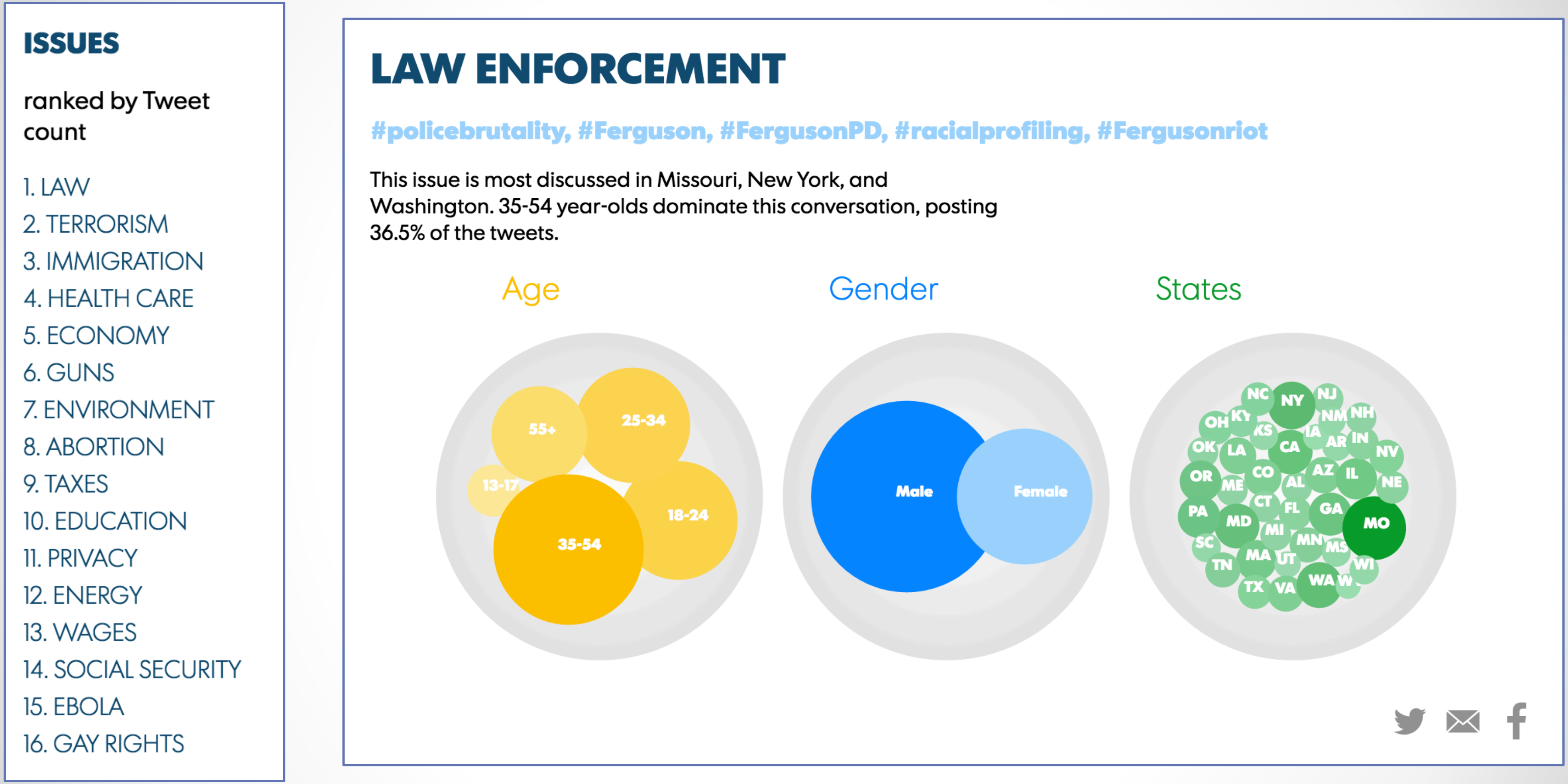
Algorithms week 2-1:
Text Analysis

Jonathan Stray
Columbia Lede Program
July 23, 2018

Text Analysis in Journalism



ISSUES

ranked by Tweet count

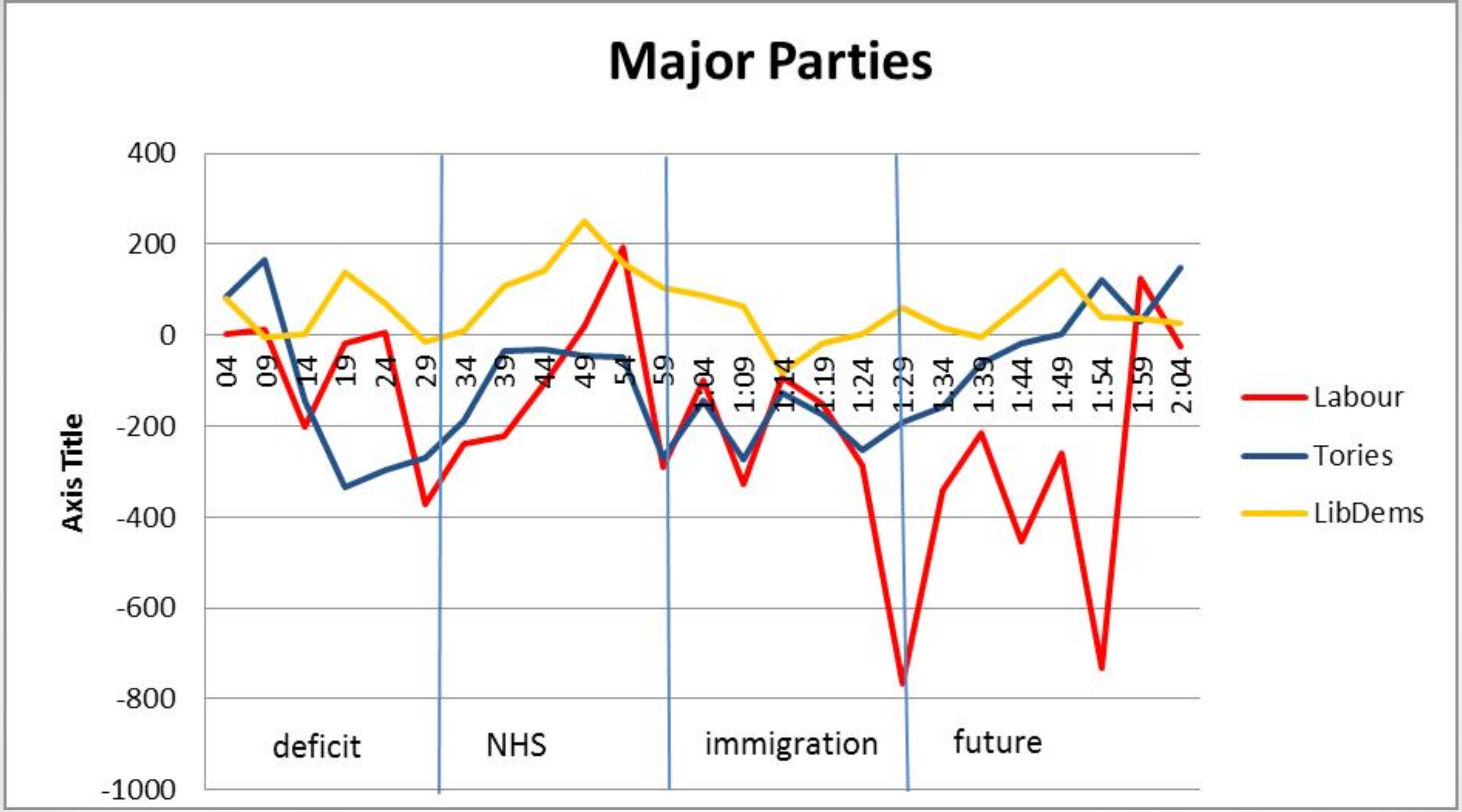
1. LAW
2. TERRORISM
3. IMMIGRATION
4. HEALTH CARE
5. ECONOMY
6. GUNS
7. ENVIRONMENT
8. ABORTION
9. TAXES
10. EDUCATION
11. PRIVACY
12. ENERGY
13. WAGES
14. SOCIAL SECURITY
15. EBOLA
16. GAY RIGHTS

