Natural Language Processing

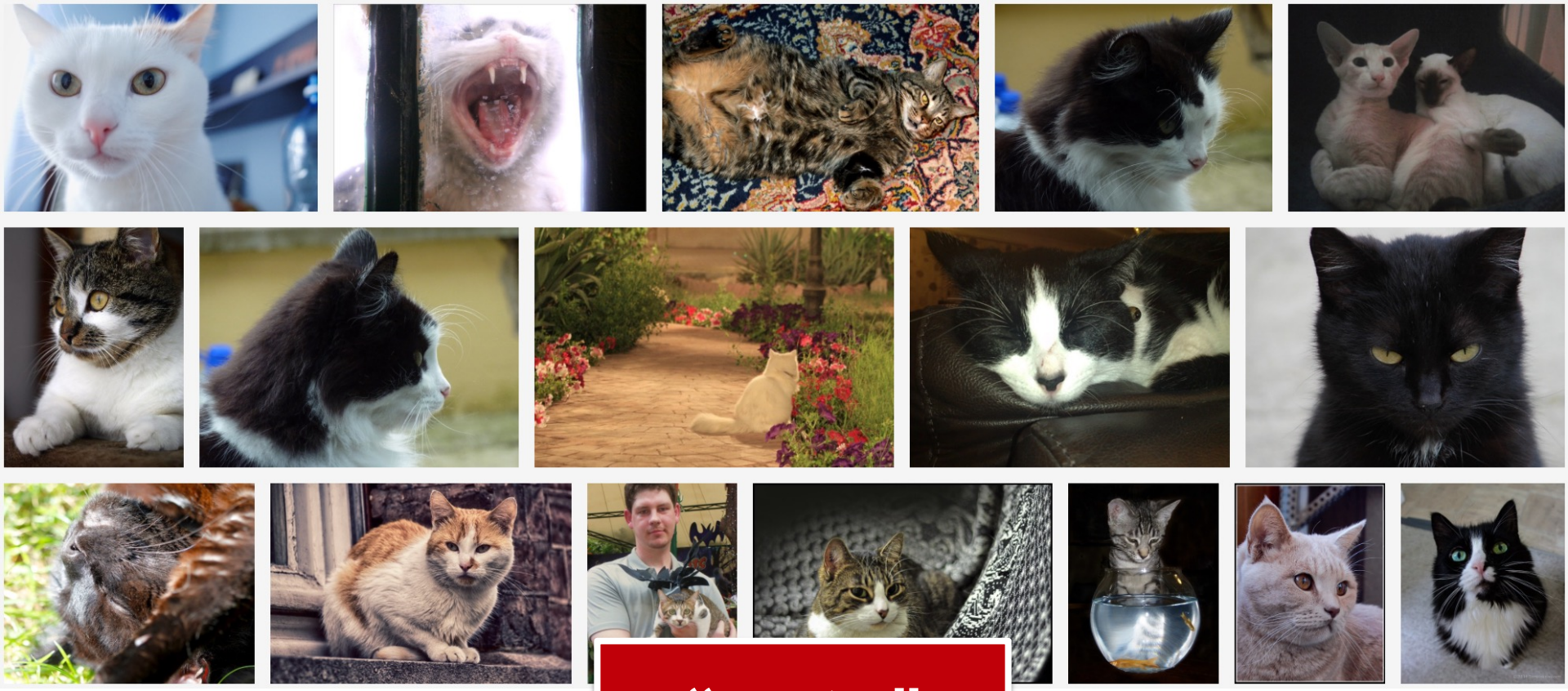
02: Basic Text Processing

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2022-04-01

Technology
Arts Sciences
TH Köln

Words as dimensionality reduction



https://search.creativecommons.org/search?q=cats

Words

- One morning I shot an elephant in my pajamas
- I didn’t shoot an elephant
- Imma let you finish but Beyonce had one of the best videos of all time
- 一天早上我穿着睡衣射了一只大象

Words

- @phschaer have you seen this :) http://popvssoda.com

Tokenization before Twitter

```
@
phschaer
have
you
seen
this
:
)
http
:
//popvssoda.com
```

