WELCOME to the Portland R User Group event

Unsupervised Politics, Machine Learning, & R

Winston Saunders

WELCOME to the Portland R User Group event

Unsupervised Politics, Machine Learning, and R with Winston Saunders

SPEAKER BIO:

Winston Saunders

- reformed-physicist
- Work at Intel. Security Technology Program Mgr.
- Began learning R, AI, and Data Science ~two years ago thru Coursera.

WELCOME to the Portland R User Group event

Unsupervised Politics, Machine Learning, and R with Winston Saunders

ABOUT PdxData.org:

- We're all about openness. We're a federation of independent Meetups, part of the larger data science conversation in Portland.
- Check out our website: http://pdxdata.org/
- We use http://pdxdata.slack.com/ for a back-channel during MeetUps and in between. Invite yourself here: http://pdxdata.org/slack/ ...
- **Join us for a beeR afterwards...**

WELCOME to the Portland R User Group event

Unsupervised Politics, Machine Learning, and R with Winst

A big thanks to our host, Mozilla!

...if you spend a whole Saturday studying rather than watching TV, there's no one there to pat you on the back or tell you you did a good job. Chances are what you learned studying all Saturday won't make you that much better at your job the following Monday. There are very few, almost no short-term rewards for these things. But it's a fantastic long-term investment.

- Andrew Ng

Agenda

- Candidates as Bags of Words
- Vector Representations
- Politics Meets Analytics
- Is AI possible in politics?

Data Practice

- Text of presidential debates downloaded from the UCSB Presidency Project.
- Transcripts cut/pasted to Apple Pages. Store as .txt files Subsequent processing in R.

```
PARTICIPANTS:
Former Secretary of State Hillary Clinton;
Former Governor Martin O'Malley (MD);
Senator Bernie Sanders (VT);
MODERATORS:
Lester Holt (NBC News)
Andrea Mitchell (NBC News)

HOLT: Good evening and welcome to the NBC News Youtube Democratic candidate's debate.
After all the campaigning, soon, Americans will have their say with the first votes of the
2016 campaign just 15 days away in Iowa. And New Hampshire not far behind.

Tonight will be the final opportunity to see these candidates face to face before the
voting begins. Our purpose here tonight is to highlight and examine the differences among
the three Democratic candidates. So let's get started.

Please welcome Secretary Hillary Clinton, Senator Bernie Sanders and Governor Martin
O'Malley. [applause]
```

Bags of Words

WordCloud

Clinton-Fiorina

```
c_wordcloud(clinton_all)
```

```
c_wordcloud(fiorina_all)
```

```
color_map2<-c("#00003B", "#041851", "#4E0812", "#051E65", "#650A16", "#103374", "#9F1A2D", "#482B25")

wordcloud(t_c, scale = c(4, 0.4), max.words = 150, min.freq = 20, random.order = FALSE, rot.per = 0.2,
          use.r.layout = FALSE, colors = color_map2)
```

WordCloud

Sanders - Trump

```
c_wordcloud(trump_all)
```

```
c_wordcloud(sanders_all)
```

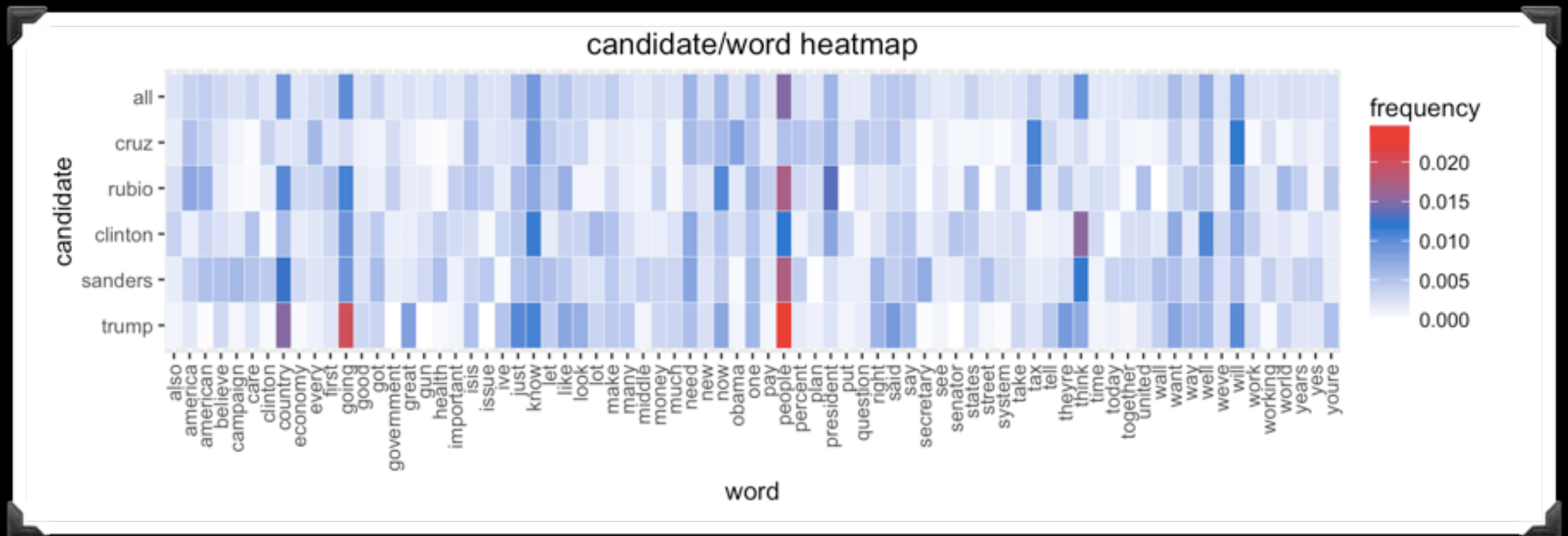
Vocabulary and Word Count

There are a total of 4853 words in the combined vocabulary of the candidates.

