lecture 14: visualizing text

November 27, 2017

Portions of this slide deck adapted from Jordan Crouser of Smith College, in turn adapted from J. Chuang of the University of Washington

What is text data?

- Documents
  - Articles, books and novels
  - E-mails, web pages, blogs
- Text snippets
  - Tweets, SMS messages
  - Tags, comments, profiles
- And more...
  - Computer programs, logs
  - Collections of documents
  - This slide!

via Jordan Crouser

What are some characteristics of text data?

- Often high dimensional (over 228,000 words in OED)

- Packed with meaning and relationships:

  - **Correlations**: Hong Kong, San Francisco, Bay Area

  - **Order**: April, February, January, June, March, May

  - **Membership**: Tennis, Running, Swimming, Hiking, Piano

  - **Hierarchy**, antonyms & synonyms, entities, …

via Jordan Crouser

Why visualize text data?

- **Understand** – read a document
- **Summarize** – get the "gist" of a document
- **Cluster** – group together similar contents
- **Quantify** – convert to numerical measures
- **Correlate** – compare patterns in text to those in other data, e.g., test scores with conversations on social media

via Jordan Crouser

“Bag of words” model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
  - Each dimension corresponds to a term (10,000+)
  - Each value represents the relevance
- For example, simple term counts
- Aggregate into a document-term matrix

	Antony and Cleopatra	Julius Caesar
Antony	157	73
Brutus	4	157
Caesar	232	227
Calpurnia	0	10
Cleopatra	57	0

via Jordan Crouser

An example: heath care speeches

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you about an issue that is central to that future – and that is the issue of health care.

I am not the first President to take up this cause, but I am determined to be the last. It has now been nearly a century since Theodore Roosevelt first called for health care reform. And ever since, nearly every President and Congress, whether Democrat or Republican, has attempted to meet this challenge in some way. A bill for comprehensive health reform was first introduced by John Dingell Sr. in 1943. Sixty-five years later, his son continues to introduce that same bill at the beginning of each session.

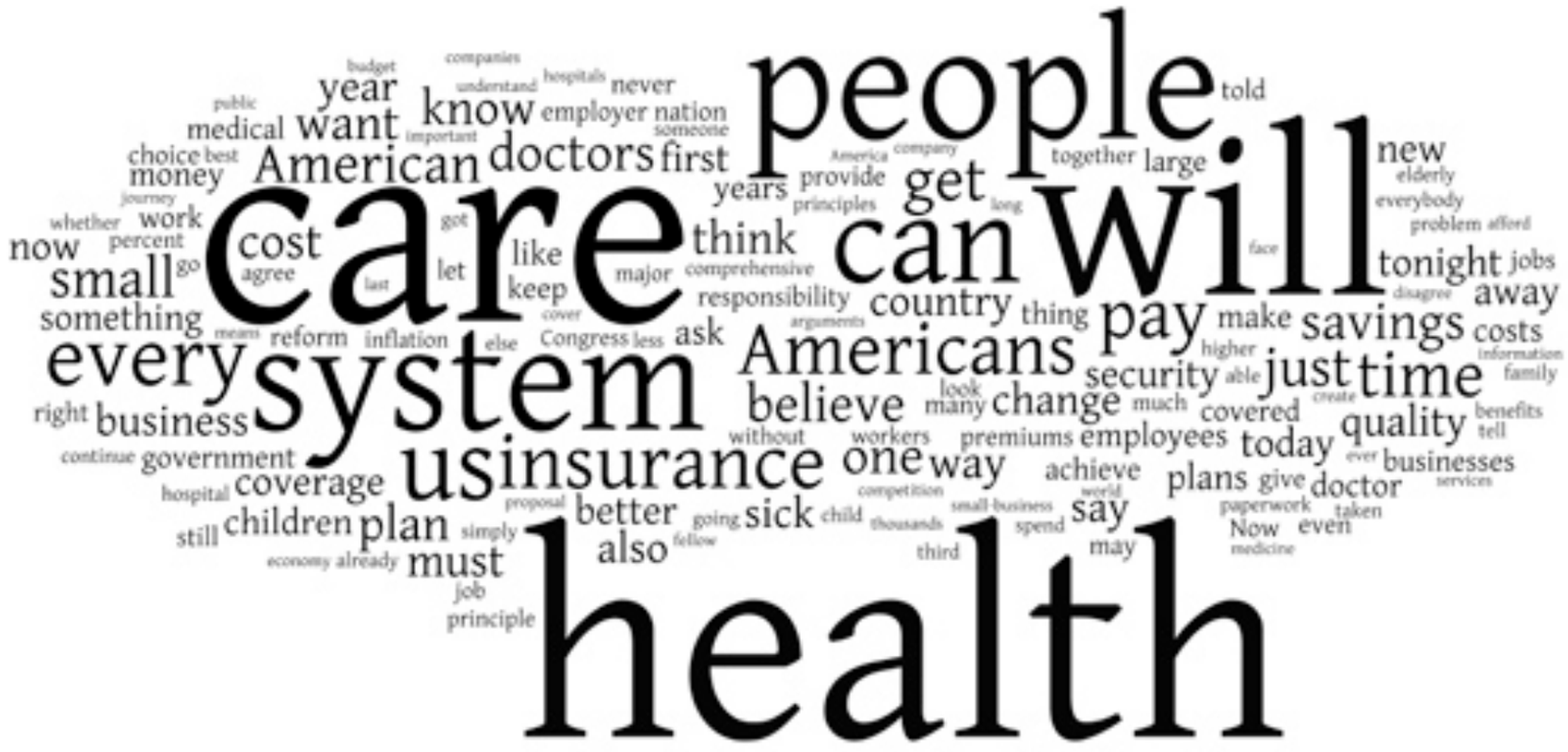
Our collective failure to meet this challenge – year after year, decade after decade – has led us to a breaking point. Everyone understands the extraordinary hardships that are placed on the uninsured, who live every day just one accident or illness away from bankruptcy. These are not primarily people on welfare. These are middle-class

