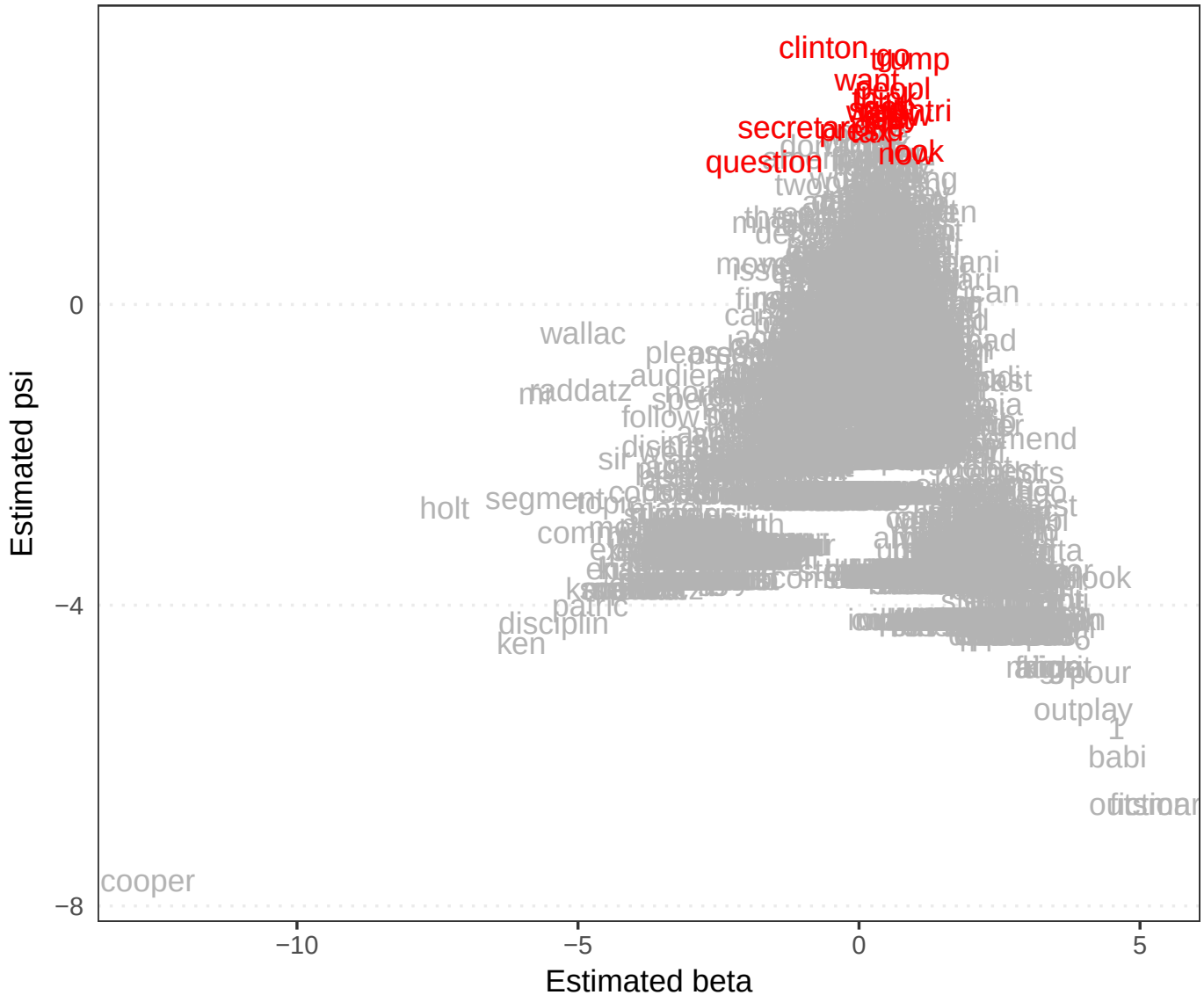
Wordfish!

```
> WF <- textmodel_wordfish(DDFM,dir=c(1,2))
> summary(WF)

Call:
textmodel_wordfish.dfm(x = DDFM, dir = c(1, 2))

Estimated Document Positions:
                         theta     se
Clinton: First Debate  -0.213 0.0218
Trump: First Debate     1.333 0.0224
Holt                   -0.889 0.0123
Clinton: Second Debate -0.192 0.0227
Trump: Second Debate    1.300 0.0243
Raddatz                -0.818 0.0177
Cooper                 -0.974 0.0115
Clinton: Third Debate  -0.256 0.0204
Trump: Third Debate     1.522 0.0233
Wallace                -0.813 0.0119

Estimated Feature Scores:
     clinton donald applaus  well thank lester hofstra   host    us central
beta  -0.626 -0.579  0.0833 0.252 -1.36  0.606   -1.43 -0.578 0.211   -1.15
psi    3.425  2.119  0.5430 2.571  1.17  0.139   -2.13 -1.427 1.948   -1.61
     question  elect realli   kind countri  want  futur build togeth  today
beta    -1.69 -0.513   0.57 -0.147   0.821 0.145 -0.561 0.679 -0.375 0.4548
psi      1.90  1.225   1.72  1.326   2.583 2.992 -0.442 0.441  0.553 0.0914
     granddaught second birthday think   lot   first economi    work
beta      -0.451  0.214   -0.451  0.46 0.383 -0.0648   -0.25 -0.0746
psi       -2.502  1.159   -2.502  2.75 2.077  1.9229    1.01  2.1613
     everyon  just
beta  -0.267 0.504
psi    0.351 2.478
```