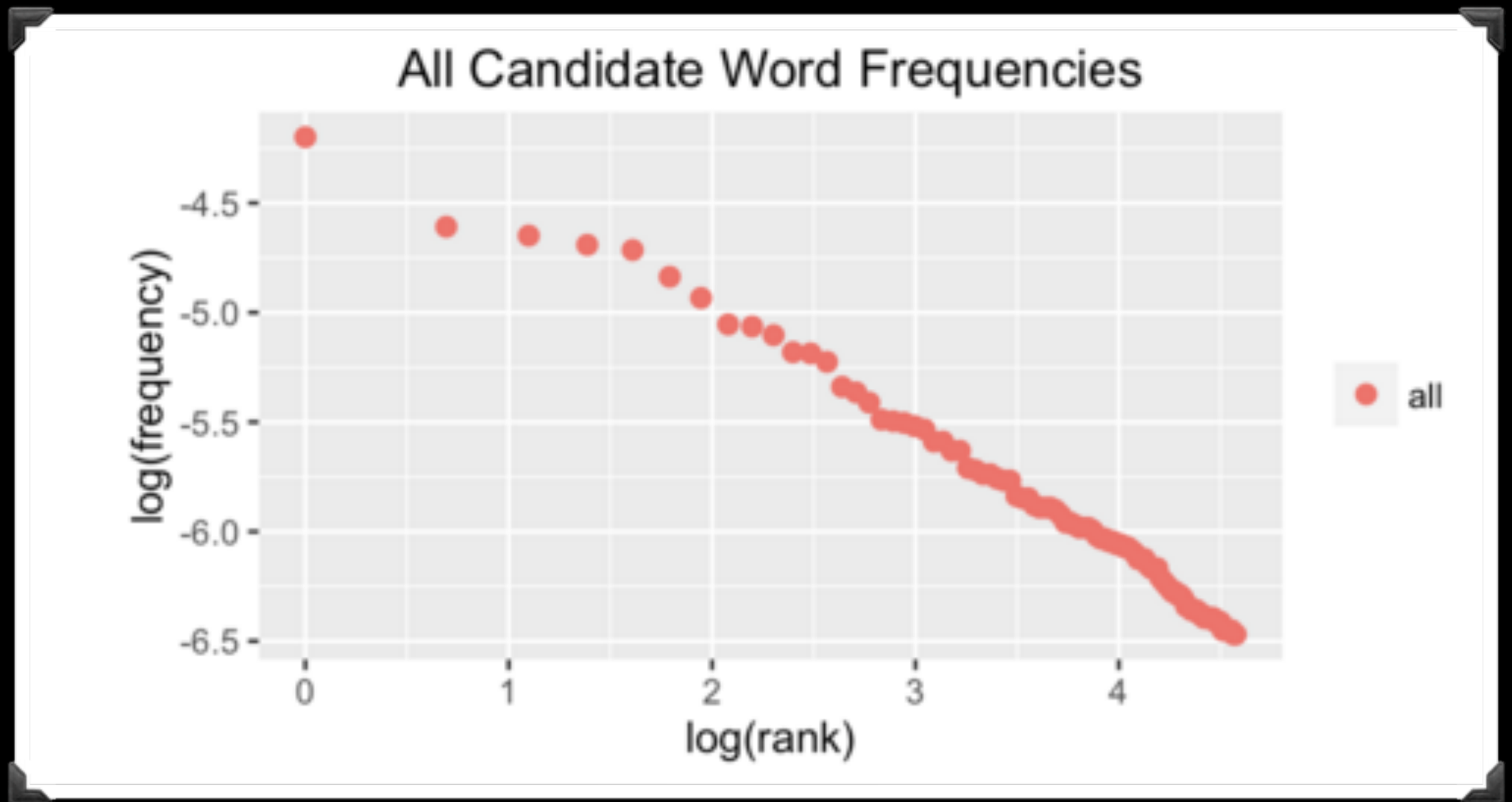
word	trump	sanders	clinton	rubio	cruz	all
people	105	162	117	86	24	494
going	84	85	90	56	13	328
think	31	111	150	10	13	315
country	64	119	57	52	10	302
know	47	55	115	37	41	295
will	43	48	70	45	55	261
well	26	58	104	23	26	237
need	23	72	73	15	27	210
president	10	25	75	68	30	208
now	32	40	48	53	27	200
just	44	45	47	28	13	177
tax	9	15	8	47	51	130
obama	4	5	26	10	37	82
SUM	4224	9306	9742	4607	5033	32912

- Hilary Clinton spoke a total of 9742 and had a vocabulary of 2477 words.
- Bernie Sanders spoke 9306 total words, with a vocabulary of 2109.
- Donald Trump spoke 4224 and with a vocabulary of 1206.
- Ted Cruz spoke 4607 and with a vocabulary of 1688.
- Marco Rubio spoke 5033 and with a vocabulary of 1583.