Emoticons

Icon											Meaning
:-) :)	:-] :]	:-3 :3	:->> :>	8-) 8)	:-} :}	:o)	:c)	:^)	=]	=)	Smiley or happy face.[4][5][6]
:-D :D	8-D 8D	x-D xD	X-D XD	=D	=3	B^D					Laughing,[4] big grin,[5][6] laugh with glasses,[7] or wide-eyed surprise[8]
:-))											Very happy or double chin[7]
:-(:(	:-c :c	:-< :<	:-[:[	:-II	>:[	:{	:@	>:(			Frown,[4][5][6] sad,[9] angry,[7] pouting
:'-(:'(											Crying[9]
:'-) :')											Tears of happiness[9]
D-':	D:<	D:	D8	D;	D=	DX					Horror, disgust, sadness, great dismay[5][6] (right to left)
:-O :O	:-o :o	:-0	8-0	>:O							Surprise,[3] shock,[4][10] yawn[11]
:-* :*	:x										Kiss
;-) ;)	*-) *)	;-] ;]	;^)	:-,	;D						Wink,[4][5][6] smirk[10][11]
:-P :P	X-P XP	x-p xp	:-p :p	:-Þ :Þ	:-þ :þ	:-b :b	d:	=p	>:P		Tongue sticking out, cheeky/playful,[4] blowing a raspberry

Types and tokens

- Type = abstract descriptive concept
- Token = instantiation of a type

To be or not to be

- 6 tokens (to, be, or, not, to, be)
- 4 types (to, be, or, not)

- Types = The **vocabulary**; the unique tokens.

Types and tokens

- Type = abstract descriptive concept
- Token = instatiation of a type

How can we use types and tokens to measure vocabulary richness?

Whitespace

```
text.split(" ")
```

- As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.

Whitespace

```
text.split(" ")
```

- As much mud in the streets as if the waters had but newly retired from the face of the **earth,** and it would not be wonderful to meet a **Megalosaurus,** forty feet long or **so,** waddling like an elephantine lizard up Holborn **Hill.**

What do we lose with whitespace tokenization?

368	earth
135	earth,
68	earth.
26	earth
24	earth.
18	earth."
16	earth;
14	earth,
9	earth's
5	earth!"
5	earth!
4	earth;
4	earth,"
3	earth.”
3	earth?

3	earth!"
2	earth--to
2	earth--if
2	earth--and
2	earth:
2	earth,'
1	earth-worms,
1	earth-worm.
1	earth--which
1	earth--when
1	earth--something
1	earth-smeared,
1	earth-scoops,
1	earth's
1	earth--oh,

“earth” in sample of 100 books (Project Gutenberg)

Punctuation

- We typically don't want to just strip all punctuation, however.
  - Punctuation signals **boundaries** (sentence, clausal boundaries, parentheticals, asides)
  - Some punctuation has **expressive force**, like exclamation points (!) and question marks (?)
  - Emoticons are strong signals of e.g. sentiment

- Most tokenization algorithms (for languages typically delimited by whitespace) use **regular expressions** to segment a string into discrete tokens.

Regular expressions

regex	matches	doesn't match
`/the/`	the, isothermally	The
`/[Tt]he/`	the, isothermally, The	
`/\b[Tt]he\b/`	the, The	—The

- Python
  - **re.findall(regex, text)** finds all non-overlapping matches for a target regex.
  - `re.findall(r"[Tt]he", "The dog barked at the cat")`
  - ["The", "the"]

NLTK for the win

```
import nltk
tokens=nltk.word_tokenize(text)
```

Tokenizes following the conventions of the **Penn Treebank**:

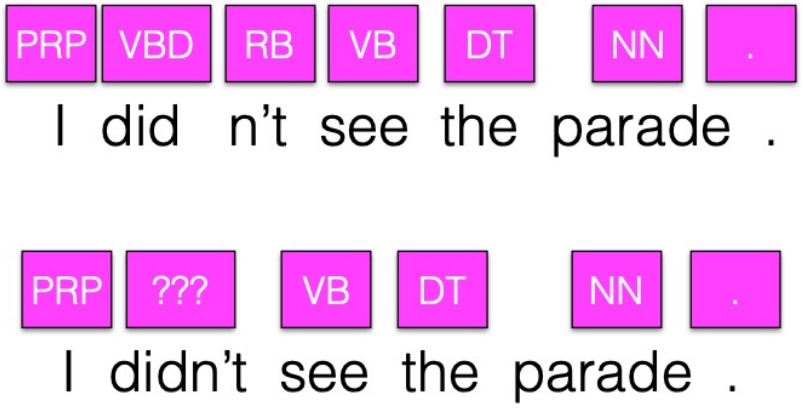
- punctuation split from adjoining words double quotes (“) changes to forward/ backward quotes based on on their location in word (``the”)
- verb contractions + ’s split into separate
- tokens: (`did_n’t, children_’s`)

ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html

NLTK for the win

```
import nltk
tokens=nltk.word_tokenize(text)
```

- Penn Treebank tokenization is important because a lot of downstream NLP is trained on annotated data that uses Treebank tokenization!

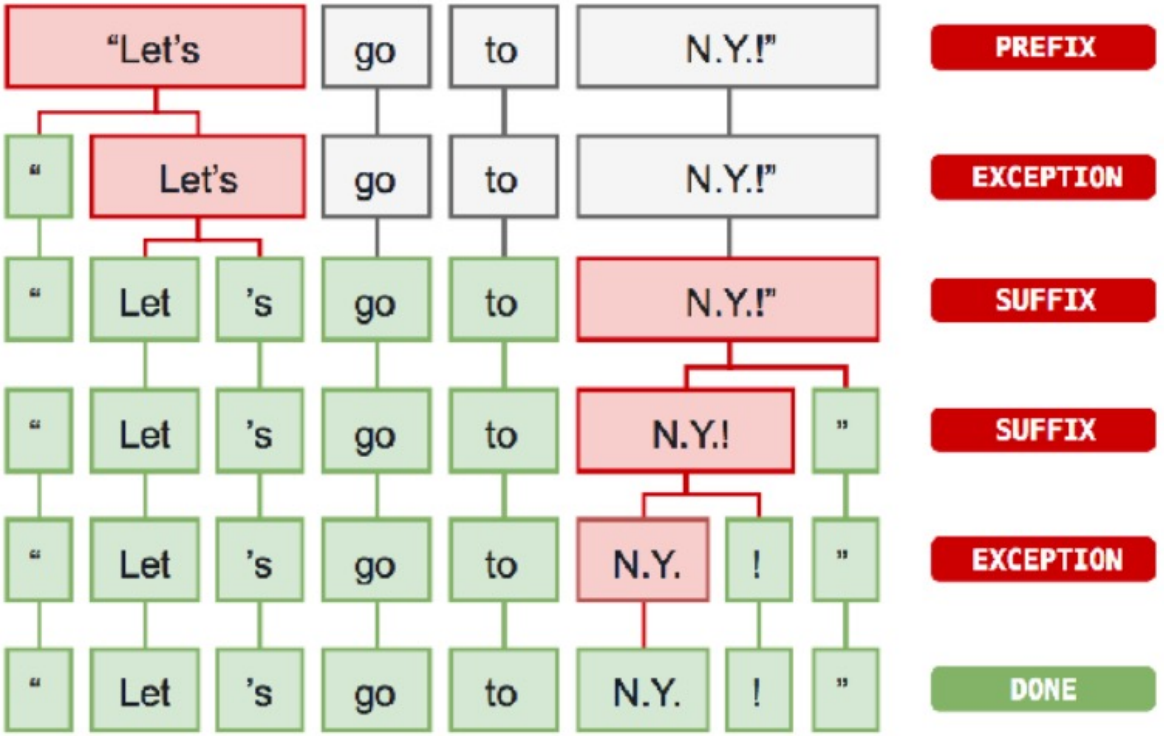


https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html

Spacy

```
import spacy
nlp = spacy.load('en')
tokens=[token.text for token in nlp(text)]
```



https://spacy.io/usage/spacy-101#annotations-token

Sentence segmentation

- Word tokenization presumes a preprocessing step of sentence segmentation – Identifying the boundaries between sentences.

- Lots of NLP operates at the level of the sentence (POS tagging, **parsing**), so really important to get it right.

- Harder to write regexes to delimit these, since there are many cases where the usual delimiters (periods, question marks) serve double duty.