LAW ENFORCEMENT

#policebrutality, #Ferguson, #FergusonPD, #racialprofiling, #Fergusonriot

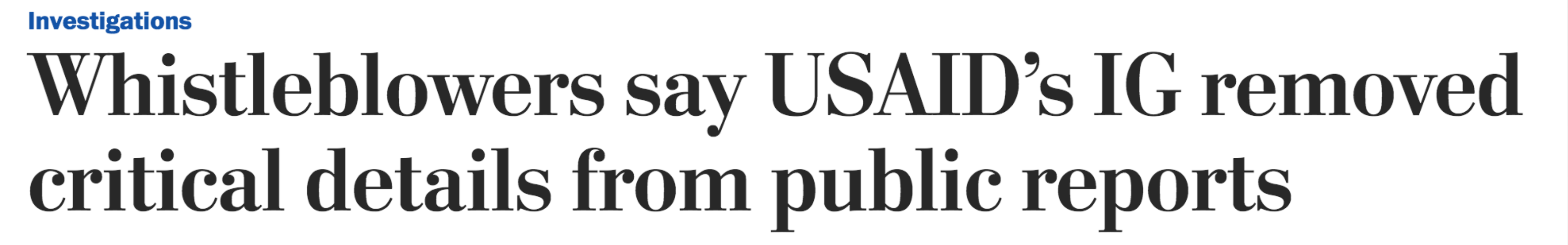
This issue is most discussed in Missouri, New York, and Washington. 35-54 year-olds dominate this conversation, posting 36.5% of the tweets.

USA Today/Twitter Political Issues Index



Twitter sentiment index

Post-match analysis of public attitudes on Twitter, University of Reading, 2015



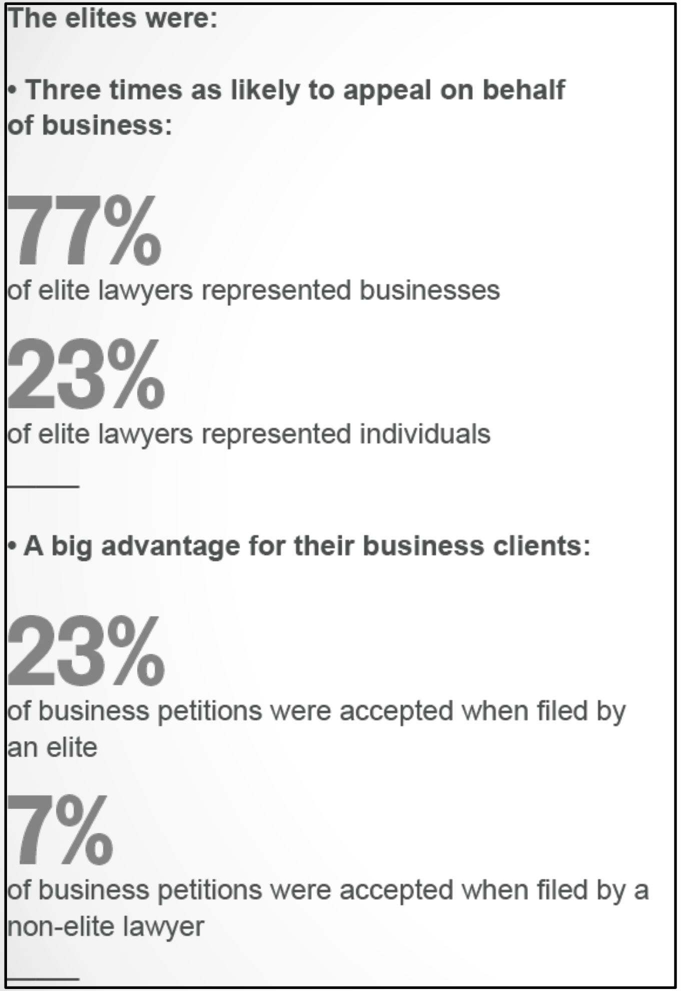
Investigations

Whistleblowers say USAID's IG removed critical details from public reports

The Post obtained draft versions of 12 audits by the inspector general's office, covering projects from the Caribbean to Pakistan to the Republic of Georgia between 2011 and 2013. The drafts are confidential and rarely become public. The Post compared the drafts with the final reports published by the inspector general's office and interviewed former and current employees. E-mails and other internal records also were reviewed.

The Post tracked changes in the language that auditors used to describe USAID and its mission offices. The analysis found that more than 400 negative references were removed from the audits between the draft and final versions.

Sentiment analysis used by *Washington Post*, 2014



The elites were:

- **Three times as likely to appeal on behalf of business:**

77%

of elite lawyers represented businesses

23%

of elite lawyers represented individuals

- **A big advantage for their business clients:**

23%

of business petitions were accepted when filed by an elite

7%

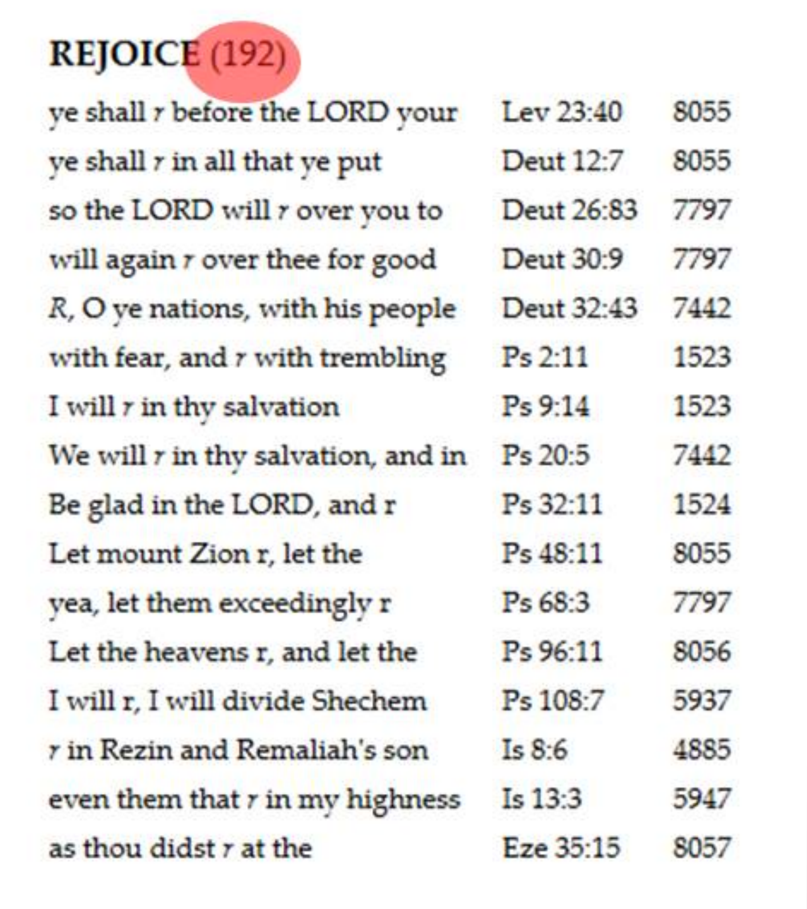
of business petitions were accepted when filed by a non-elite lawyer

We used a machine-learning method known as latent Dirichlet allocation to identify the topics in all 14,400 petitions and to then categorize the briefs. This enabled us to identify which lawyers did which kind of work for which sorts of petitioners. For example, in cases where workers sue their employers, the lawyers most successful getting cases before the court were far more likely to represent the employers rather than the employees.