Wordfish: Speaker Locations



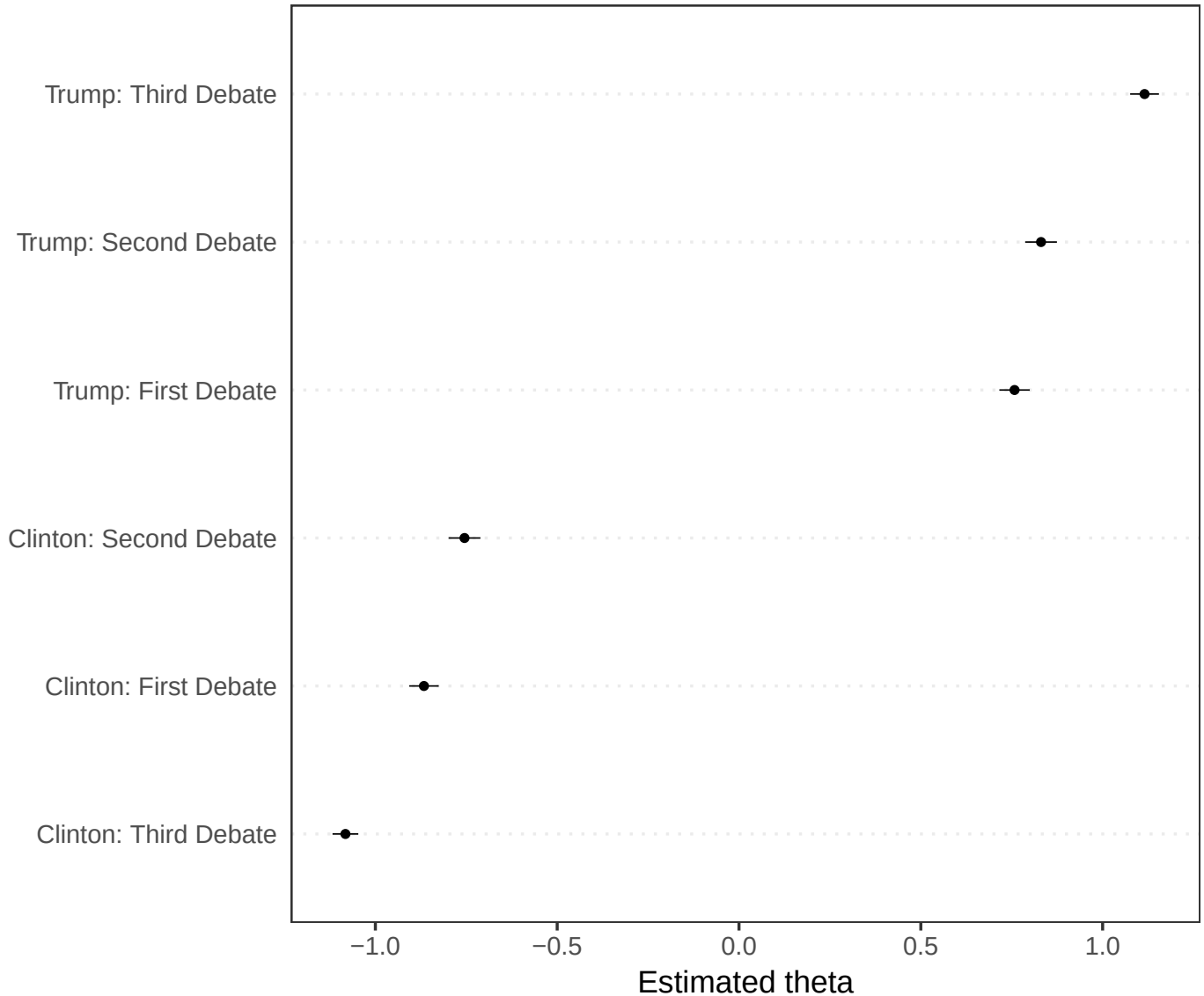
Wordfish: Word Locations



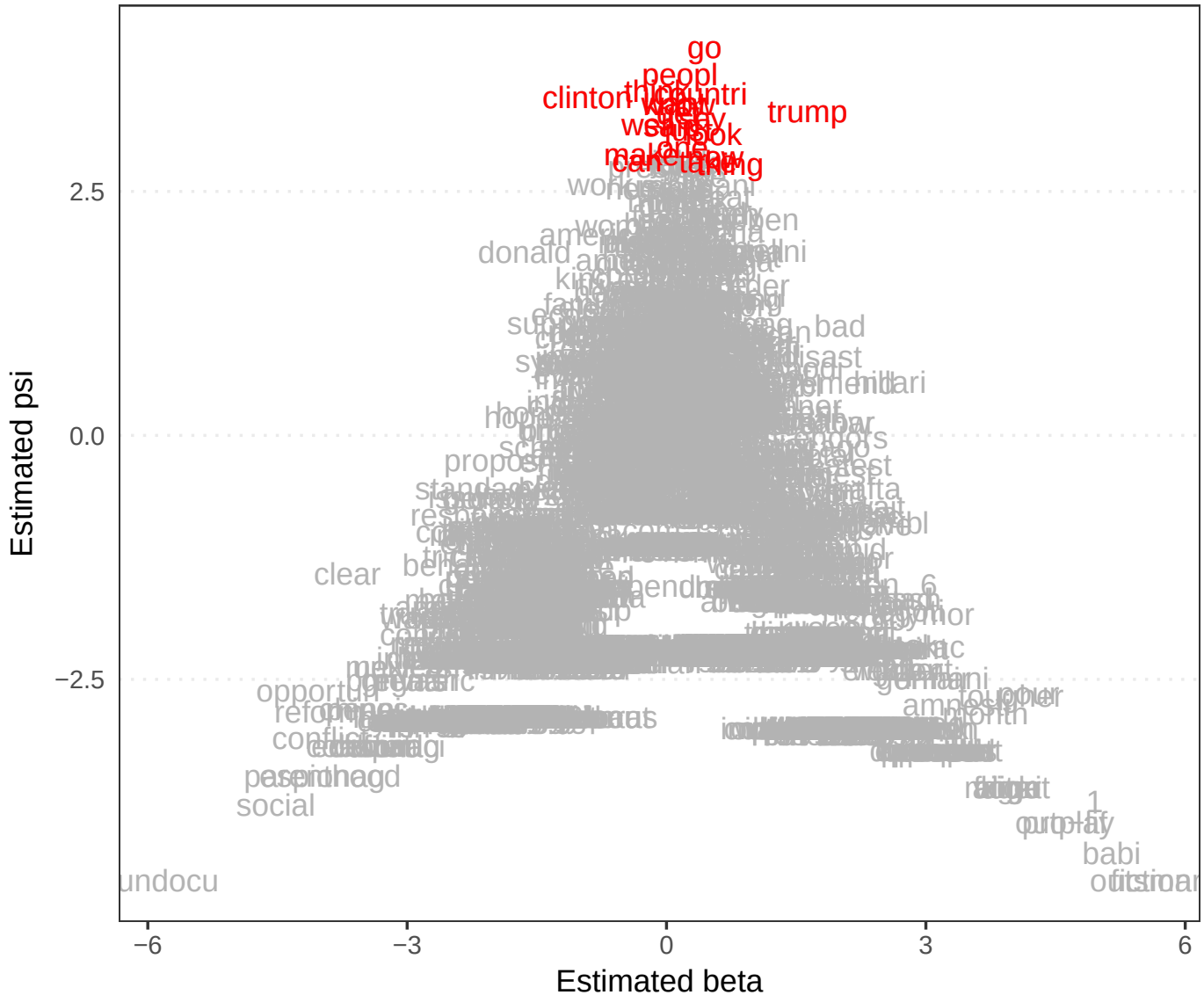
Wordfish: Candidates Only

```
> WF2 <- textmodel_wordfish(CTDFM,dir=c(1,2))
Note: removed the following zero-token documents: Holt
Note: removed the following zero-token documents: Raddatz
Note: removed the following zero-token documents: Cooper
Note: removed the following zero-token documents: Wallace

> summary(WF2)

Call:
textmodel_wordfish.dfm(x = CTDFM, dir = c(1, 2))