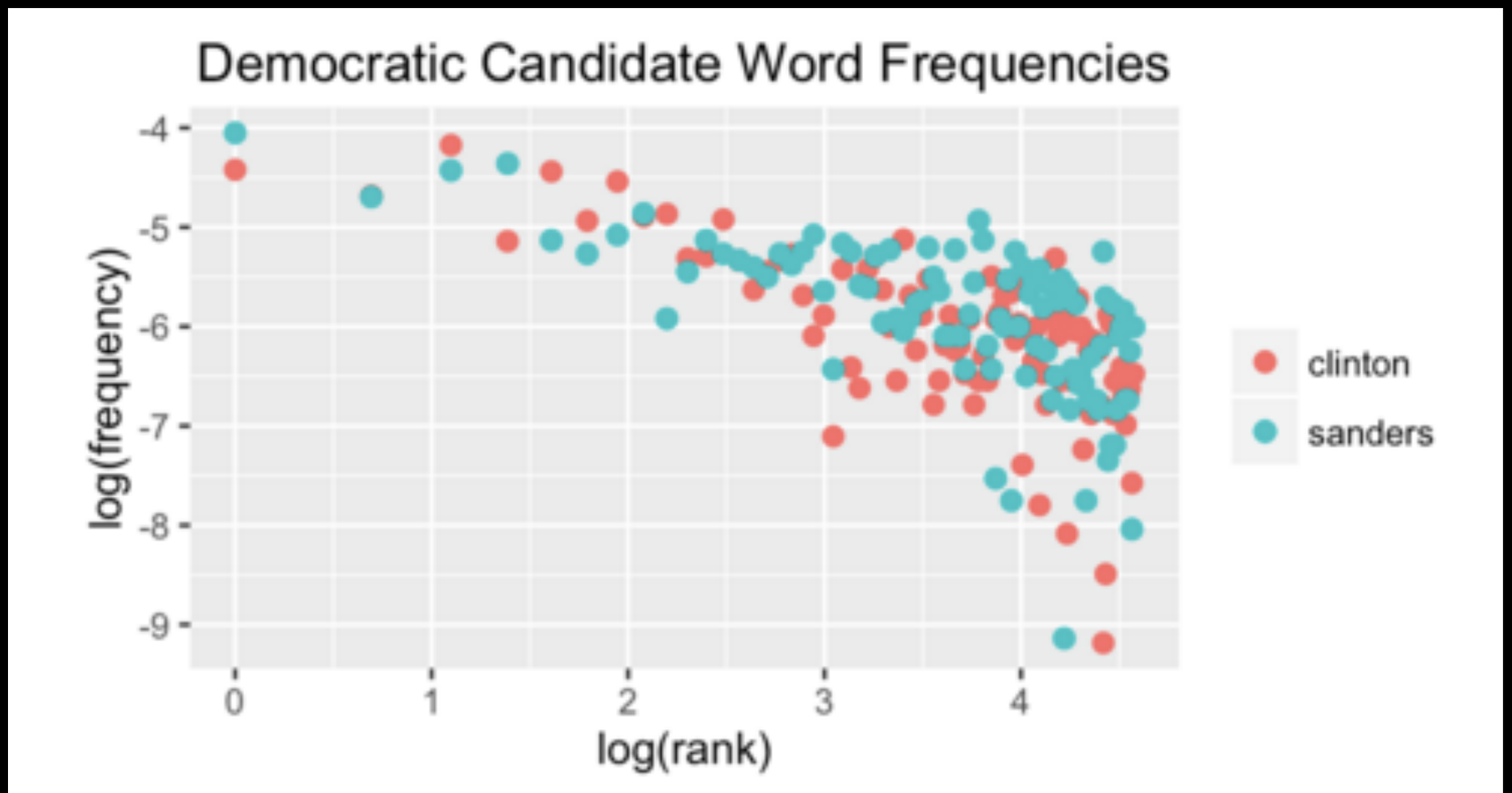
Heat Map


candidate/word heatmap
candidate
all
cruz
rubio
clinton
sanders
trump
word
also
america
american
believe
campaign
care
clinton
country
economy
every
first
going
good
got
government
great
gun
health
important
isis
issue
ive
just
know
let
like
look
lot
make
many
middle
money
much
need
new
now
obama
one
pay
people
percent
plan
president
put
question
right
said
say
secretary
see
senator
states
street
system
take
tax
tell
theyre
think
time
today
together
united
wall
want
way
well
weve
will
work
working
world
years
yes
youre
frequency
0.020
0.015
0.010
0.005
0.000