via Jordan Crouser

New York Times: Obama 2009

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93, via Jordan Crouser

New York Times: Clinton 1993



economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93, via Jordan Crouser

Comparison

Obama 2009

Clinton 1993

Rep. Charles Boustany of Louisiana 2009

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93, via Jordan Crouser

Strengths and weaknesses of word clouds

- Strengths
  - Familiar to many people
  - Can help with “gisting” and initial query formation
- Weaknesses
  - Does not show the structure of the text
  - Sub-optimal visual encoding (position is not meaningful)
  - Inaccurate size encoding (long words are bigger)
  - May not facilitate comparison (unstable layout)
  - Term frequency may not be meaningful

via Jordan Crouser

Weighting words

Term Frequency

tf_{td} = # of times term t appears in document d

TF-IDF: Term Frequency by Inverse Document Frequency

$$tf\text{-}idf_{td} = \frac{\text{\# of times term } t \text{ appears in document } d}{\text{\# of times term } t \text{ appears in all documents}}$$

via Jordan Crouser

Published: July 3, 2011

Inaugural Words: 1789 to the Present

A look at the language of presidential inaugural addresses. The most-used words in each address appear in the interactive chart below, sized by number of uses. Words highlighted in yellow were used significantly more in this inaugural address than average. (Related Article)

2005

George W. Bush

Full text of the address

Article from the Times archive (pdf)

Mr. Bush began his second term without using the words "Iraq," "Afghanistan," "Sept. 11" or "terrorism." He instead cast the crises and controversies of his first four years as a struggle in defense of the nation's founding creed: freedom. "The best hope for peace in our world is the expansion of freedom in all the world," he said, pledging himself to "the ultimate goal of ending tyranny in our world."

