CS/INFO 6742, lightly adapted from a section of Danescu-Niculescu-Mizil and Lee Neurips 2016 tutorial, http://www.cs.cornell.edu/~cristian/index_files/NIPS_NLP_for_CSS_tutorial.pdf

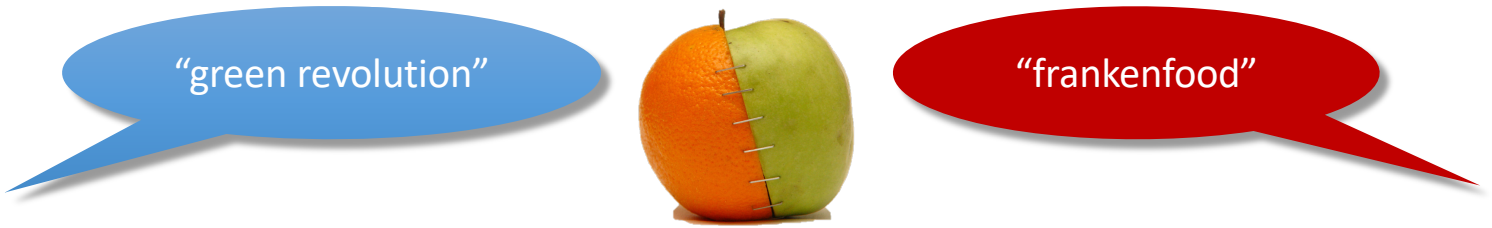
Exploring differences in two "languages"

Issues analyzed in Kleinberg (2004, *Data Stream Management* 2016), with a Markov model applied for temporal analysis.

Presentation/figures from slides 4 on follow Monroe, Colaresi and Quinn, *Political Analysis* (2008)

Example application: *frame* competition

Example: public discussion of GMOs in food

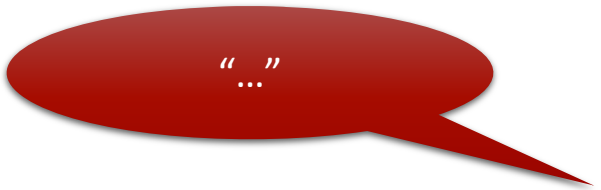


http://www.ourbreathingplanet.com/control-the-world-through-genetically-modified-food/

Additional applications: Differentiating the language of ….

- successful vs. unsuccessful persuaders
- language in one time period vs. another…
- *your experimental condition A vs. your experimental condition B!!*

Also good for sanity-checking your data…



Example: 106th U.S. Senate speeches on abortion

"Frames" → words we might expect from Democrats:

… women's rights …
… privacy …

"Frames" → words we might expect from Republicans:

… unborn children …
… murder …

Assume a joint vocabulary of terms v_i .

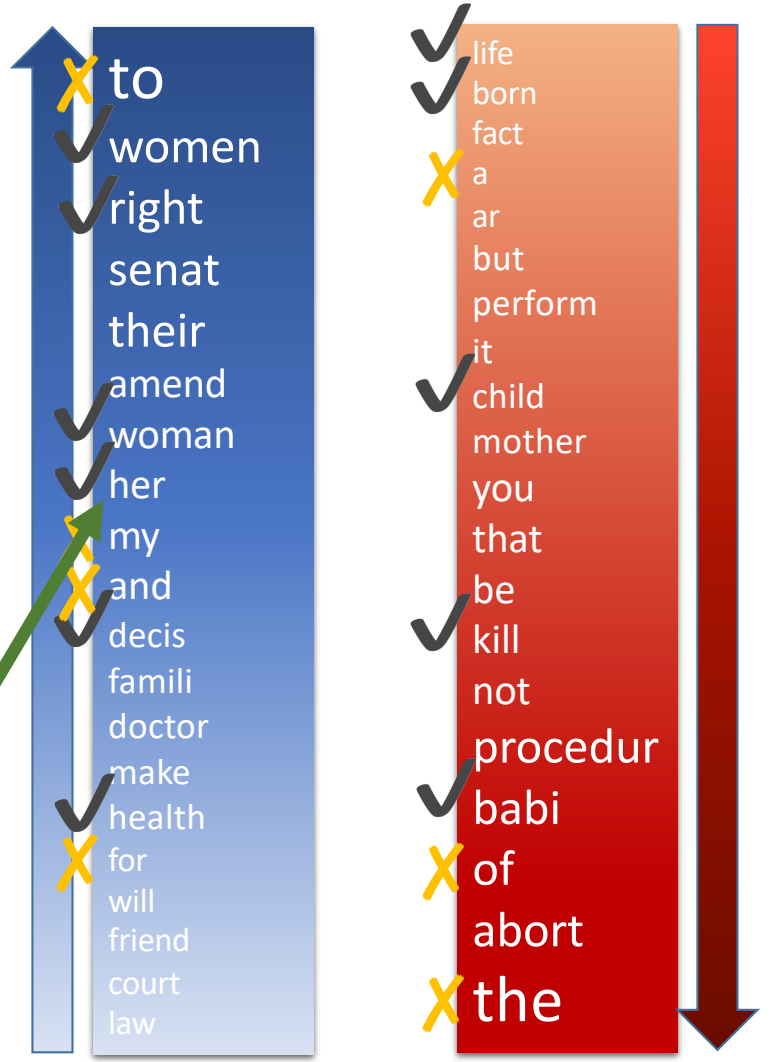
$p(v_i)$ and $p(v_i)$: **observed** relative frequency of v_i in the blue and red samples