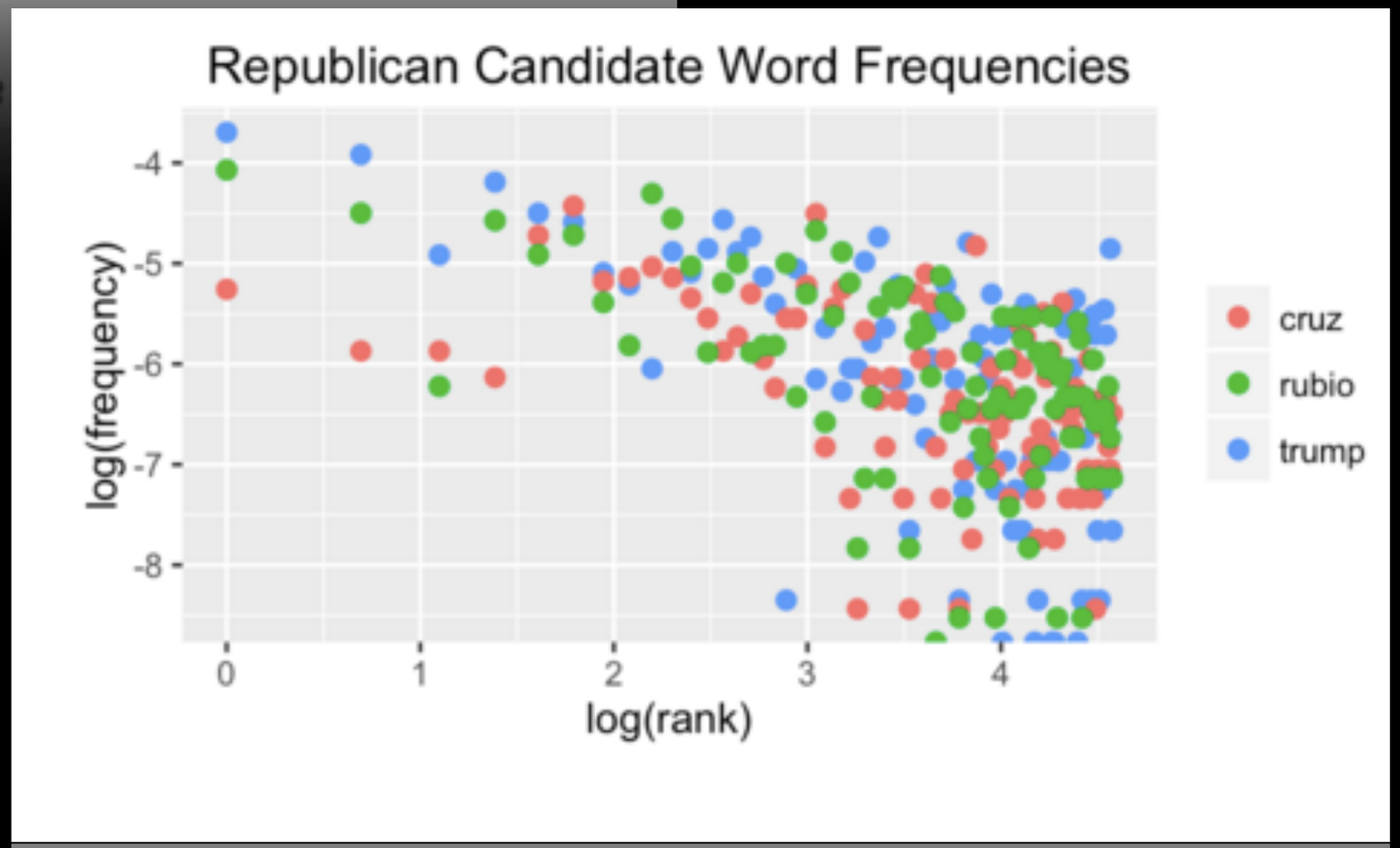
Zipf's (remarkable) Law



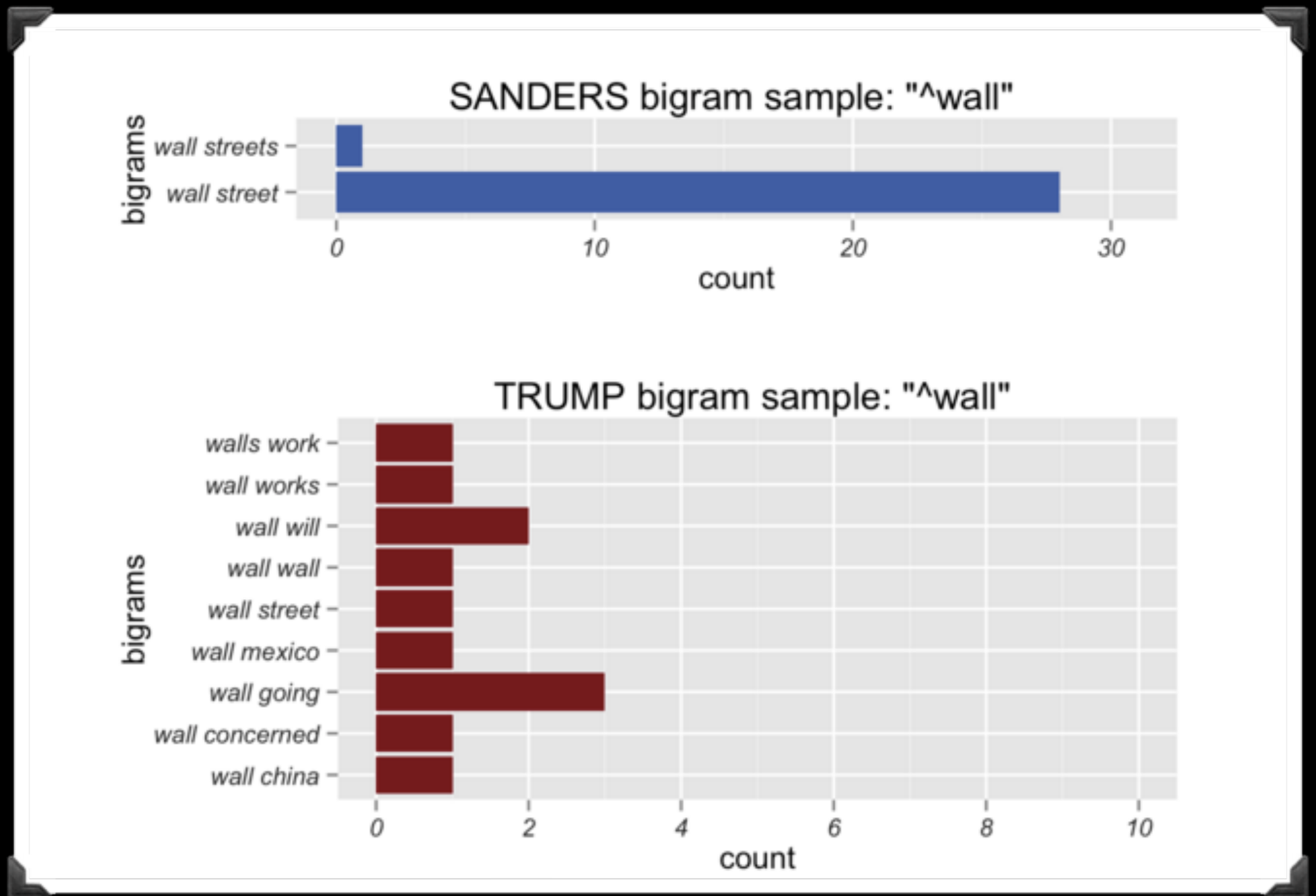
Observe that exponent = -1/2



Greatest differences occur for least-frequently used words (in most cases)



Bi-Gram Tokenization



...but

limited context

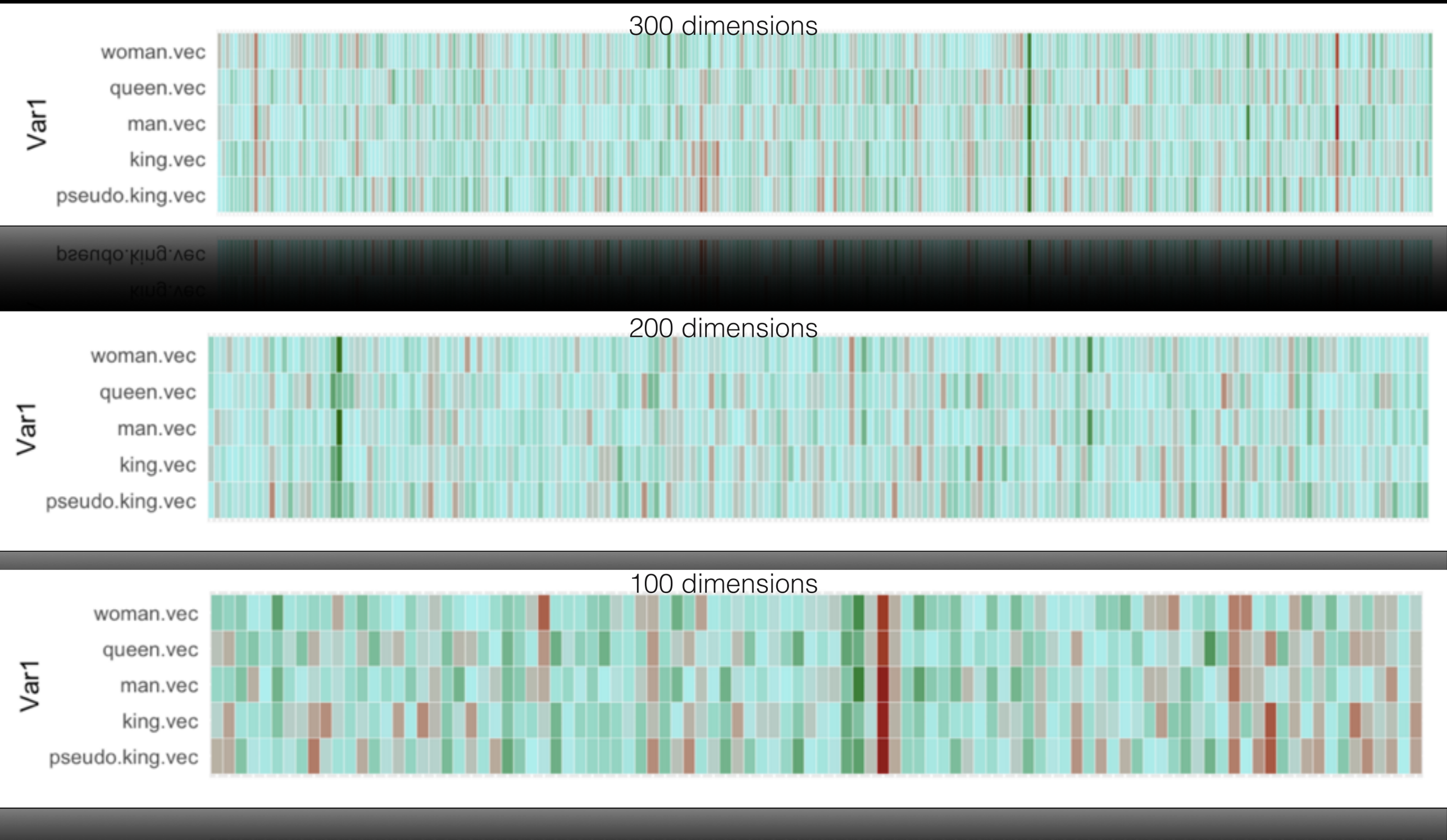
Vector Representations of Words: Using pre-computed GloVe Word Vectors

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib]