freedom America liberty nation American country world time free citizen hope history people day human right seen ideal work unite justice cause government move choice tyranny live act life accept defend duty generation great question honor states president fire character force power fellow enemy century witness excuse soul God division task define advance speak institution independence society serve determine

http://www.nytimes.com/interactive/2009/01/17/washington/20090117_ADDRESSES.html

Published: January 25, 2011

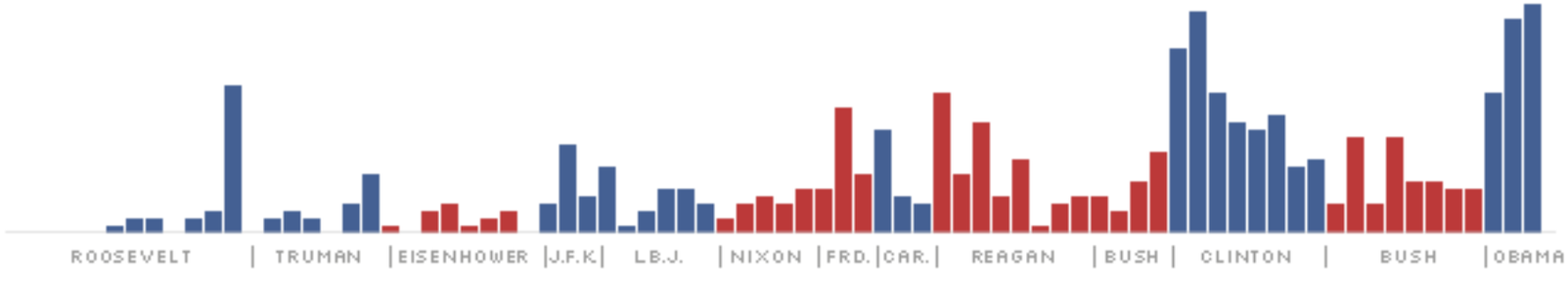
Patterns of Speech: 75 Years of the State of the Union Addresses

In 2010, President Obama was the first modern president to use the words "bubble," "supermajority" and "obesity" in a State of the Union speech. But other words have a longer history. Below, a historical look at the number of times presidents have used selected words in their State of the Union addresses (or analogous speeches) from 1934 to 2011.

'jobs'

With unemployment above 9 percent, jobs were a focus of President Obama's speech. Historically, jobs get mentioned in the speech in rough correlation to the economic cycle, with spikes around 1975, 1981, 1991 and 2002.

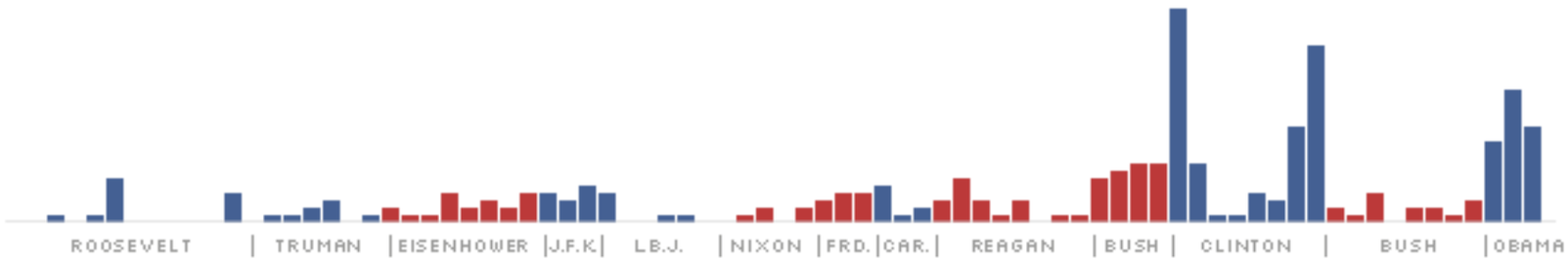
Words included: JOB, JOBS



'invest'

Historically, Democrats use this word more than Republicans, and they mean "public" investment, or new government programs. Bill Clinton used it a lot at the beginning and end of his term, first to propose new programs, and last to take credit for successful ones. Mr. Obama mentioned the word 13 times, proposing new investments in information technology, clean energy and science research.