Ranking idea

Top and bottom 20 words according to

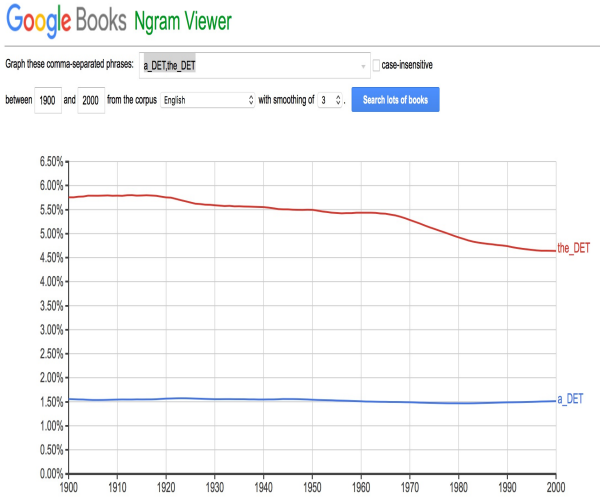
$$p(v_i) - p(v_i)$$

important, but would be lost with stopword filtering



Aside: “stopword removal” not recommended

- Very-frequent terms have been proving “increasingly” useful, e.g., for stylistic or psychological cues
- “a” vs “the” is surprising