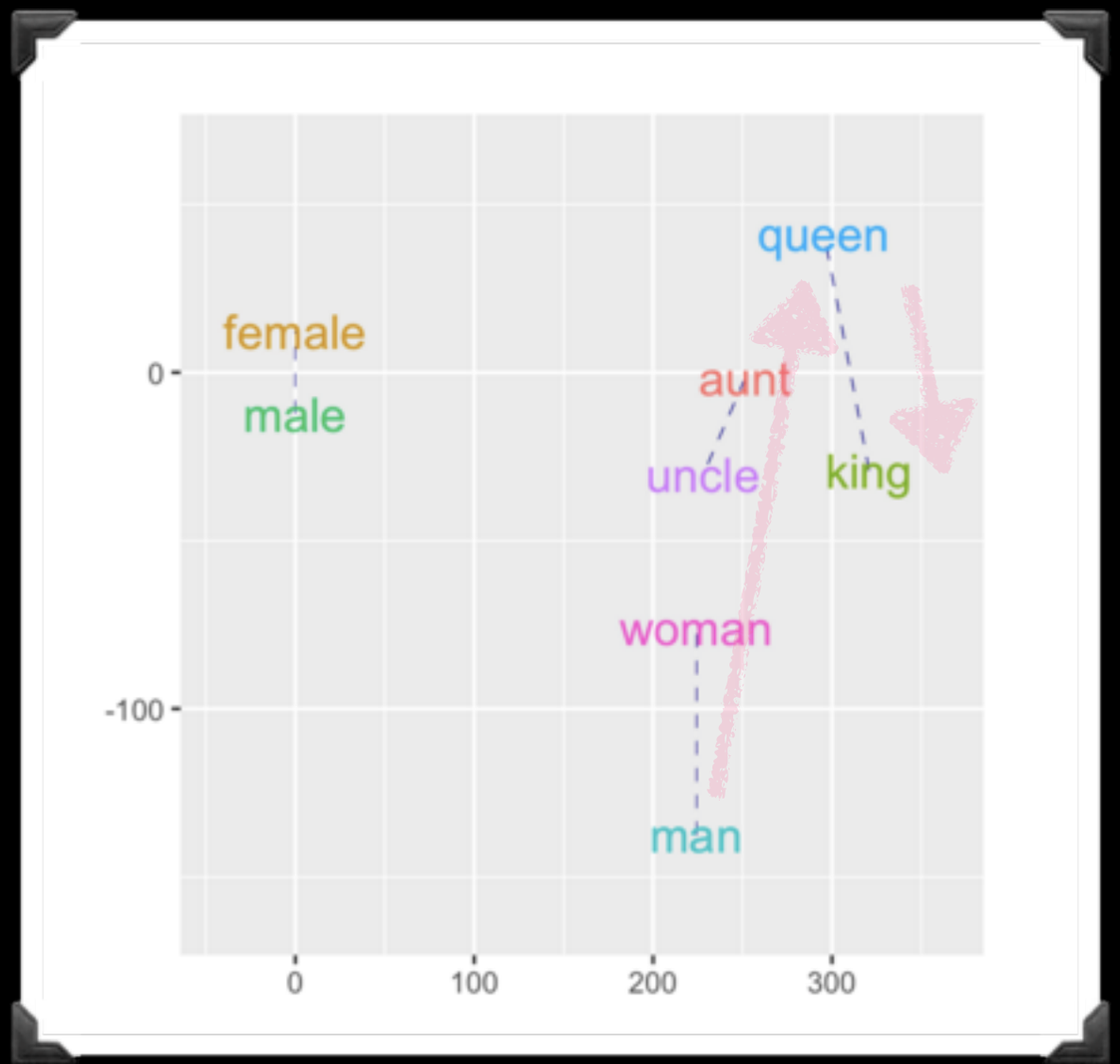
- Computed from 6Billion Tokens and a 400k word vocabulary. (other, larger sets also available)
- The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.
- This objective associates ratios of co-occurrence probabilities with vector differences in the word vector space.
- For this reason, the resulting word vectors perform very well on word analogy tasks, such as those examined in the word2vec package.

Word Vectors?