Estimated Document Positions:
                         theta     se
Clinton: First Debate  -0.866 0.0207
Trump: First Debate     0.758 0.0213
Clinton: Second Debate -0.755 0.0223
Trump: Second Debate    0.831 0.0223
Clinton: Third Debate  -1.082 0.0181
Trump: Third Debate     1.115 0.0202

Estimated Feature Scores:
     clinton donald applaus   well thank lester hofstra  host     us
beta  -0.916  -1.64   0.166 -0.212 0.230  0.113   -1.17 -2.41 -0.293
psi    3.466   1.88   0.834  3.188 0.895  0.971   -2.32 -2.27  2.558
     central question  elect realli  kind countri   want  futur build
beta   -1.84    -0.14 -0.874 0.0492 -0.96   0.394 0.0868 -1.905 0.181
psi    -2.17     1.77  1.317 2.5562  1.61   3.506 3.4032 -0.842 1.329
     togeth today granddaught second birthday  think     lot  first economi
beta  -1.01 0.646       -1.17  0.206    -1.17 -0.111 -0.0186 -0.118  -0.889
psi    0.74 0.463       -2.32  1.558    -2.32  3.536  2.7721  2.248   1.241
       work everyon  just
beta -0.764  -0.904 0.257
psi   2.574   0.637 3.139
```