Sentence segmentation

- “Do you want to go?” said Jane.
- Mr. Collins said he was going.
- He lives in the U.S. John, however, lives in Canada.

Sentence segmentation

- **NLTK**: Punkt sentence tokenizer – unsupervised method to learn common abbreviations, collocations, sentence-initial words. Can be trained on data from new domain.
  - https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt
  - Kiss, Tibor and Strunk, Jan (2006): Unsupervised Multilingual Sentence Boundary Detection (*Computational Linguistics*)

- **spaCy**: Relies on dependency parsing to find sentence boundaries.

Stemming and lemmatization

- Many languages have some inflectional and derivational morphology, where similar words have similar forms:

  organizes, organized, organizing

- Stemming and lemmatization reduce this variety to a single common **base form**.

Stemming

- Heuristic process for chopping off the inflected suffixes words

  organizes, organized, organizing → **organ**

- Lower precision, higher recall

- Remember Information Retrieval, **Porter stemmer**: Sequence of rules for removing suffixes from words
  - EMENT → ∅
  - SSES → SS
  - IES → I
  - SS → ∅
  - S→∅

Lemmatization

- Using morphological analysis to return the dictionary form of a word (the entry in a dictionary you'd find all forms under)

organizes, organized, organizing → **organ**

```
import spacy
nlp = spacy.load('en')
lemmas=[token.lemma_ for token in nlp(text)]
```