The Echo Chamber, Reuters

Counting Words

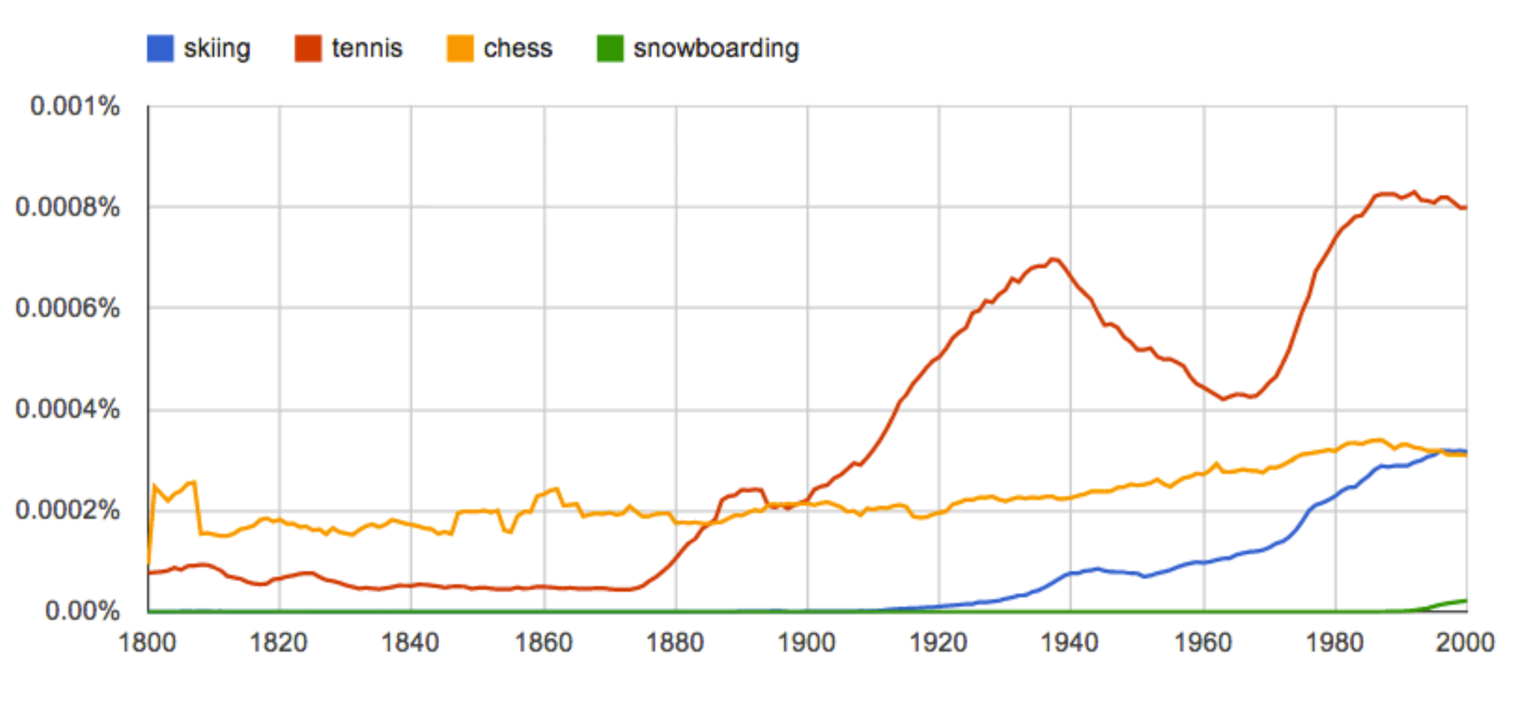
Concordance: classical word counting



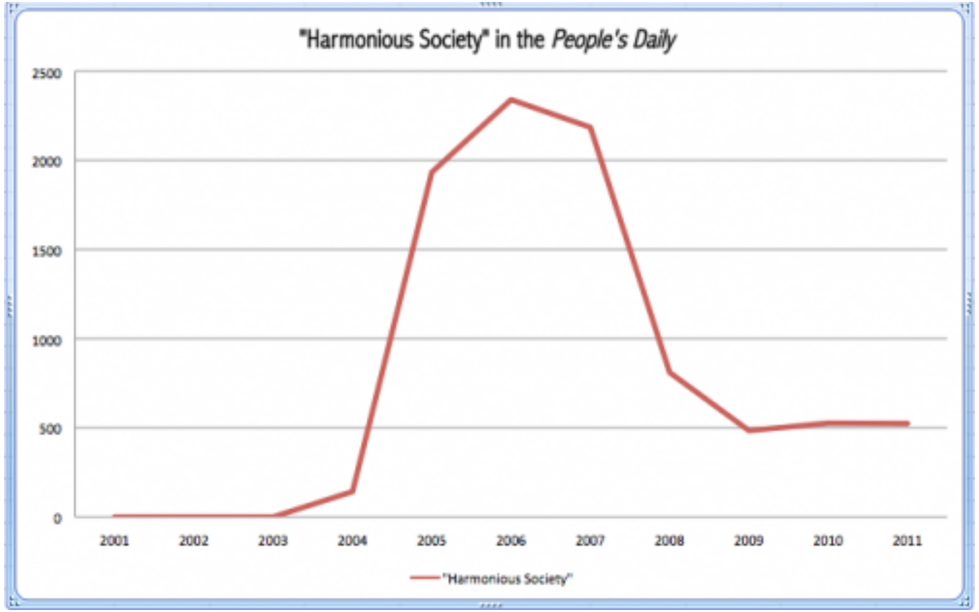
REJOICE (192)

ye shall *r* before the LORD your	Lev 23:40	8055
ye shall *r* in all that ye put	Deut 12:7	8055
so the LORD will *r* over you to	Deut 26:83	7797
will again *r* over thee for good	Deut 30:9	7797
R, O ye nations, with his people	Deut 32:43	7442
with fear, and *r* with trembling	Ps 2:11	1523
I will *r* in thy salvation	Ps 9:14	1523
We will *r* in thy salvation, and in	Ps 20:5	7442
Be glad in the LORD, and r	Ps 32:11	1524
Let mount Zion r, let the	Ps 48:11	8055
yea, let them exceedingly r	Ps 68:3	7797
Let the heavens r, and let the	Ps 96:11	8056
I will r, I will divide Shechem	Ps 108:7	5937
r in Rezin and Remaliah's son	Is 8:6	4885
even them that *r* in my highness	Is 13:3	5947
as thou didst *r* at the	Eze 35:15	8057

Google ngram viewer
12% of all English books



Stories from counting



When Hu Jintao came to power in 2002, China was already experiencing a worsening social crisis. In 2004, President Hu offered a rhetorical response to growing internal instability, trumpeting what he called a "harmonious society." For some time, this new watchword burgeoned, becoming visible everywhere in the Party's propaganda.

- Qian Gang, *Watchwords: Reading China through its Party Vocabulary*

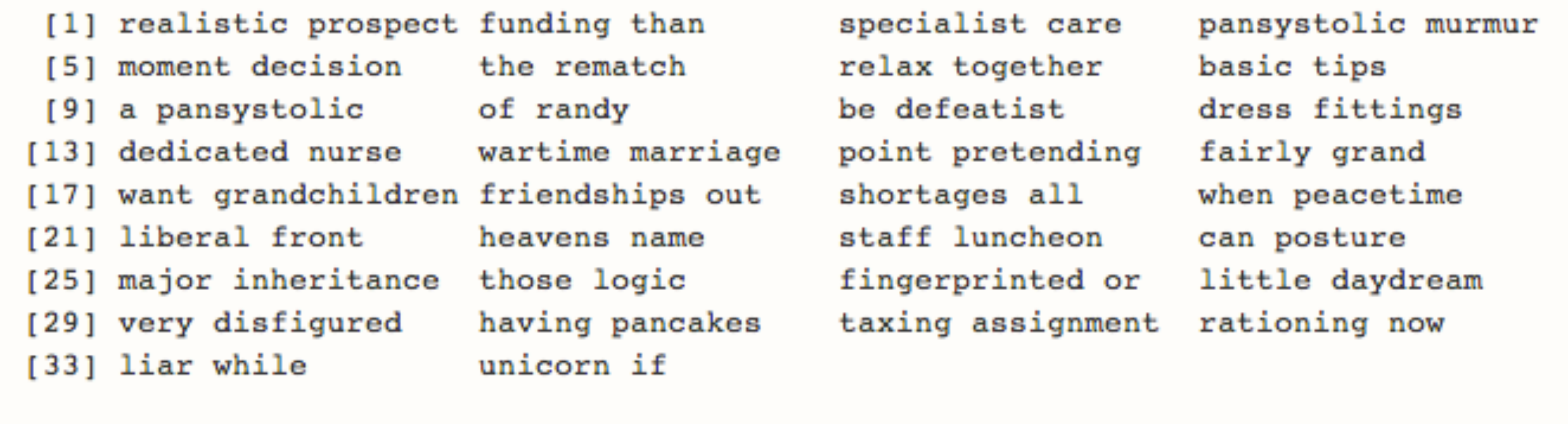
Data can give a wider view

Let me talk about Downton Abbey for a minute. The show's popularity has led many nitpickers to draft up lists of mistakes. ... But all of these have relied, so far as I can tell, on finding a phrase or two that sounds a bit off, and checking the online sources for earliest use.

I lack such social graces. So I thought: why not just check every single line in the show for historical accuracy? ... So I found some copies of the Downton Abbey scripts online, and fed every single two-word phrase through the Google Ngram database to see how characteristic of the English Language, c. 1917, Downton Abbey really is.