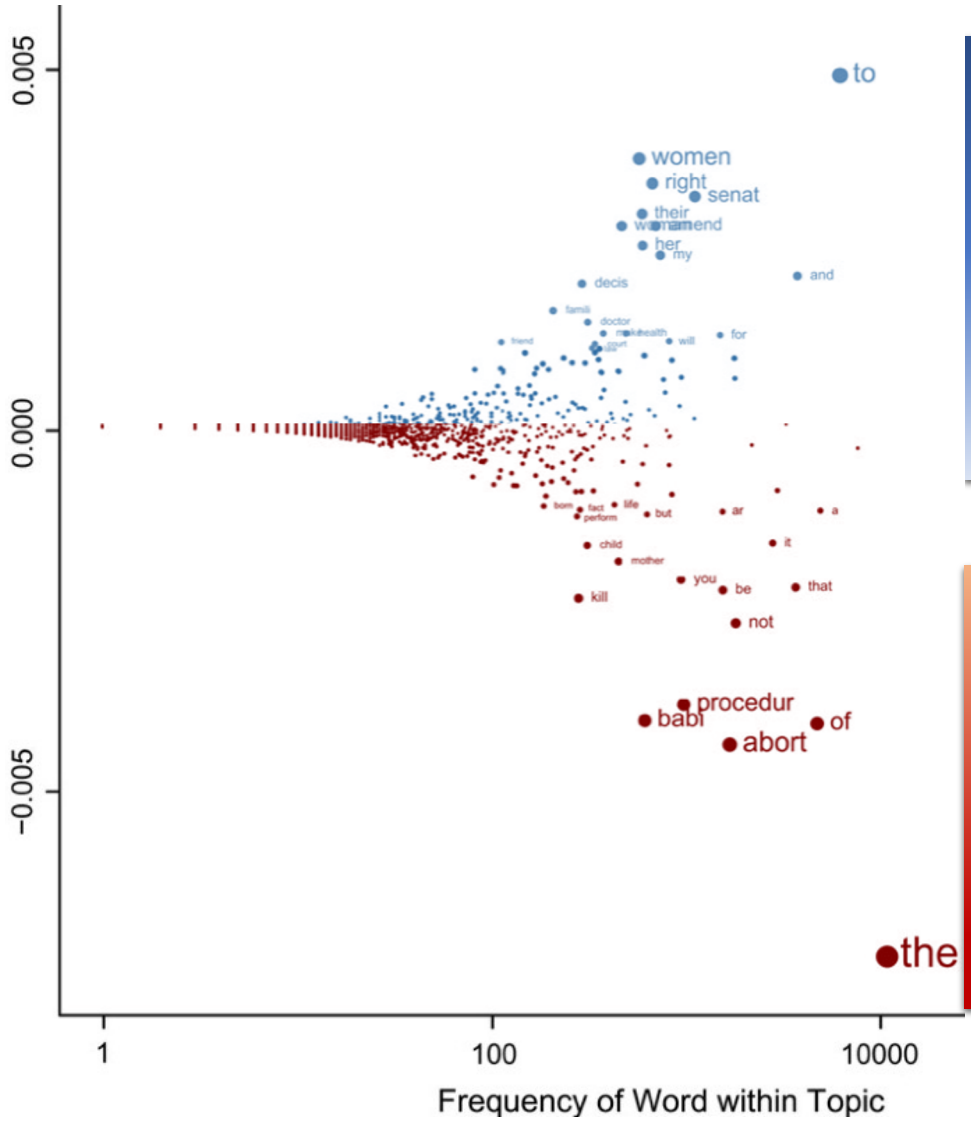


[for years LL assumed this was a bug, but see Language Log, Jan 3 2016: “The case of the missing determiners”]

$p(vi)$ vs. count

$p(v_i) - p(v_i)$ favors big counts, i.e., v_i towards the righthand side of this plot

(can't have a large difference between two small differences)



to
women
right
senat
their
amend
woman

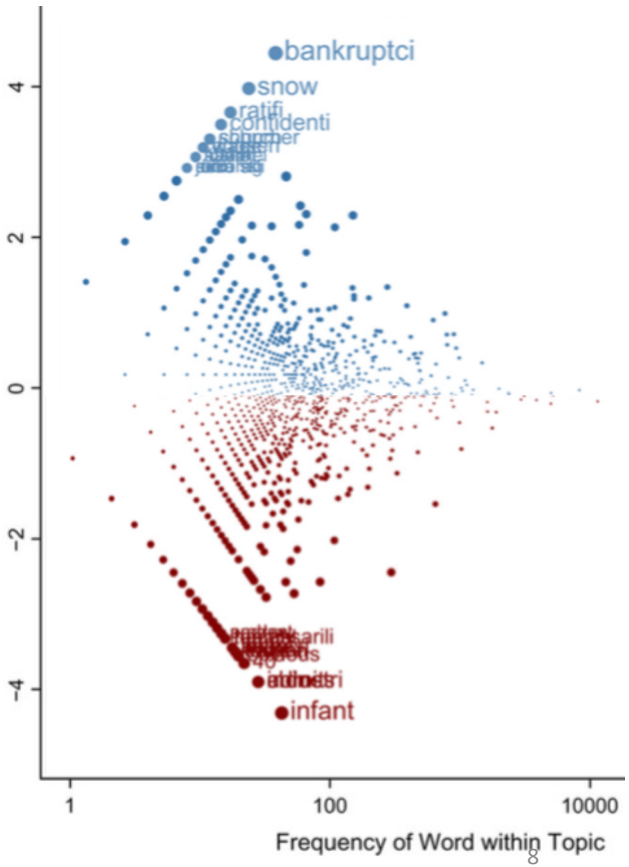
kill
not
procedur
babi
of
abort
the

Ranking by log odds-ratio

$$\log \frac{p(v_i)/(1-p(v_i))}{p(v_i)/(1-p(v_i))}$$

bankruptc
snow
ratifi
confidenti
church
schumer
chosen
voter
wage
1974
attach
attornie
idaho
sadli
coverag
d
juri
mikulsi