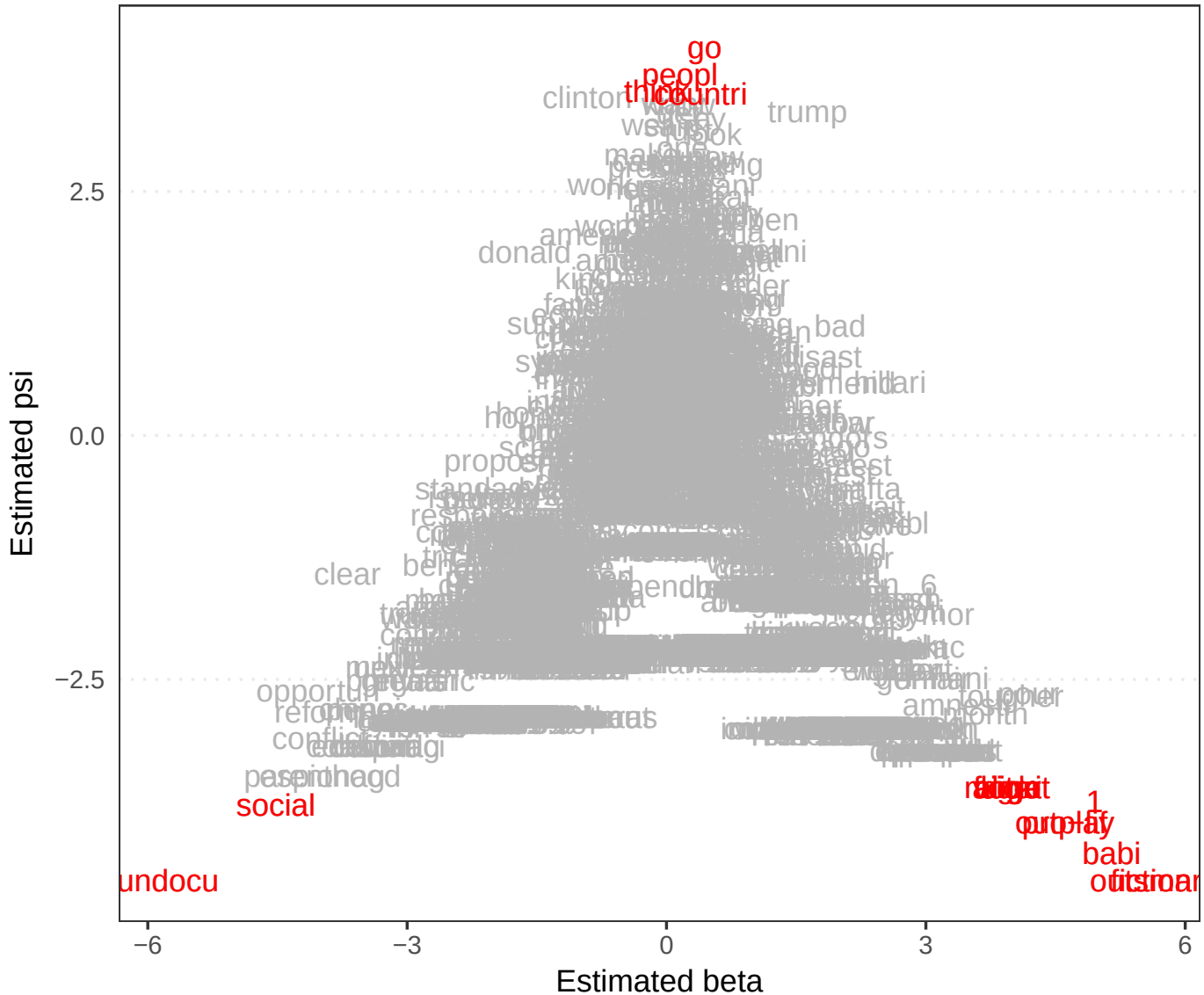
Wordfish: Speaker Locations (Candidates Only)



Wordfish: Word Locations (Frequent Words)



Wordfish: Word Locations (large $|\hat{\psi}|$)



Scaling Texts: Other Approaches...

- E.g., factor analysis / SEM, unfolding, IRT...
- The "Class Affinity Model"
  - Perry and Benoit (2017)
  - "...a text modeling framework that allows actors to take latent positions on a 'gray' spectrum between 'black' and 'white' polar opposites."
  - In `quanteda`
- They're all kinda the same.
- (Read this paper by Will Lowe: `http://dl.conjugateprior.org/preprints/all-on-the-line.pdf`)

Scaling Texts: Things to Think About

- Interpretation: What do the scales mean?
- What does it mean to "validate"?
  - Compare to human / expert coding?
  - Compare to "numerical" position estimates (D-NOMINATE / Martin-Quinn / etc.)?
  - Cross-validate?
  - Predicting other phenomena
- Propagating (measurement and estimation) uncertainty...