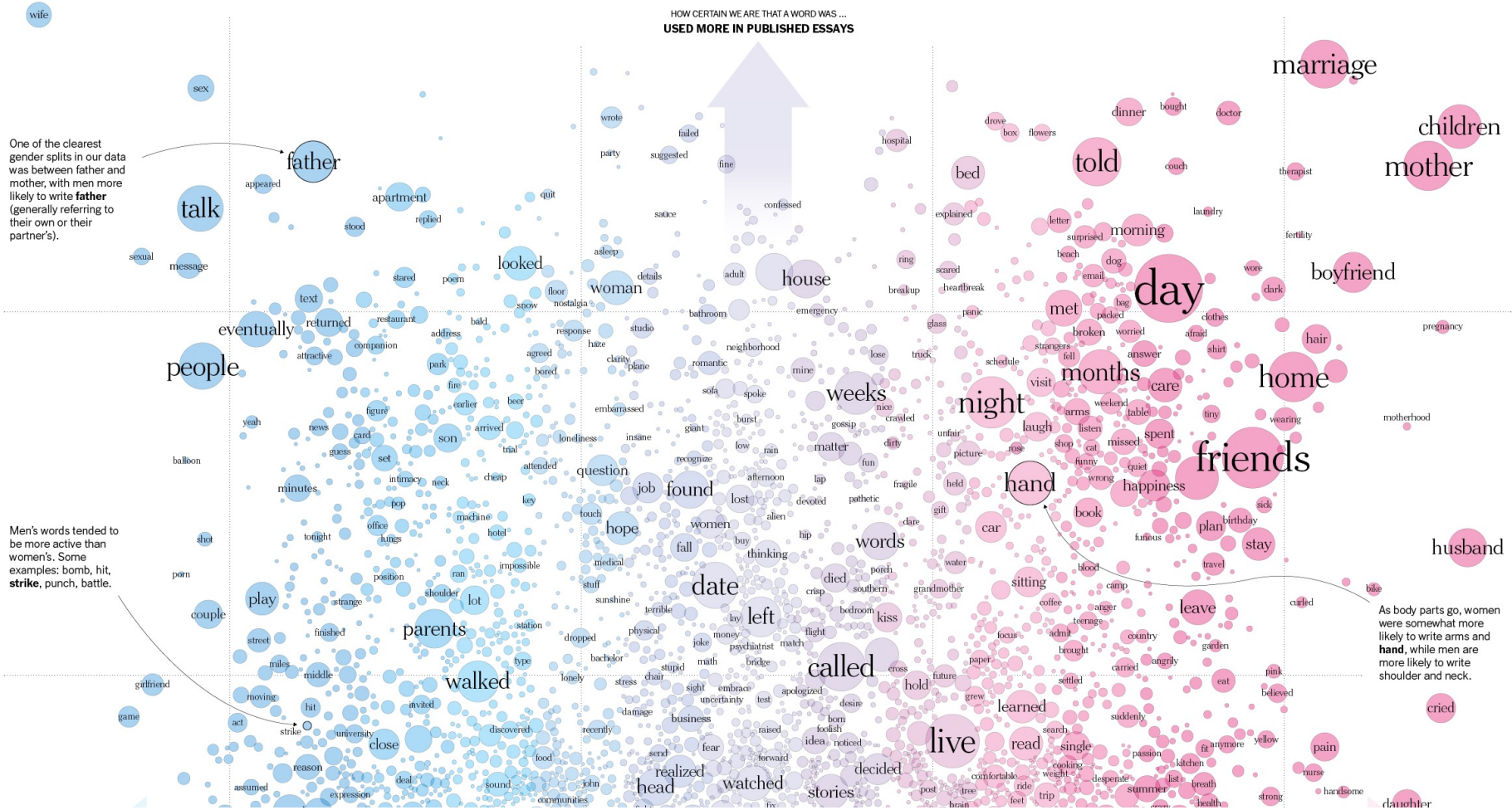
Difficulties

- When does **punctuation** disrupt the desired boundaries of a token?

Emoticons	:) :D \o/ o_O
URLs	http://www.google.com
Prices	$19.99
Decimals	19.99
Hyphens	state-of-the-art
Usernames	@phschaer
Hashtags	#blacklivesmatter