tonight
necessarili
martin
peter
leg
harvest
frist
bright
anim
trade
taught
dayton
obvious
40
industri
chines
admit
infant



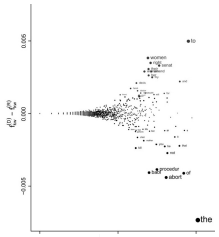
(Move to handout: model choices)

Aside: warning on ignoring (language) history

Should we really write $P(v_i)$, with no conditioning on context?

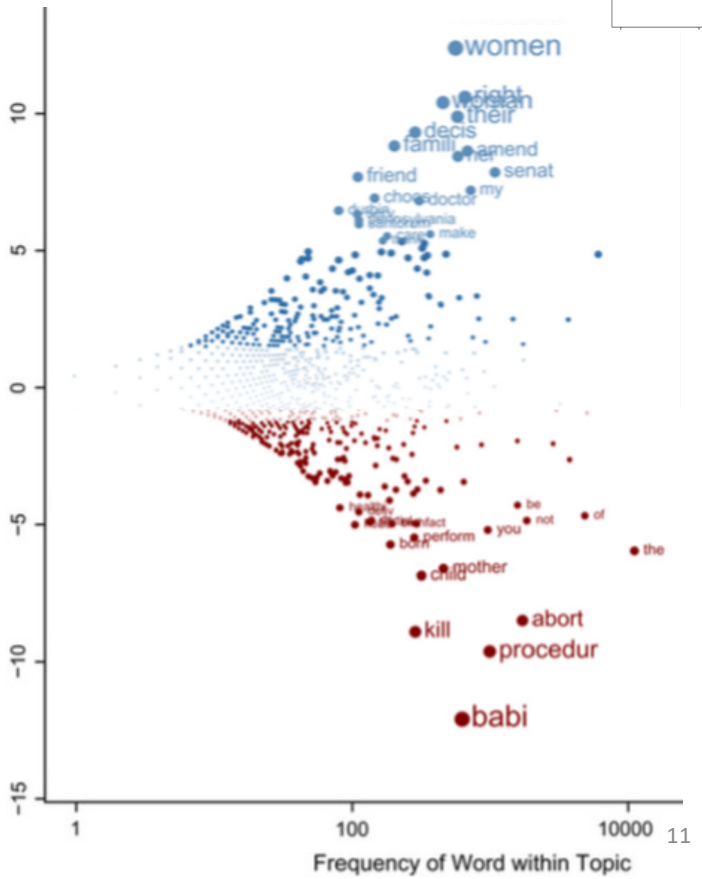
- Previous lectures: language accommodation/coordination
- Church 2000: "Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to p / 2 than p^2 ". COLING.
  - "Finding a rare word like *Noriega* in a document is like lightning. We might not expect lightning to strike twice, but it happens all the time, especially for good keywords."

Ranking by z-score of log odds-ratio, with model of variance (uninformative prior)

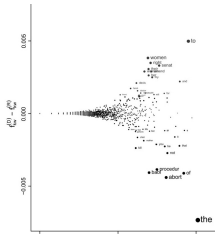


women
right
woman
their
decis
famili
amend
her
senat
friend
my
choos
doctor
durbin
serv
pennsylvania
santorum

of
dr
not
partial
fact
birth
head
you
perform
born
the
mother
child
abort
kill
procedur
babi



Ranking by z-score of log odds-ratio, with model of variance (informative prior)



women
woman
right
decis
her
doctor
durbin
choos
santorum
v
pennsylvania
pregnanc
viabil
friend
privaci
their
famili

aliv
deliv
dr
head
perform
head
perform
birth
healthi
partial
child
born
mother
abort
procedur
kill
babi