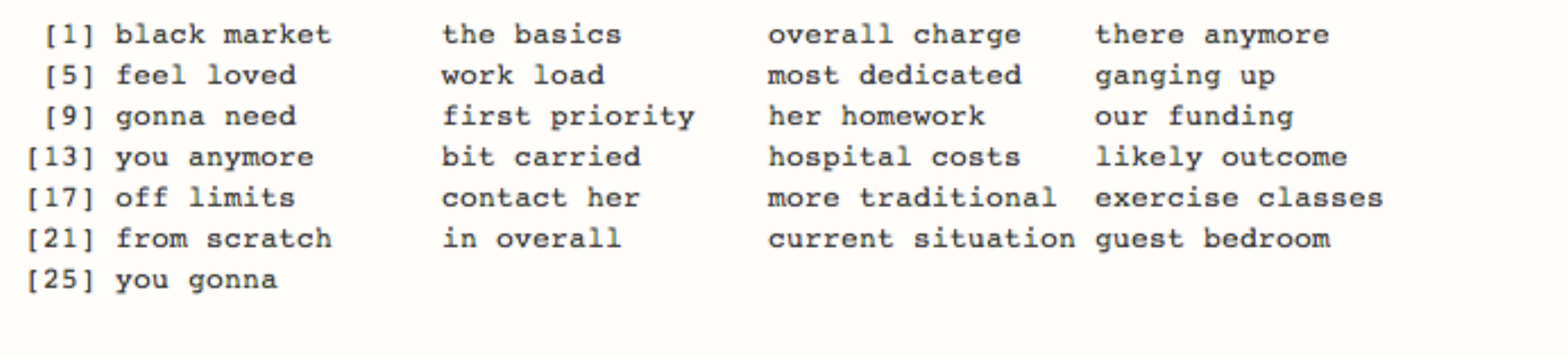
- - Ben Schmidt, *Making Downton more traditional*
-

Bigrams that do not appear in English books between 1912 and 1921.

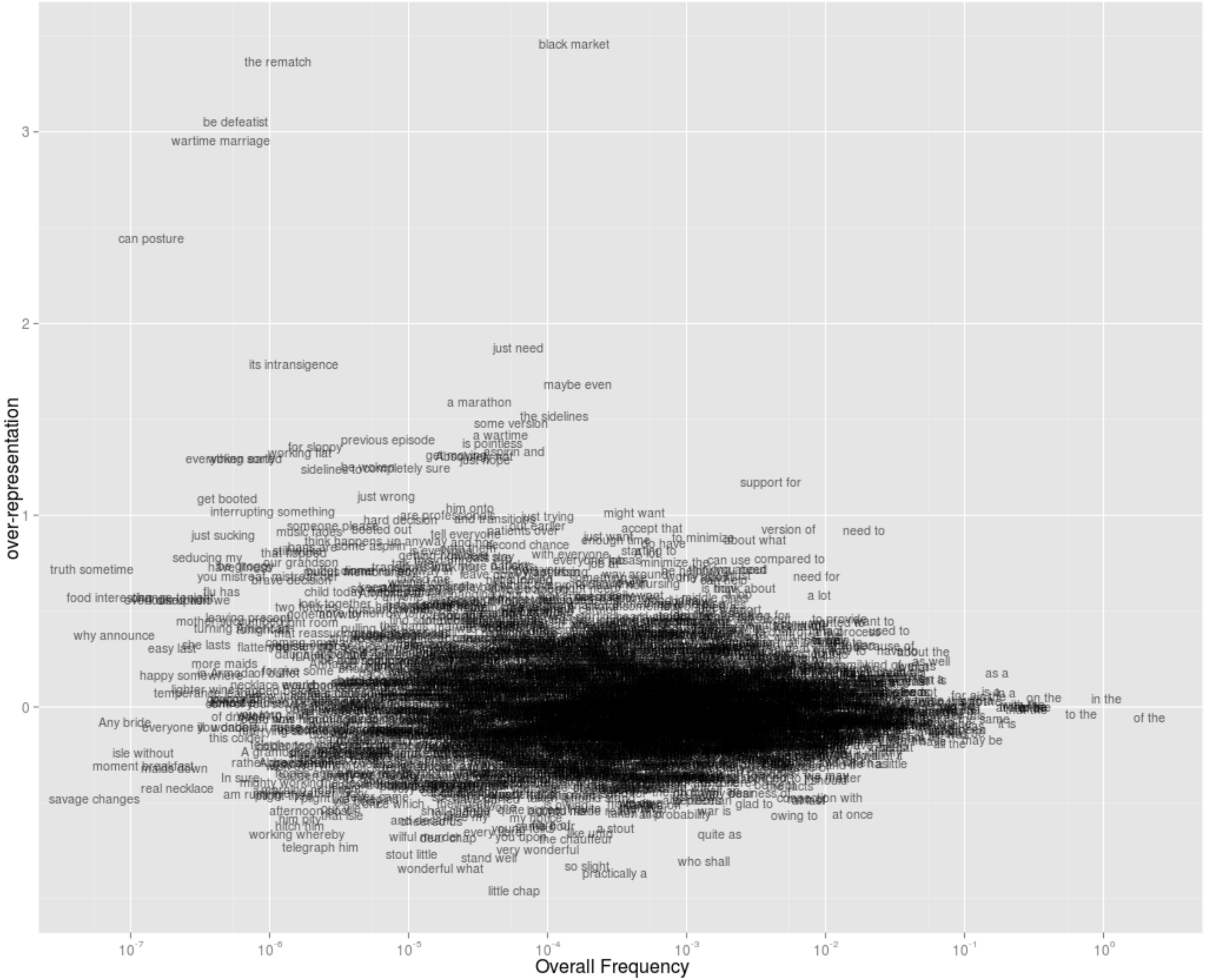


```
 [1] realistic prospect funding than       specialist care    pansystolic murmur
 [5] moment decision    the rematch        relax together     basic tips
 [9] a pansystolic      of randy           be defeatist       dress fittings
[13] dedicated nurse    wartime marriage   point pretending   fairly grand
[17] want grandchildren friendships out    shortages all      when peacetime
[21] liberal front      heavens name       staff luncheon     can posture
[25] major inheritance  those logic        fingerprinted or   little daydream
[29] very disfigured    having pancakes    taxing assignment  rationing now
[33] liar while         unicorn if
```

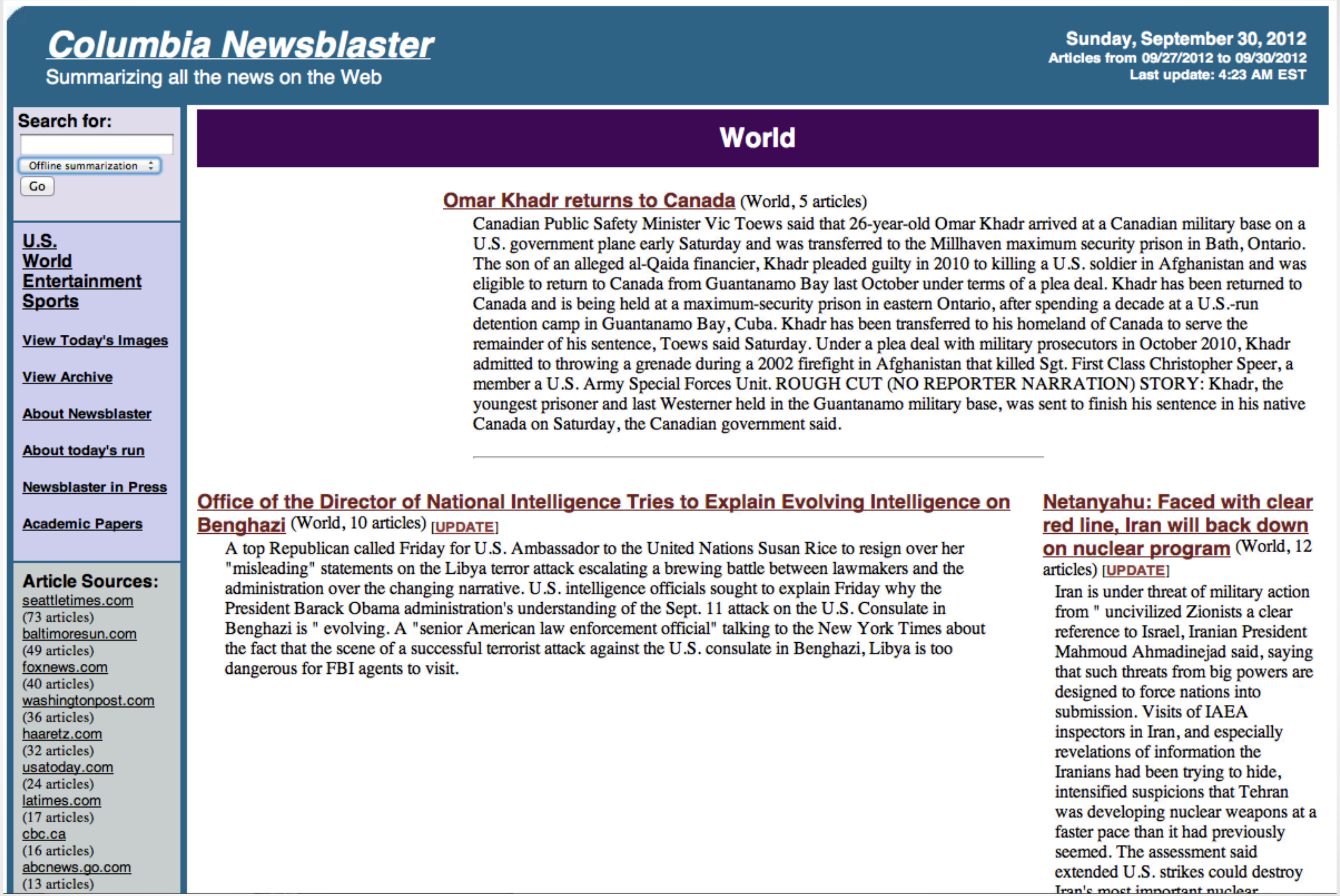
Bigrams that are at least 100 times more common today than they were in 1912-1921



```
 [1] black market      the basics        overall charge     there anymore
 [5] feel loved        work load         most dedicated     ganging up
 [9] gonna need        first priority    her homework       our funding
[13] you anymore       bit carried       hospital costs     likely outcome
[17] off limits        contact her       more traditional   exercise classes
[21] from scratch      in overall        current situation  guest bedroom
[25] you gonna
```