INVEST, INVESTED, INVESTING, INVESTMENT, INVESTMENTS, INVESTS



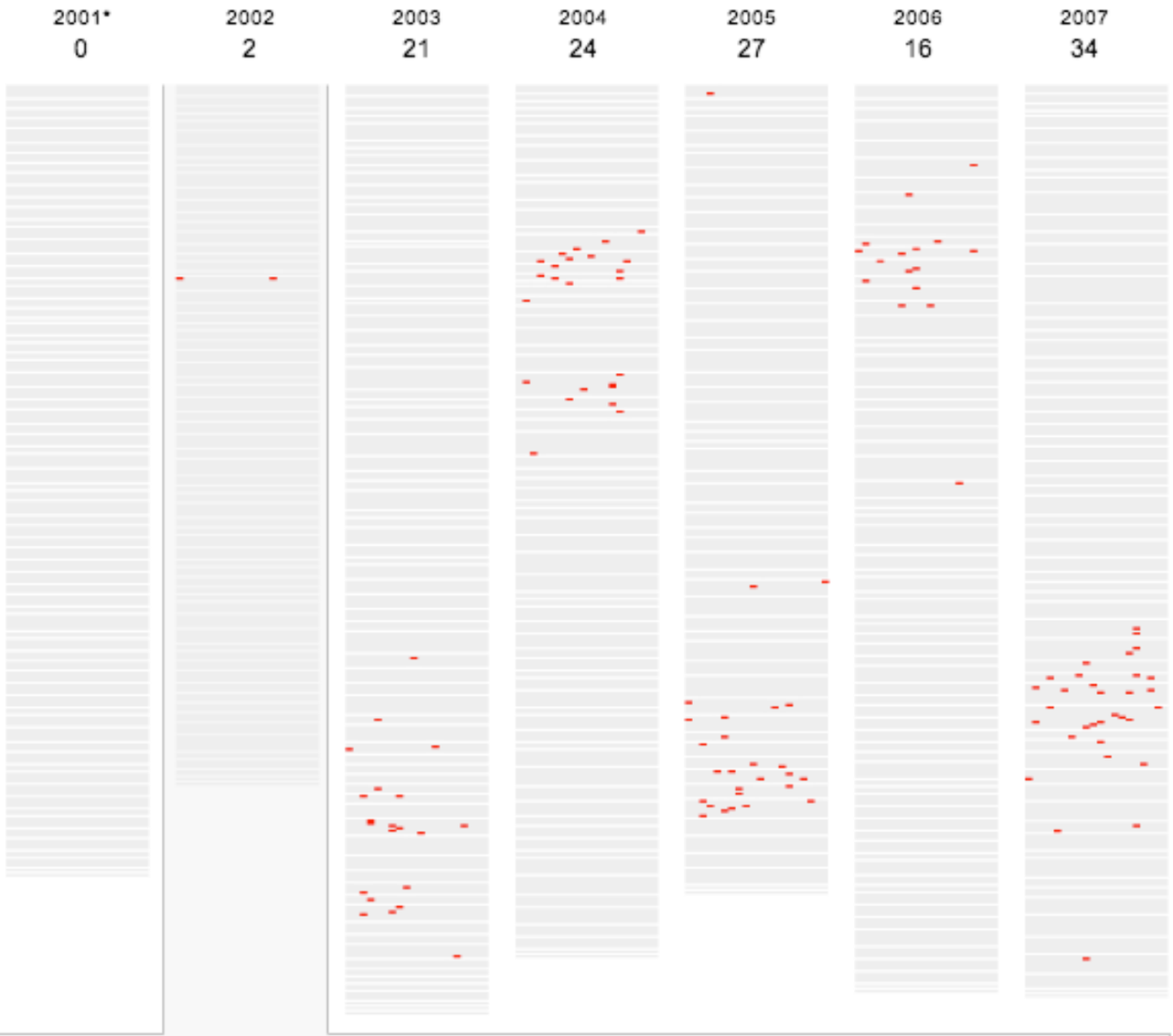
http://www.nytimes.com/interactive/2011/01/25/us/politics/state-of-the-union-words-used.html

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Search | or choose a word here.