pseudo-king ~ man + queen - woman

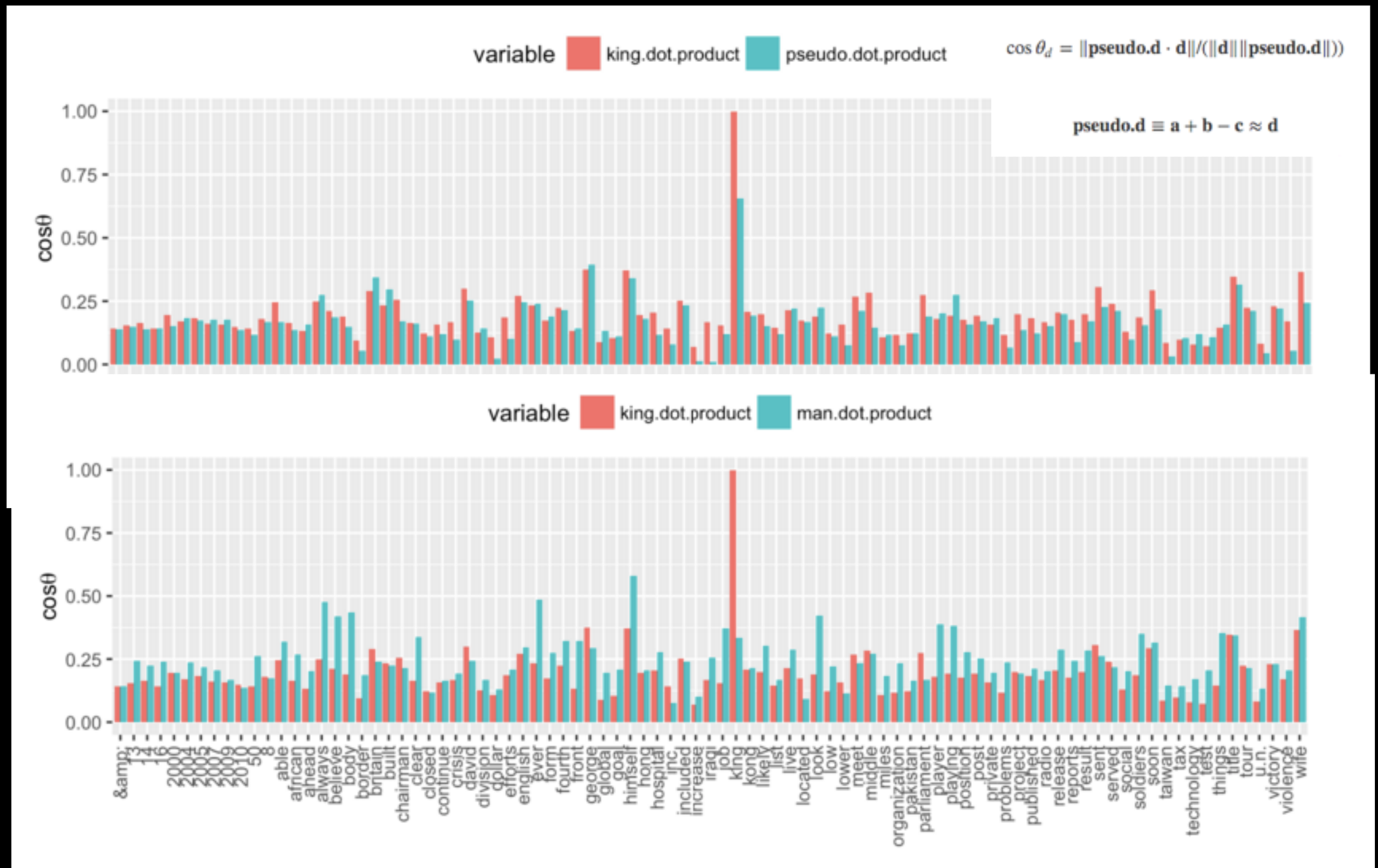


Word analogies



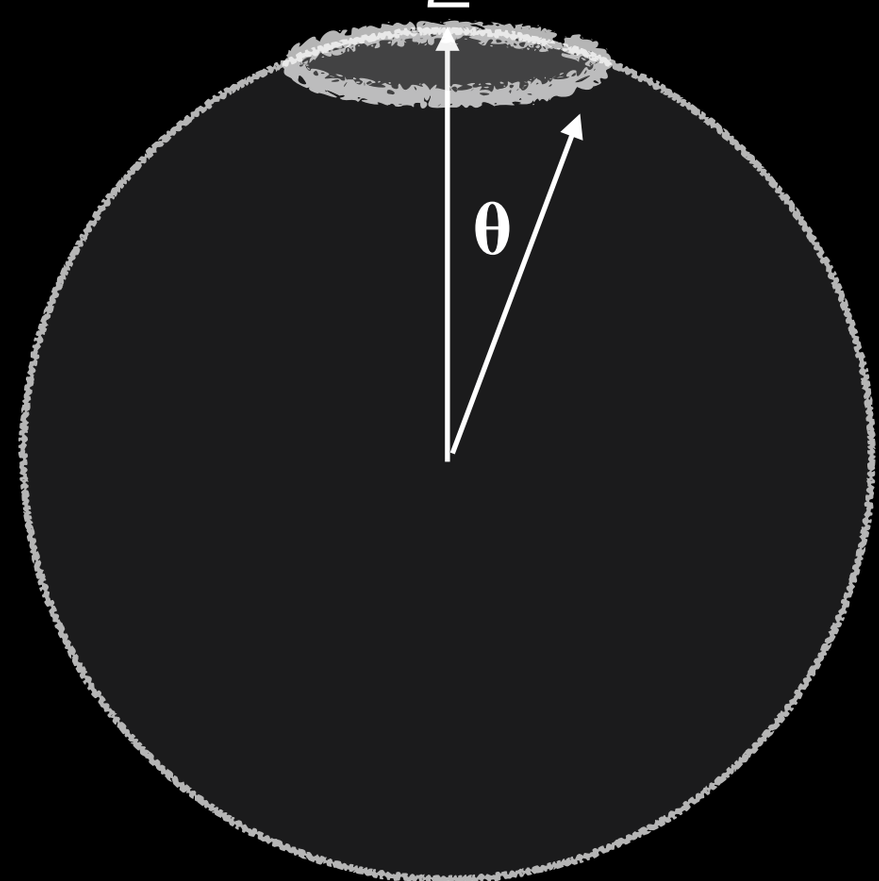
pseudo-king ~ man + queen - woman

Scalar Products



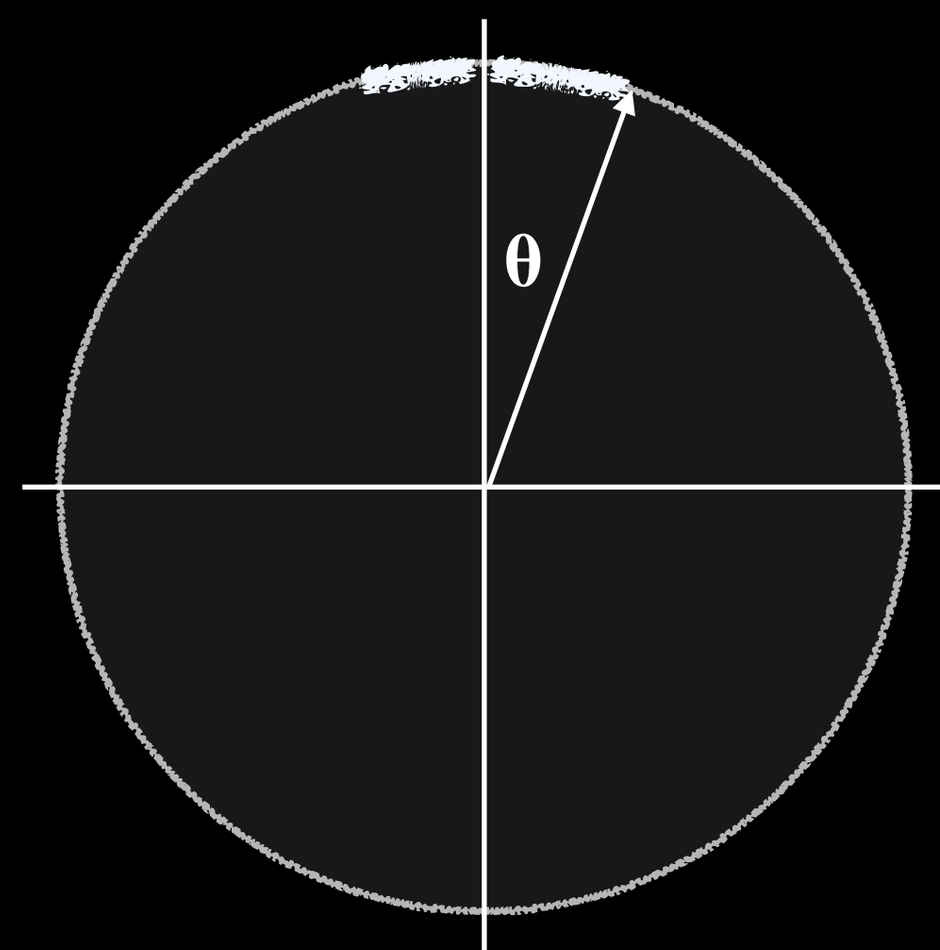
n-Dimensions

$$\sim \frac{\theta^n(2n+1)}{2^{2n}}$$



n-sphere

$$\sim \theta/\mathrm{pi}$$



2-sphere

Diving into Politics

When I was a boy I was told that anybody could become President; I'm beginning to believe it. ~Clarence Darrow