over-representation
Overall Frequency
3
2
1
0
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
black market
the rematch
be defeatist
wartime marriage
can posture
just need
its intransigence
maybe even
a marathon
the sidelines
some version
previous episode
a wartime
is pointless
for sloppy
working flat
aspirin and
just hope
completely sure
support for
get booted
just wrong
interrupting something
him onto
just trying
might want
accept that
version of
need to
just sucking
booted out
second chance
to minimize
about what
truth sometime
seducing my
our grandson
with everyone
can use
compared to
need for
brave decision
flu has
child today
about
a lot
food interest
why announce
easy last
she lasts
more maids
happy somewhere
to provide
used to
want to
about the
as well
as a
on the
in the
to the
of the
Any bride
everyone ill
this cold
isle without
moment breakfast
maids down
In sure
real necklace
savage changes
am running
afternoon people
him pity
hitch him
working whereby
telegraph him
stout little
wonderful what
stand well
little chap
wilful murder
dear chap
very wonderful
the chauffeur
so slight
practically a
who shall
quite as
a stout
war is
glad to
owing to
at once
connection with
all probability
may be
it is
all the
a little

Document Vector Space Model



Columbia Newsblaster

Summarizing all the news on the Web

Sunday, September 30, 2012
Articles from 09/27/2012 to 09/30/2012
Last update: 4:23 AM EST

Search for:

U.S.
World
Entertainment
Sports

View Today's Images

View Archive

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

Article Sources:

seattletimes.com	(73 articles)
baltimoresun.com	(49 articles)
foxnews.com	(40 articles)
washingtonpost.com	(36 articles)
haaretz.com	(32 articles)
usatoday.com	(24 articles)
latimes.com	(17 articles)
cbc.ca	(16 articles)
abcnews.go.com	(13 articles)

World

Omar Khadr returns to Canada (World, 5 articles)

Canadian Public Safety Minister Vic Toews said that 26-year-old Omar Khadr arrived at a Canadian military base on a U.S. government plane early Saturday and was transferred to the Millhaven maximum security prison in Bath, Ontario. The son of an alleged al-Qaida financier, Khadr pleaded guilty in 2010 to killing a U.S. soldier in Afghanistan and was eligible to return to Canada from Guantanamo Bay last October under terms of a plea deal. Khadr has been returned to Canada and is being held at a maximum-security prison in eastern Ontario, after spending a decade at a U.S.-run detention camp in Guantanamo Bay, Cuba. Khadr has been transferred to his homeland of Canada to serve the remainder of his sentence, Toews said Saturday. Under a plea deal with military prosecutors in October 2010, Khadr admitted to throwing a grenade during a 2002 firefight in Afghanistan that killed Sgt. First Class Christopher Speer, a member a U.S. Army Special Forces Unit. ROUGH CUT (NO REPORTER NARRATION) STORY: Khadr, the youngest prisoner and last Westerner held in the Guantanamo military base, was sent to finish his sentence in his native Canada on Saturday, the Canadian government said.

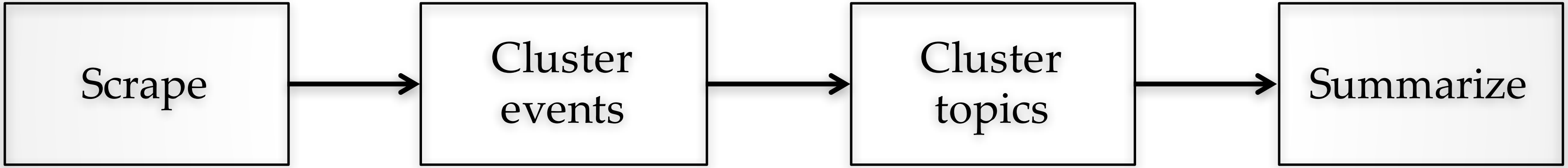
Office of the Director of National Intelligence Tries to Explain Evolving Intelligence on Benghazi (World, 10 articles) [UPDATE]

A top Republican called Friday for U.S. Ambassador to the United Nations Susan Rice to resign over her "misleading" statements on the Libya terror attack escalating a brewing battle between lawmakers and the administration over the changing narrative. U.S. intelligence officials sought to explain Friday why the President Barack Obama administration's understanding of the Sept. 11 attack on the U.S. Consulate in Benghazi is " evolving. A "senior American law enforcement official" talking to the New York Times about the fact that the scene of a successful terrorist attack against the U.S. consulate in Benghazi, Libya is too dangerous for FBI agents to visit.

Netanyahu: Faced with clear red line, Iran will back down on nuclear program (World, 12 articles) [UPDATE]

Iran is under threat of military action from " uncivilized Zionists a clear reference to Israel, Iranian President Mahmoud Ahmadinejad said, saying that such threats from big powers are designed to force nations into submission. Visits of IAEA inspectors in Iran, and especially revelations of information the Iranians had been trying to hide, intensified suspicions that Tehran was developing nuclear weapons at a faster pace than it had previously seemed. The assessment said extended U.S. strikes could destroy Iran's most important nuclear

System Description



Scrape

Cluster events

Cluster topics

Summarize

What is this document "about"?

Most commonly occurring words a pretty good indicator.

```
30  the
23  to
19  and
19  a
18  animal
17  cruelty
15  of
15  crimes
14  in
14  for
11  that
8   crime
7   we
```

animal
cruelty
crimes
crime
Enforcement
National
Dogfighting
Law
Reporting
Menendez
violent
Senator
Michael
report
Program
Establish
animals
criminal
legislation
prevent
category
Humane
separate
important
strong
people
annual
year
activity
offenders
commit
shown
found
problem
making
federal
Direct
help
need
state
FBI
Vick
violence
public
System
agencies
trends
understand
societal
track
statistics
policy
introduced
Uniform
patterns
gang
officials
makes
market
consequences
lucrative
brought
gambling
adding
director
await
proven
everyone
overdue
highprofile
monitor
gathered
impossible
America
twelve
percent
Attorney
involvement
General
Specifically
collect
much
tool
linked
date
strongly
lining
stop
quarterback
country
grateful
aware
cruel
allocate
become
Datatoexchange
Studies
creating
effectively
truly
make
Act
23
rates
sure
long
going
drug
United
made
incidentbased
fight
Association
single
incident
repulsiveness
communities
security
executive
gives
Perhaps
included
yearly
thrilled
months
safety
bloodsport
escalating
Bureau
drugs
policies
Jersey
finally
momentum
ring
dog
last
aid
repulsive
give
silver
bunker
anticrime
add
difficult
severe
local
Washington
period
month
abuse
behavior
experts
episode
politics
States
Based
raid
along
sake
forms
tools
include
DNJ
dogs
also
scope
Society
New
59
showing
cases
NFL
underground
currently
tracked
bill
ability
today
recurrence
combating
selling
president
terrible
propensity
efficiently
established
exceeding
factually
illegal
champion
ultimately
selfadmitted
Markarian
critical
Chicago
Police
gangs
Department
$25000
$100000
preventing
assigned
closely
across
American
participate
implementation
enactment
prosecuted
consultation
window
accepting
information
resources
members
pastime
enable
Research
black
battle
support
investigation
makers
trying
prison
Tracking
anywhere
overall
sentenced
FBI's
bets
interconnected
address
precursor
introducing
enterprise
study
disaggregate

Map documents to vectors

Encode each document as the list of words it contains.