Use of the phrase "Iraq" in past State of the Union Addresses



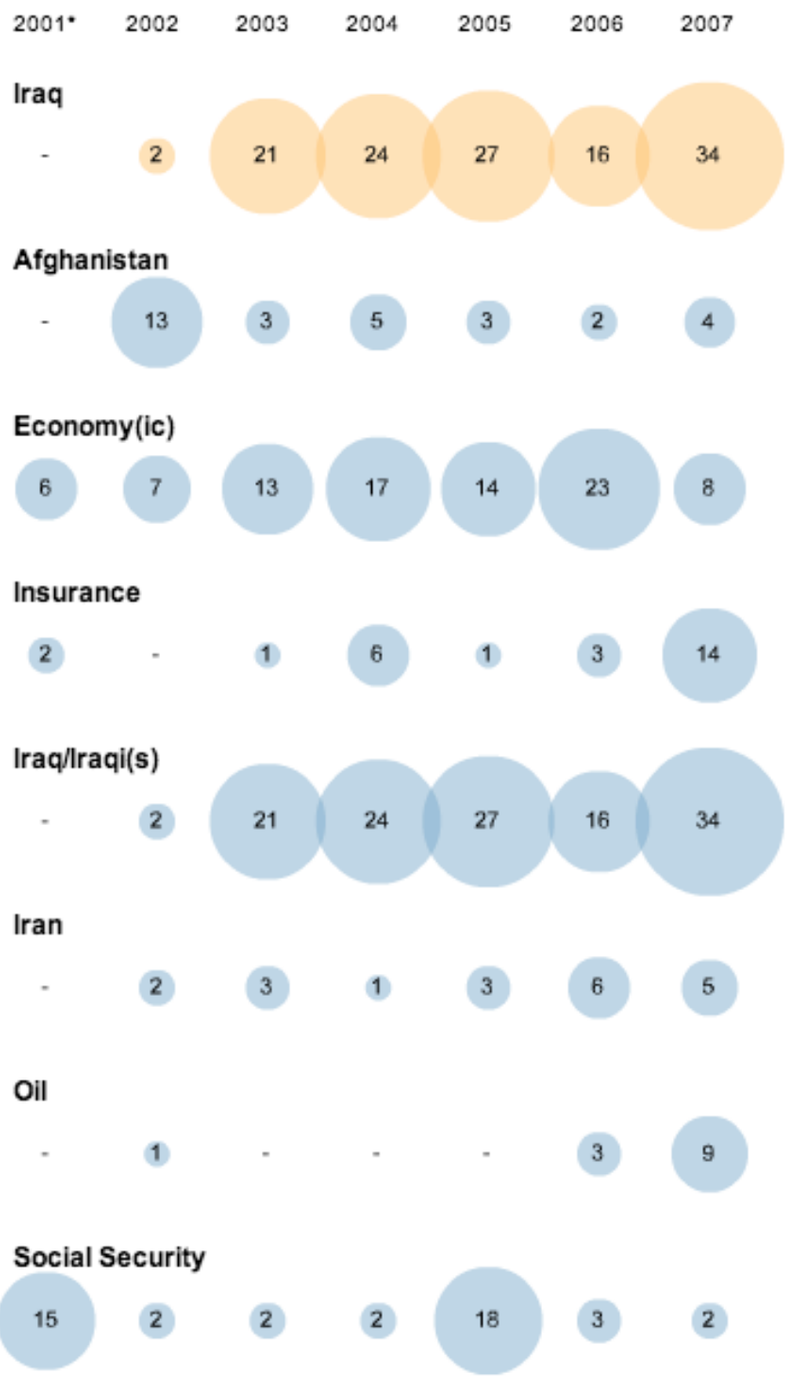
The word in context

Next Instance of 'Iraq'

IRAQ continues to flaunt its hostility toward America and to support terror. The Iraqi regime has plotted to develop anthrax, and nerve gas, and nuclear weapons for over a decade. This is a regime that has already used poison gas to murder thousands of its own citizens -- leaving the bodies of mothers huddled over their dead children. This is a regime that agreed to international inspections -- then kicked out the inspectors. This is a regime that has something to hide from the civilized world.