Candidate Cosines

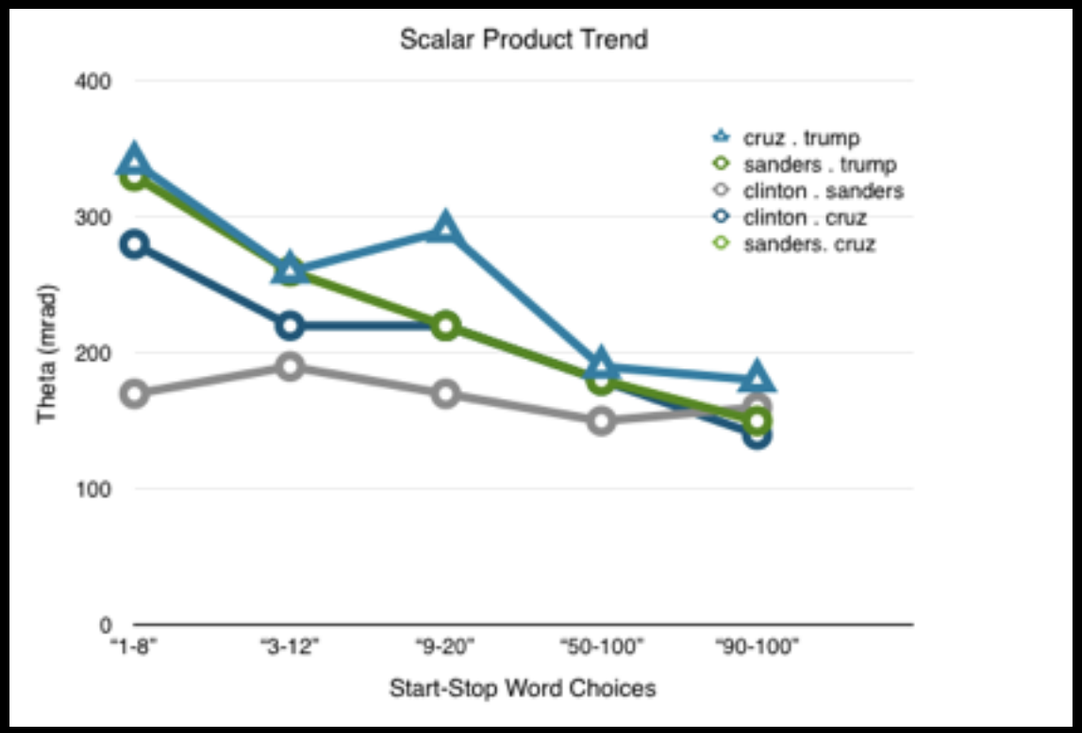
CANDIDATE "BAG OF WORDS" COSINES

We can get a more quantitative comparison between candidates by taking the scalar product of the normalized candidate vectors. The angle θ expressed in milliradians. While the angular deviations are relatively small, they follow the pattern one might expect. For example, reatlively close alignment is found between Trump and Rubio and between Sanders and Clinton. Deviations do not follow strict party lines as the angle between Cruz and Trump is the largest in the table.

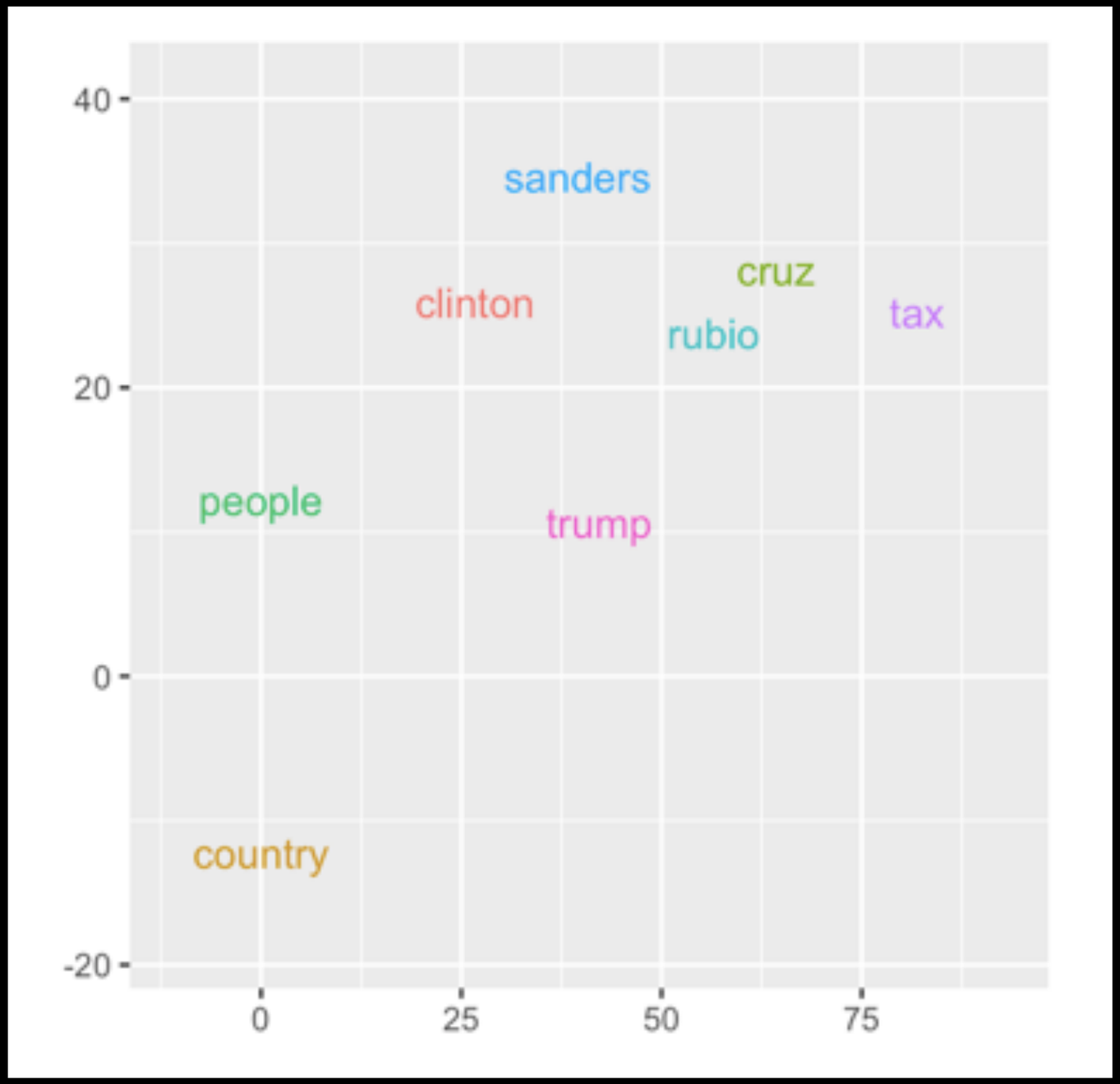
a	b	$\cos(\theta)$	θ (mrad)
cruz	trump	0.93	260
sanders	trump	0.93	260
clinton	sanders	0.96	190
clinton	cruz	0.95	220
sanders	cruz	0.93	260