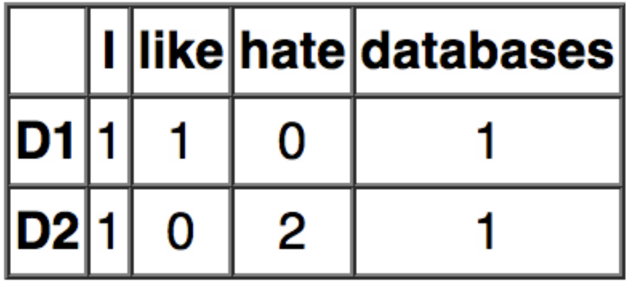
Dimensions = vocabulary of document set.

Value on each dimension = # of times word appears in document

Example

D1 = “I like databases”

D2 = “I hate hate databases”



	I	like	hate	databases
D1	1	1	0	1
D2	1	0	2	1

Each row = document vector

All rows = term-document matrix

Individual entry = tf(t,d) = “term frequency”

Aka “Bag of words” model

Throws out word order.

e.g. “soldiers shot civilians” and “civilians shot soldiers” encoded identically.

Tokenization

The documents come to us as long strings, not individual words. Tokenization is the process of converting the string into individual words, or "tokens."

For this course, we will assume a very simple strategy:

- convert all letters to lowercase
- remove all punctuation characters
- separate words based on spaces

Note that this won't work at all for Chinese. It will fail in some ways even for English. How?

Distance function

Useful for:

- clustering documents
- finding docs similar to example
- matching a search query

Basic idea: look for overlapping terms

Cosine similarity

Given document vectors a,b define

$$similarity(a,b) \equiv a \bullet b$$

If each word occurs exactly once in each document, equivalent to counting overlapping words.

Note: *not* a distance function, as similarity *increases* when documents are… similar. (What part of the definition of a distance function is violated here?)

Problem: long documents always win

Let a = “This car runs fast.”

Let b = “My car is old. I want a new car, a shiny car”

Let query = “fast car”

	this	car	runs	fast	my	is	old	I	want	a	new	shiny
a	1	1	1	1	0	0	0	0	0	0	0	0
b	0	3	0	0	1	1	1	1	1	1	1	1
q	0	1	0	1	0	0	0	0	0	0	0	0

Problem: long documents always win

similarity(a,q) = 1*1 [car] + 1*1 [fast] = 2
similarity(b,q) = 3*1 [car] + 0*1 [fast] = 3

Longer document more “similar”, by virtue of repeating words.

Normalize document vectors

$$similarity(a,b) \equiv \frac{a \bullet b}{\|a\|\|b\|}$$

$$= \cos(\Theta)$$

returns result in [0,1]

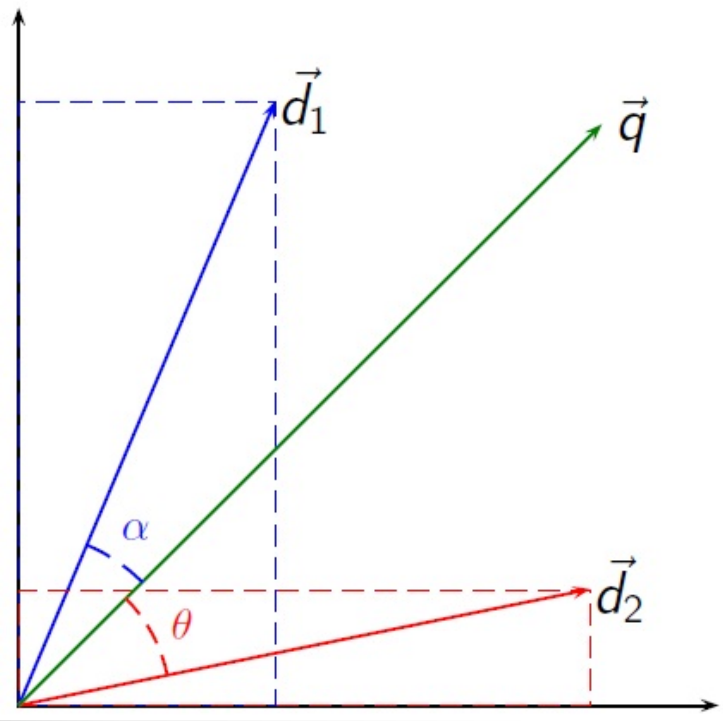
Normalized query example

	this	car	runs	fast	my	is	old	I	want	a	new	shiny
a	1	1	1	1	0	0	0	0	0	0	0	0
b	0	3	0	0	1	1	1	1	1	1	1	1
q	0	1	0	1	0	0	0	0	0	0	0	0

$$similarity(a,q) = \frac{2}{\sqrt{4}\sqrt{2}} = \frac{1}{\sqrt{2}} \approx 0.707$$

$$similarity(b,q) = \frac{3}{\sqrt{17}\sqrt{2}} \approx 0.514$$

Cosine similarity



$$\cos\theta = similarity(a,b) \equiv \frac{a \bullet b}{\|a\|\|b\|}$$

Cosine distance (finally)

$$dist(a,b) \equiv 1 - \frac{a \bullet b}{\|a\|\|b\|}$$

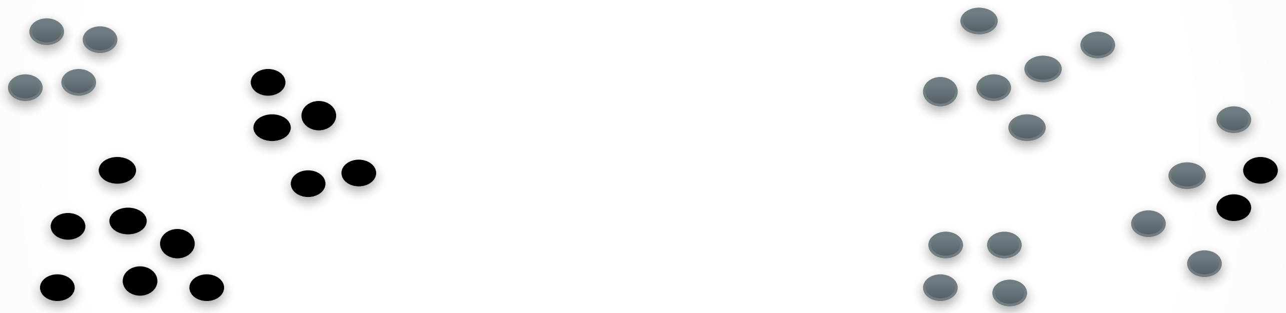
Problem: common words

We want to look at words that “discriminate” among documents.

Stopwords: if all documents contain “the,” are all documents similar?

Common words: if most documents contain “car” then car doesn’t tell us much about (contextual) similarity.

