What makes two “languages” different?


Presentation/figures follow Monroe, Colaresi and Quinn, *Political Analysis* (2008)
Persuasion: frame competition

Example: public discussion of GMOs in food

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

Example: public discussion of GMOs in food
Persuasion: frame competition

Example: public discussion of GMOs in food

“green revolution”

“frankenfood”

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...
• males vs females
• 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

\[ p(v_i) - p(v_i) \]
Ranking idea

Top and bottom 20 words according to

\[ p(v_i) - p(v_i) \]
Ranking idea

Top and bottom 20 words according to

\[ p(v_i) - p(v_i) \]
Ranking idea

Top and bottom 20 words according to

\[ p(v_i) - p(v_i) \]
Ranking idea

Top and bottom 20 words according to

\[ p(v_i) \sim p(v_i) \]
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”]
\( p(v_i) \) vs. count
$p(v_i)$ vs. count

$p(v_i) - p(v_i)$ favors big counts, i.e., $v_i$ towards the righthand side of this plot
\( p(v_i) \) 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)
Ranking by log odds-ratio

$$\log \frac{p(v_i)/(1 - p(v_i))}{p(v_i)/(1 - p(v_i))}$$
Ranking by log odds-ratio

$$\log \frac{p(v_i)/(1 - p(v_i))}{p(v_i)/(1 - p(v_i))}$$
Ranking by log odds-ratio

\[ \log \frac{p(v_i)/(1 - p(v_i))}{p(v_i)/(1 - p(v_i))} \]
(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)
Ranking by z-score of log odds-ratio, with model of variance (uninformative prior)
Ranking by z-score of log odds-ratio, with model of variance (uninformative prior)
Ranking by z-score of log odds-ratio, with model of variance (informative prior)