-- 2002 (Paragraph 20 of 67)

Compared with other words



* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

Ben Werschkul/The New York Times

http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html

Limitations of frequency statistics

- Often favors frequent (TF) or rare (IDF) terms
    - Still not clear that these provide best description
- A “bag of words” ignores additional information
    - Grammar / part-of-speech
    - Position within document
    - Recognizable entities
- Typically focus on unigrams (single terms)

via Jordan Crouser

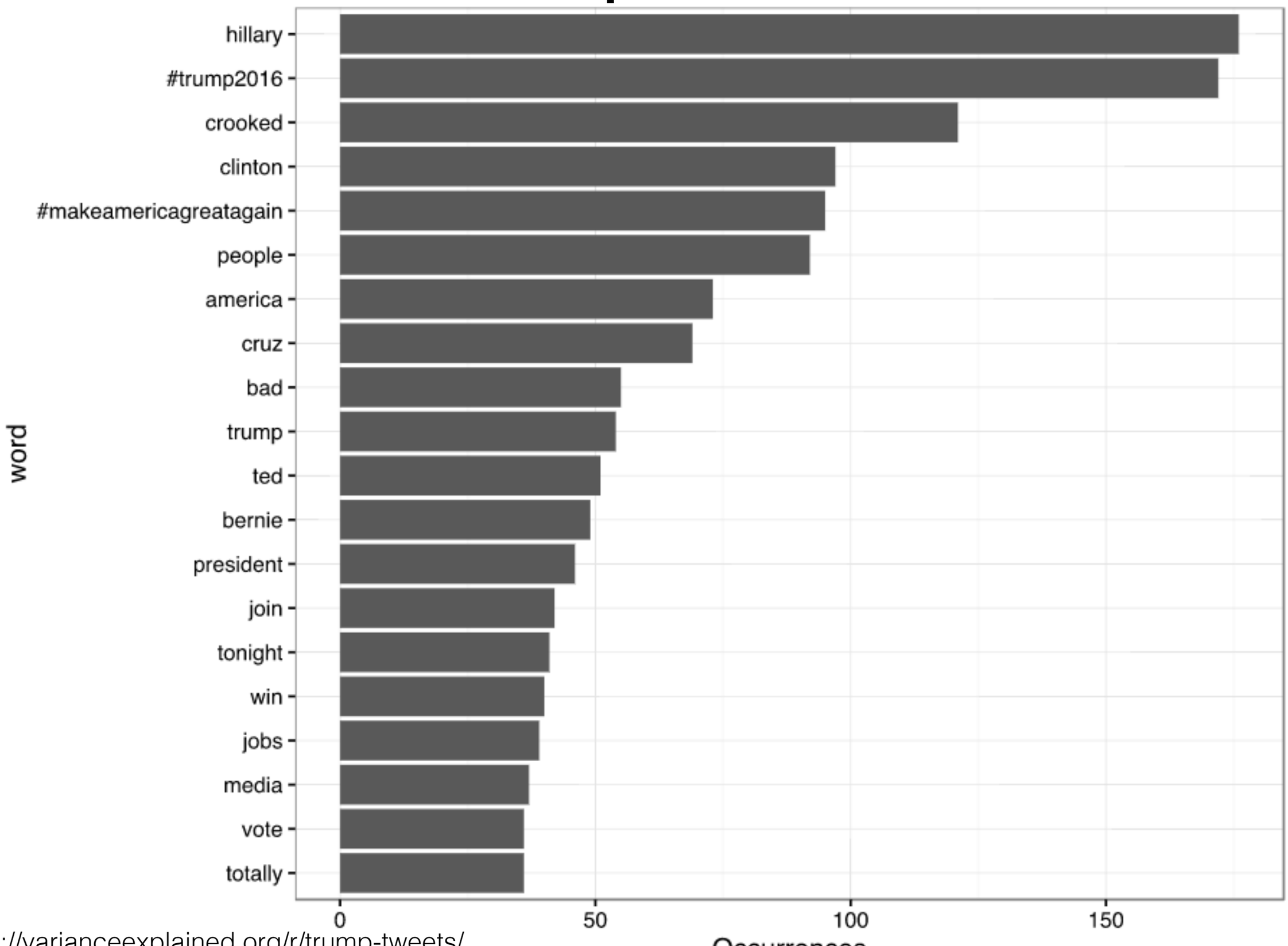
50 years of pop music

http://kaylinwalker.com/50-years-of-pop-music/

50 years of pop music

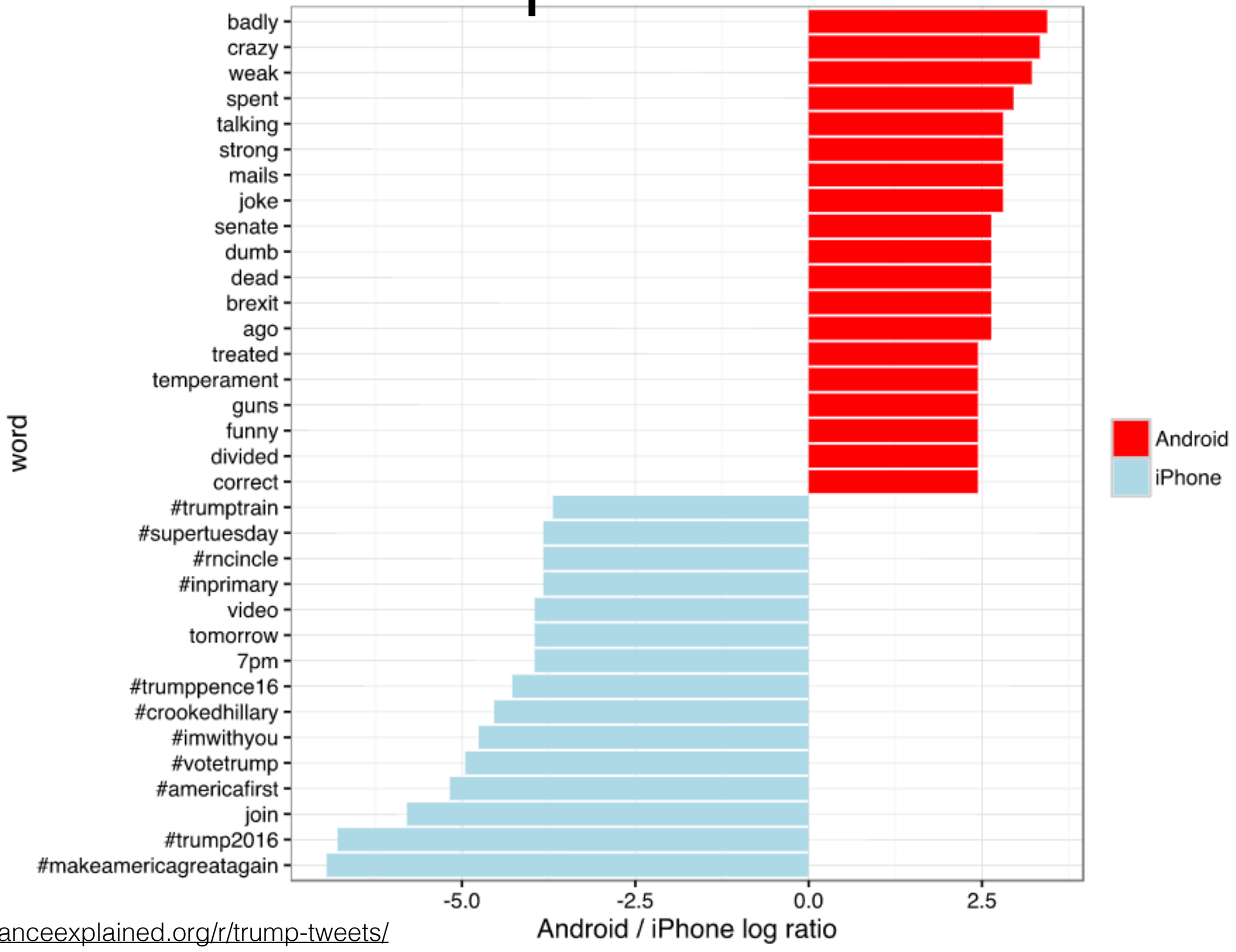
http://kaylinwalker.com/50-years-of-pop-music/