Context matters



General News

Car Reviews

◉ = contains “car”

● = does not contain “car”

Document Frequency

Idea: de-weight common words

Common = appears in many documents

$$df(t,D) = \left|d \in D : t \in d\right| / \left|D\right|$$

"document frequency" = fraction of docs containing term

Inverse Document Frequency

Invert (so more common = smaller weight) and take log

$$idf(t,D) = \log\left(|D| / |d \in D : t \in d|\right)$$

TF-IDF

Multiply term frequency by inverse document frequency

$$
\begin{aligned}
tfidf(t,d,D) &= tf(t,d) \cdot idf(d,D) \\
&= n(t,d) \cdot \log\left(|D| / n(t,D)\right)
\end{aligned}
$$

n(t,d) = number of times term t in doc d

n(t,D) = number docs in D containing t

TF-IDF depends on entire corpus

The TF-IDF vector for a document changes if we add another document to the corpus.

$$tfidf(t,d,D) = tf(t,d) \cdot idf(d,D)$$

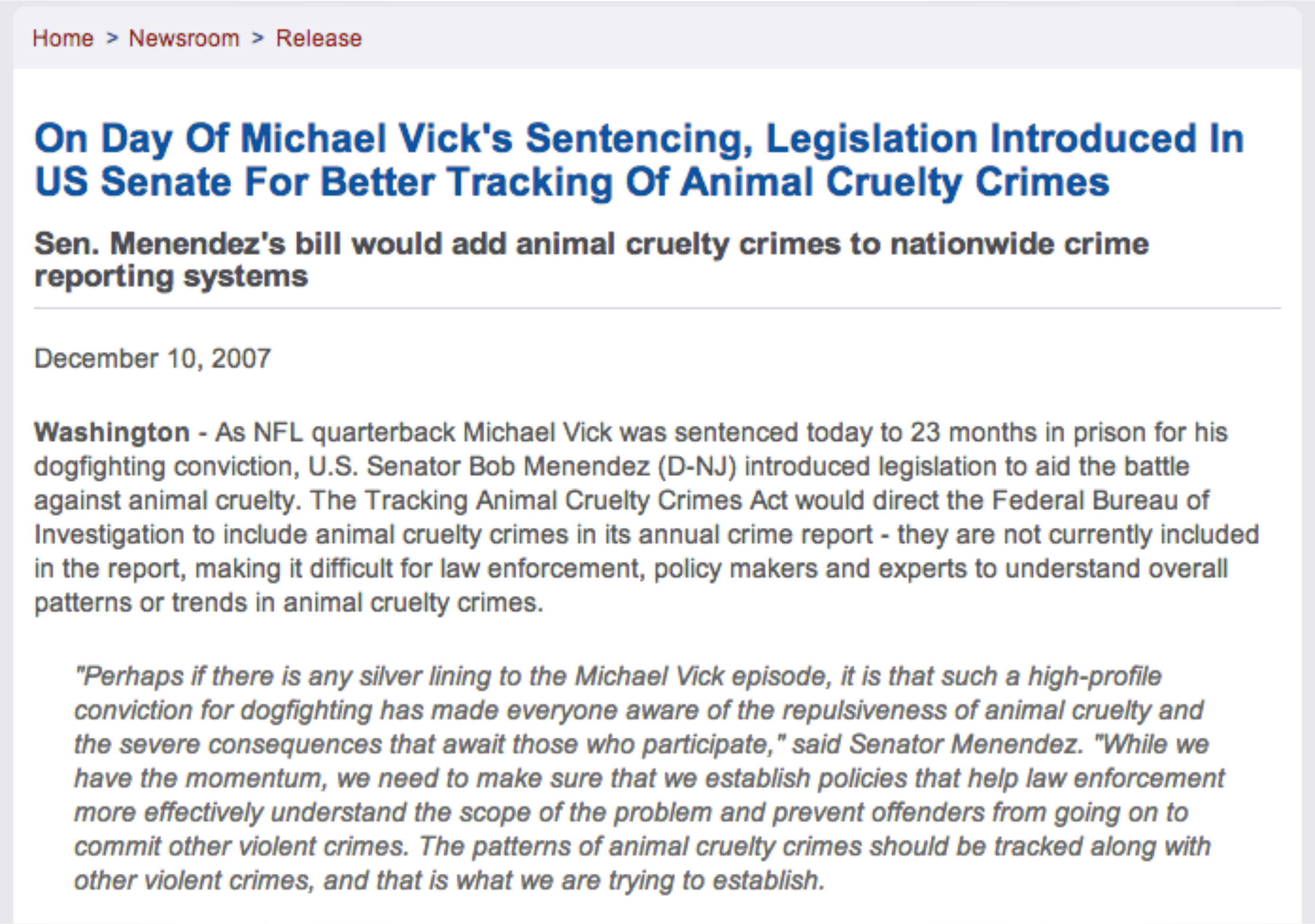
if we add a document, D changes!

TF-IDF is sensitive to *context*. The context is all other documents

What is this document "about"?

Each document is now a vector of TF-IDF scores for every word in the document. We can look at which words have the top scores.

```
crimes          0.0675591652263963
cruelty         0.0585772393867342
crime           0.0257614113616027
reporting       0.0208838148975406
animals         0.0179258756717422
michael         0.0156575858658684
category        0.0154564813388897
commit          0.0137447439653709
criminal        0.0134312894429112
societal        0.0124164973052386
trends          0.0119505837811614
conviction      0.0115699047136248
patterns        0.011248045148093
```



On Day Of Michael Vick's Sentencing, Legislation Introduced In US Senate For Better Tracking Of Animal Cruelty Crimes

Sen. Menendez's bill would add animal cruelty crimes to nationwide crime reporting systems

December 10, 2007

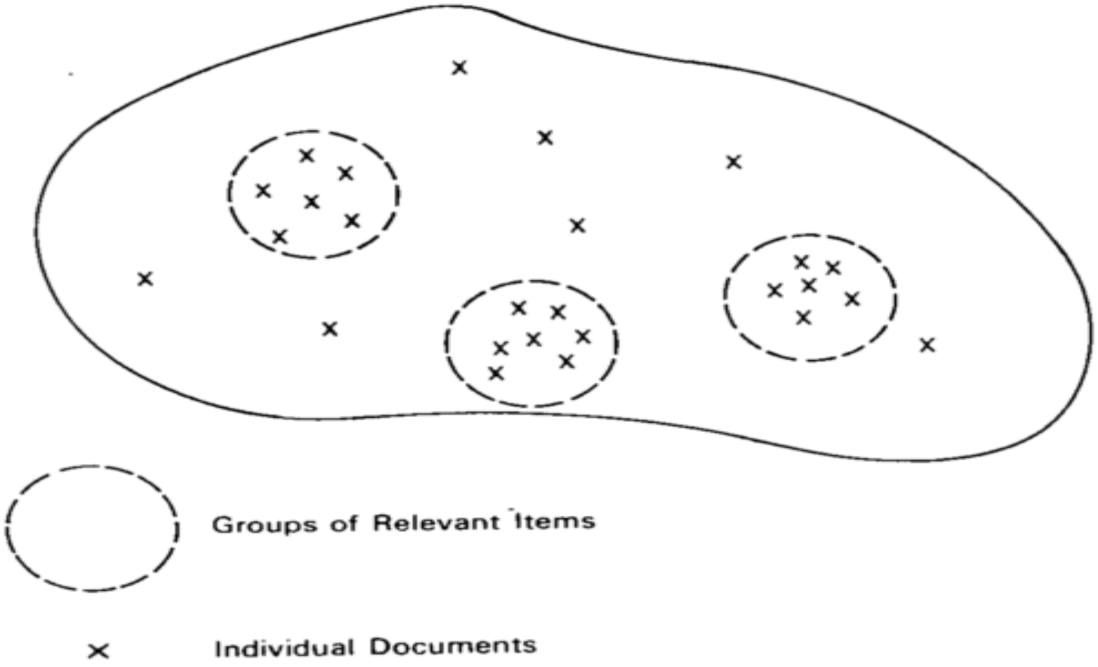
Washington - As NFL quarterback Michael Vick was sentenced today to 23 months in prison for his dogfighting conviction, U.S. Senator Bob Menendez (D-NJ) introduced legislation to aid the battle against animal cruelty. The Tracking Animal Cruelty Crimes Act would direct the Federal Bureau of Investigation to include animal cruelty crimes in its annual crime report - they are not currently included in the report, making it difficult for law enforcement, policy makers and experts to understand overall patterns or trends in animal cruelty crimes.