Candidate Vectors

300d
all contexts
nf=3-12

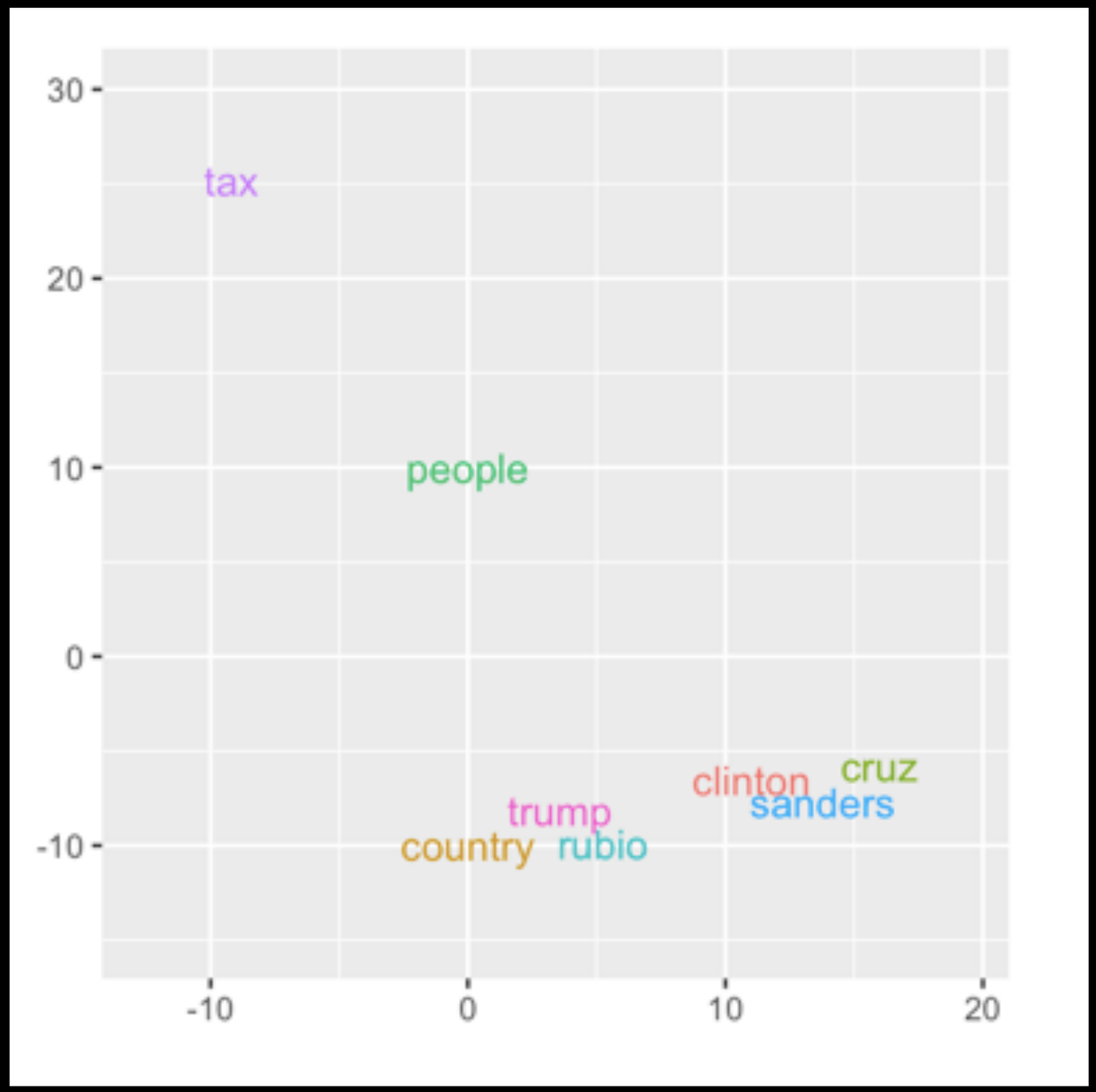


Candidate Context



Words 1 to 11
Filter *economy*
9999 to *glove.6B.300d.txt.*

Alternative View



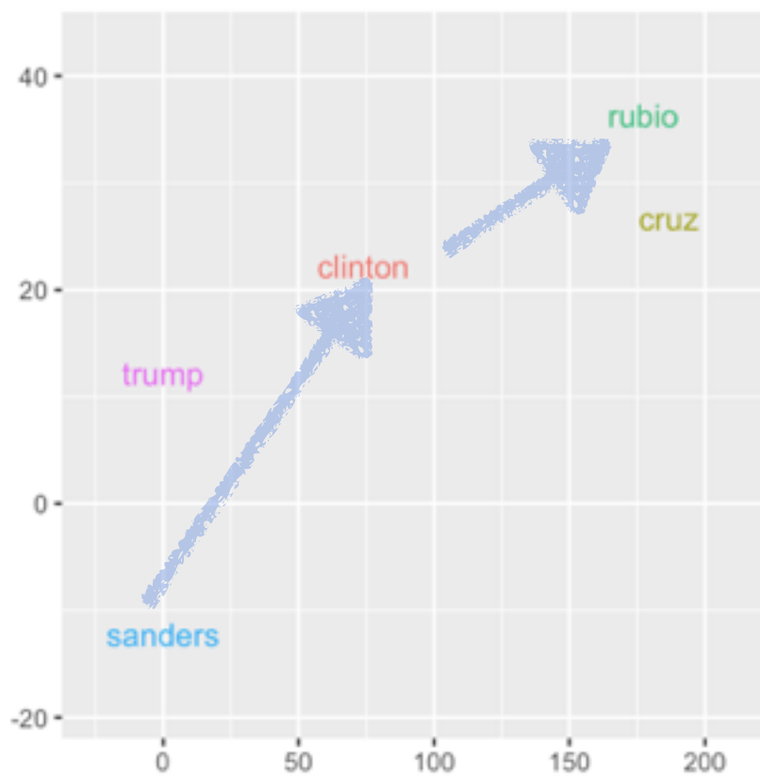
Candidate Vectors

Resulting vectors from the words 1 to 21 of the candidate "bag of words" filtered for comments containing __ and *1.210^{4}* and from *glove.6B.300d.txt*.

$$\textbf{clinton} + \textbf{sanders} - \textbf{trump} \approx \textbf{rubio}$$

and

$$\textbf{clinton} + \textbf{sanders} - \textbf{trump} \approx \textbf{cruz}$$



sanders +clinton - trump ~ rubio

Closest to “candidate”

sanders

Words 1 to 16
Filter *terror*
9999 to *glove.6B.300d.txt.*

word	theta
we	0.77
want	0.77
think	0.76
should	0.76
going	0.75
what	0.74
but	0.73
n't	0.73
because	0.73
it	0.72

…unless you dig deeper

SANDERS

Words 10 to 20
Filter *isis*
9999 to *glove.6B.300d.txt.*

word	theta
troops	0.76
forces	0.74
people	0.72
they	0.72
them	0.71
soldiers	0.70
all	0.69
so	0.69
but	0.69
only	0.68

trump_text “^wall$”

[1] "as far as the wall is concerned, were going to build a wall were going to create a border were going to let people in, but theyre going to come in legally theyre going to come in legally and its something that can be done, and i get questioned about that they built the great wall of china thats 13,000 miles here, we actually need 1,000 because we have natural barriers so we need 1,000"

[2] "we can do a wall were going to have a big, fat beautiful door right in the middle of the wall were going to have people come in, but theyre coming in legally and mexicos going to pay for the wall because mexico — i love the mexican people; i respect the mexican leaders — but the leaders are much sharper, smarter and more cunning than our leaders"

[3] "we are not talking about isolation were talking about security were not talking about religion were talking about security our country is out of control people are pouring across the southern border i will build a wall it will be a great wall people will not come in unless they come in legally drugs will not pour through that wall"

Wall

TRUMP - CRUZ

Words 5 to 25
Filter *wall*
9999 to *glove.6B.300d.txt.*

	word	dot.product
1	stuff	0.47
2	border	-0.45
3	refugees	-0.45
4	guys	0.42
5	sudanese	-0.42
6	anymore	0.40
7	pretty	0.39
8	refugee	-0.39
9	'm	0.39
10	rwandan	-0.38

Conclusions

- Candidate speech strongly distinguished by “bags of words” - both frequency and vocabulary, in some cases.
- Pre-computed word vectors == cool.
- Significant “tooling” required to manipulate vectors. R works great for that.
- Differences in speech detectable. Deconstructing political speech appears feasible.

Next Steps

- Continue exploring code.
- AI ???