Trump tweets



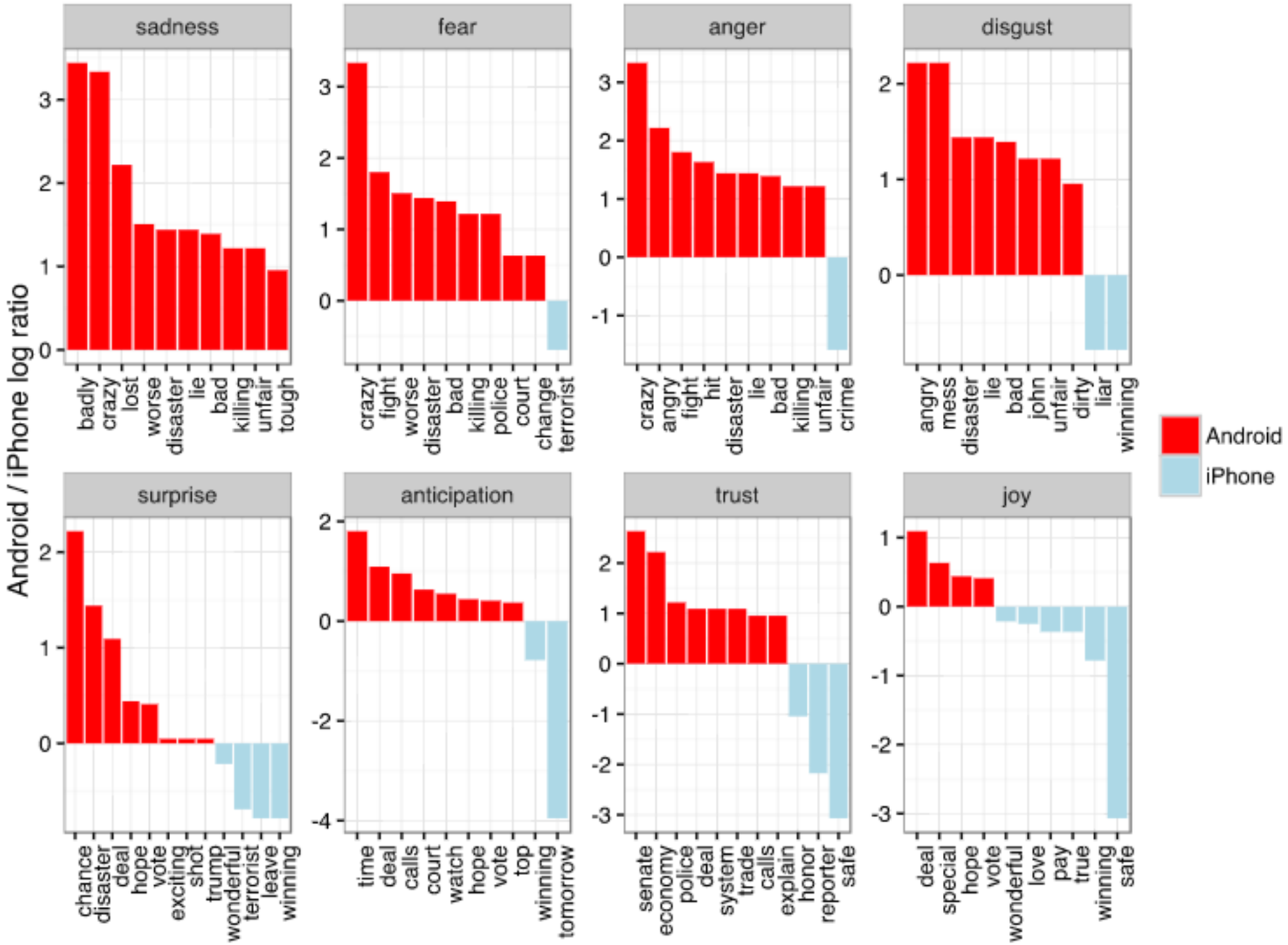
http://varianceexplained.org/r/trump-tweets/

Trump tweets



http://varianceexplained.org/r/trump-tweets/

Trump tweets



http://varianceexplained.org/r/trump-tweets/

Lab: word clouds

- We'll be using http://voyant-tools.org/ as our tool today

- You may want to use some text data from Project Gutenberg: http://voyant-tools.org/

- For example, Alice in Wonderland: http://www.gutenberg.org/cache/epub/19033/pg19033.txt

- Create a visualization to identify one trend or interesting feature in a text dataset, and post it in #lab7