> *"Perhaps if there is any silver lining to the Michael Vick episode, it is that such a high-profile conviction for dogfighting has made everyone aware of the repulsiveness of animal cruelty and the severe consequences that await those who participate," said Senator Menendez. "While we have the momentum, we need to make sure that we establish policies that help law enforcement more effectively understand the scope of the problem and prevent offenders from going on to commit other violent crimes. The patterns of animal cruelty crimes should be tracked along with other violent crimes, and that is what we are trying to establish.*

animal
cruelty
crimes
crime
Menendez
Enforcement
National
Dogfighting
Law
Reporting
violent
Program
Senator
Michael
report
Establish
animals
criminal
legislation
prevent
Category
officials
offenders
commit
found
problem
making
shown
Humane
separate
strong
important
agencies
trends
statistics
policy
introduced
understand
societal
System
public
violence
Uniform
patterns
gang
need
FBI
state
help
Direct
federal
Vick
year
annual
people
activity
track
monitor
gathered
impossible
America
twelve
General
Specifically
percent
Attorney
involvement
collect
much
fool
linked
data
strongly
tools
highprofile
proven
everyone
overdue
lining
grateful
country
quarterback
step
drug
going
long
sure
rates
made
fight
United
incidentBased
Incident
Association
repulsiveness
communities
security
single
executive
included
Perhaps
gives
yearly
thrilled
safety
months
recurrence
combating
selfadmitted
selling
illegal
factually
champion
ultimately
exceeding
propensity
president
ability
today
currently
tracked
bill
escalating
bloodsport
Bureau
Act
make
truly
creating
effectively
23
give
momentum
finally
Jersey
home
Rob
3
policies
drugs
last
aid
ring
dog
repulsive
dogs
DNJ
also
silver
bunker
anticrime
add
difficult
severe
Washington
local
period
Phillips
episode
experts
month
abuse
behavior
States
Based
along
raid
sake
forms
tools
include
makers
investigation
support
trying
enterprise
study
disaggregate
battle
black
members
resources
information
pastime
Research
enable
FBI's
sentenced
overall
anywhere
Tracking
prison
fifteen
interconnected
address
bets
precursor
Markarian
critical
Introducing
Chicago
Police
gangs
established
efficiently
terrible
underground
NFL
cases
showing
59
vice
Society
New
accepting
window
consultation
prosecuted
enactment
implementation
participate
American
across
closely
assigned
preventing
$100000
Department
$25000
director
await
market
consequences
lucrative
brought
gambling
adding
aware
cruel
allocate
become
Datatochange
Studies
scope

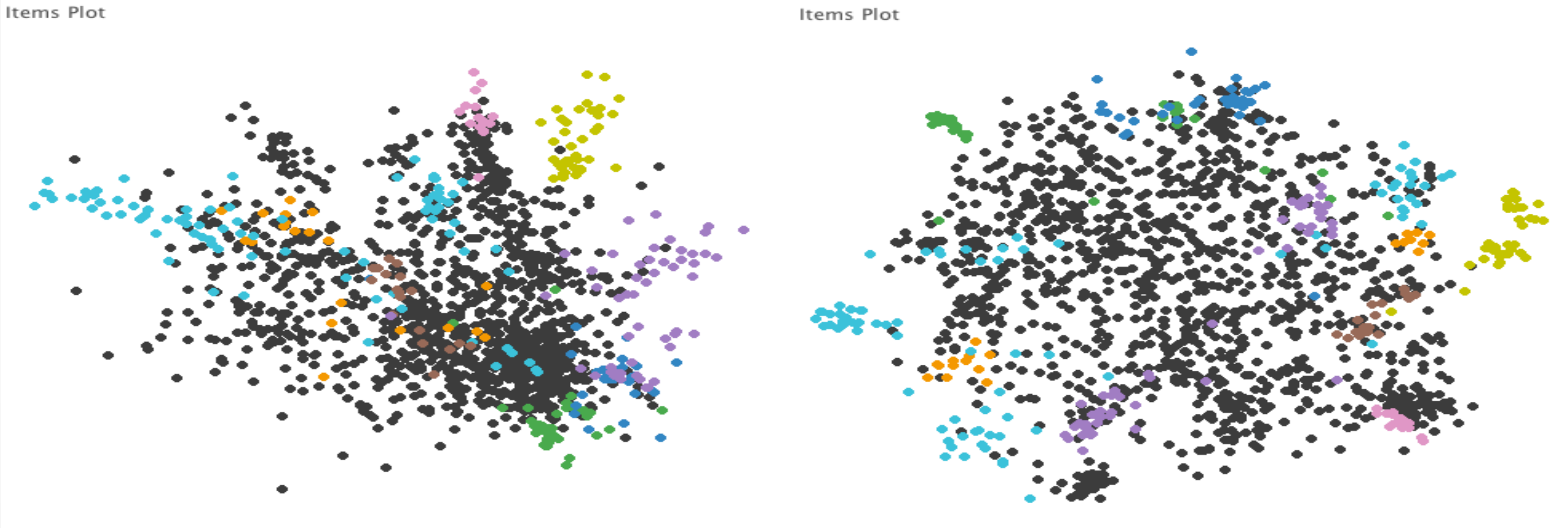
TF-IDF separates document clusters

Fig. 2. Ideal document space.



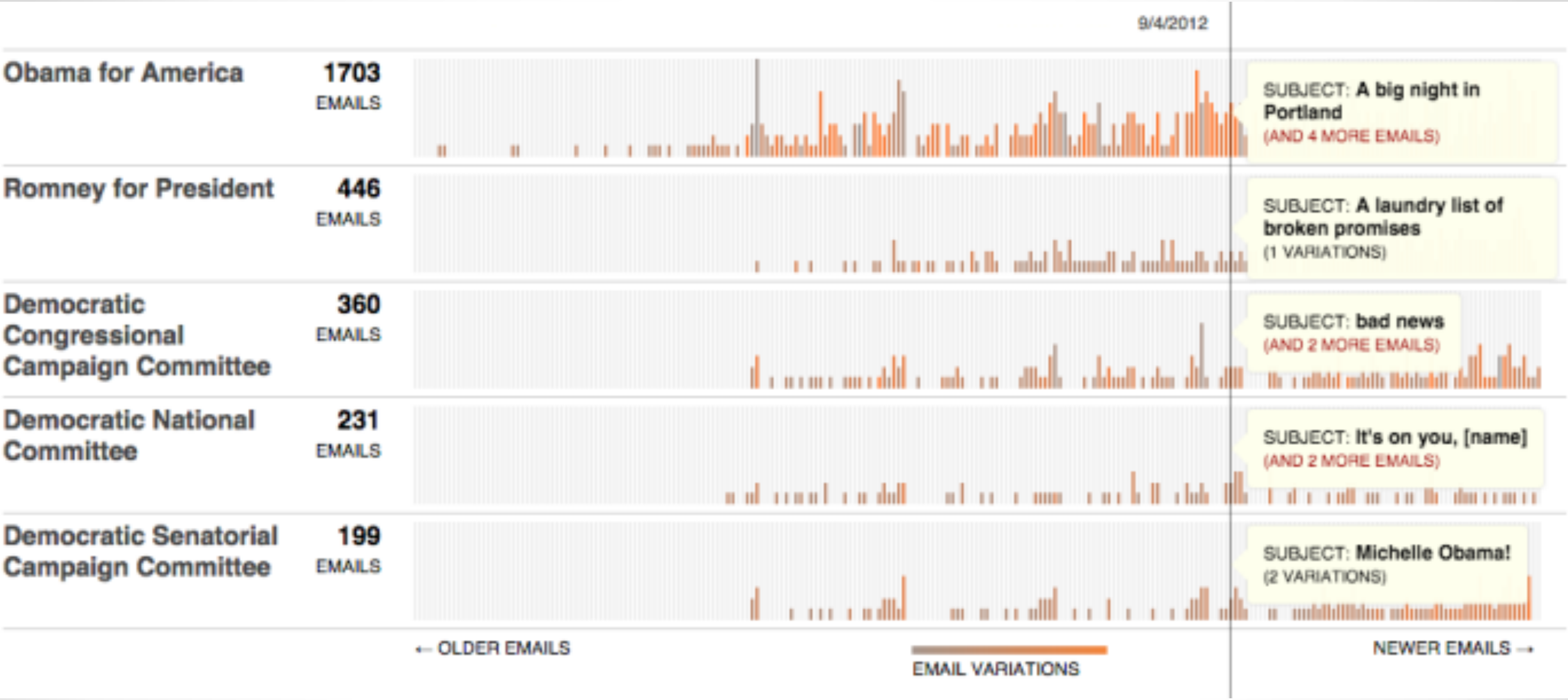
A Vector Space Model for Automatic Indexing, Salton et al, 1975

TF TF-IDF



nj-senator-menendez corpus, Overview sample files

color = human tags generated from TF-IDF clusters



SUBJECT: **A big night in Portland**
(AND 4 MORE EMAILS)

SUBJECT: **A laundry list of broken promises**
(1 VARIATIONS)

SUBJECT: **bad news**
(AND 2 MORE EMAILS)

SUBJECT: **It's on you, [name]**
(AND 2 MORE EMAILS)

SUBJECT: **Michelle Obama!**
(2 VARIATIONS)

Message Machine

Jeff Larson, Al Shaw, ProPublica, 2012