Finding Distinctive Terms

The Words Men and Women Use When They Write About Love

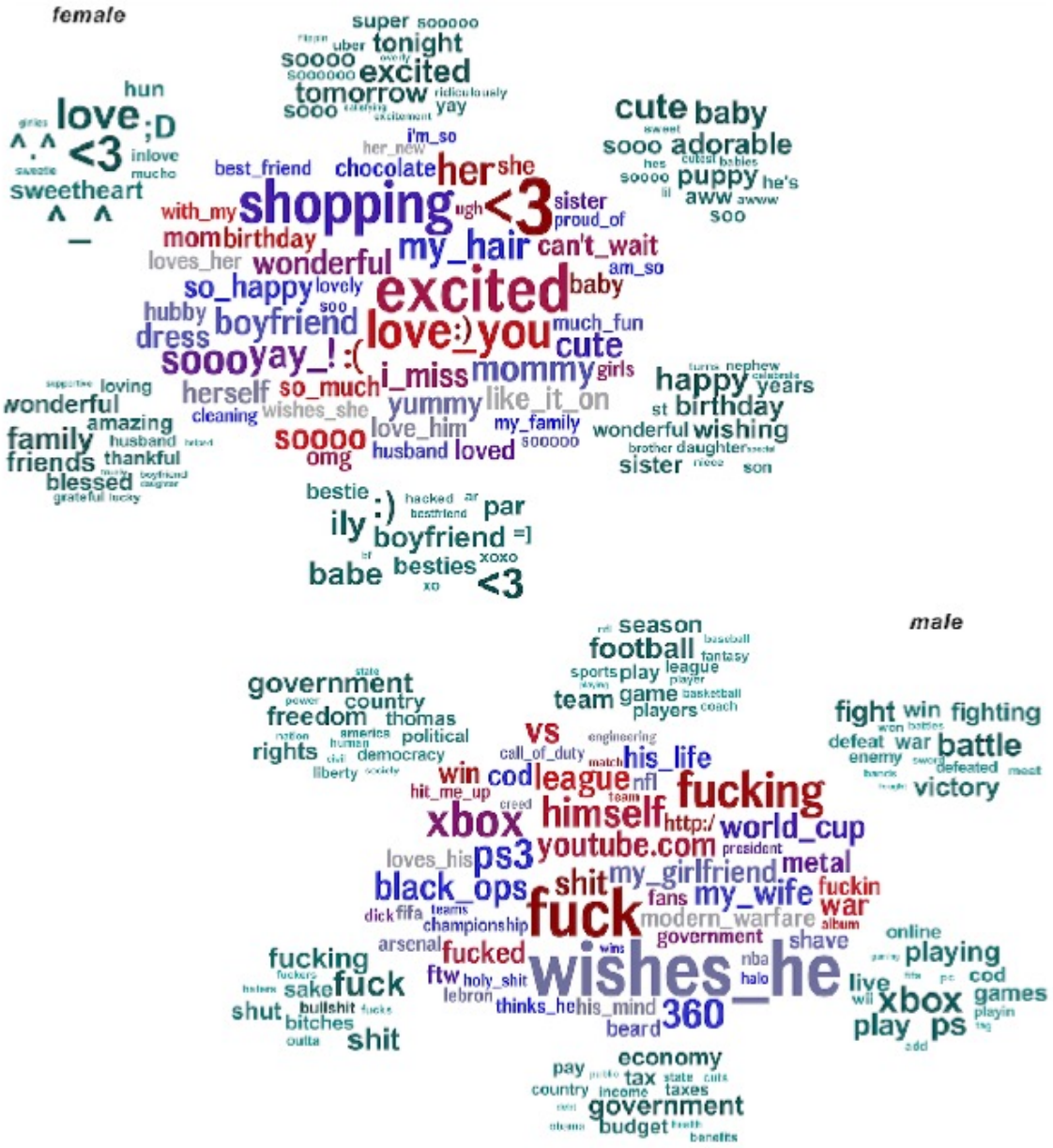


https://archive.nytimes.com/www.nytimes.com/interactive/2017/11/07/upshot/modern-love-what-we-write-when-we-write-about-love.html

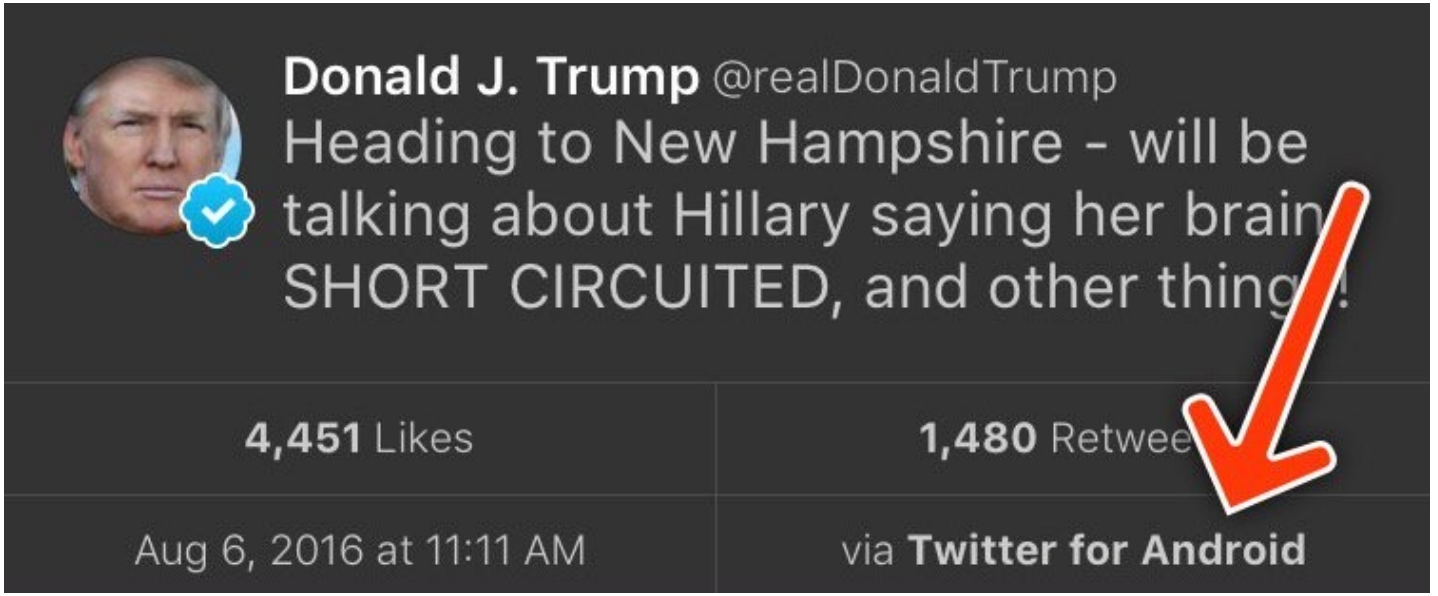
Panel B: Phrases Used More Often by Republicans

Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

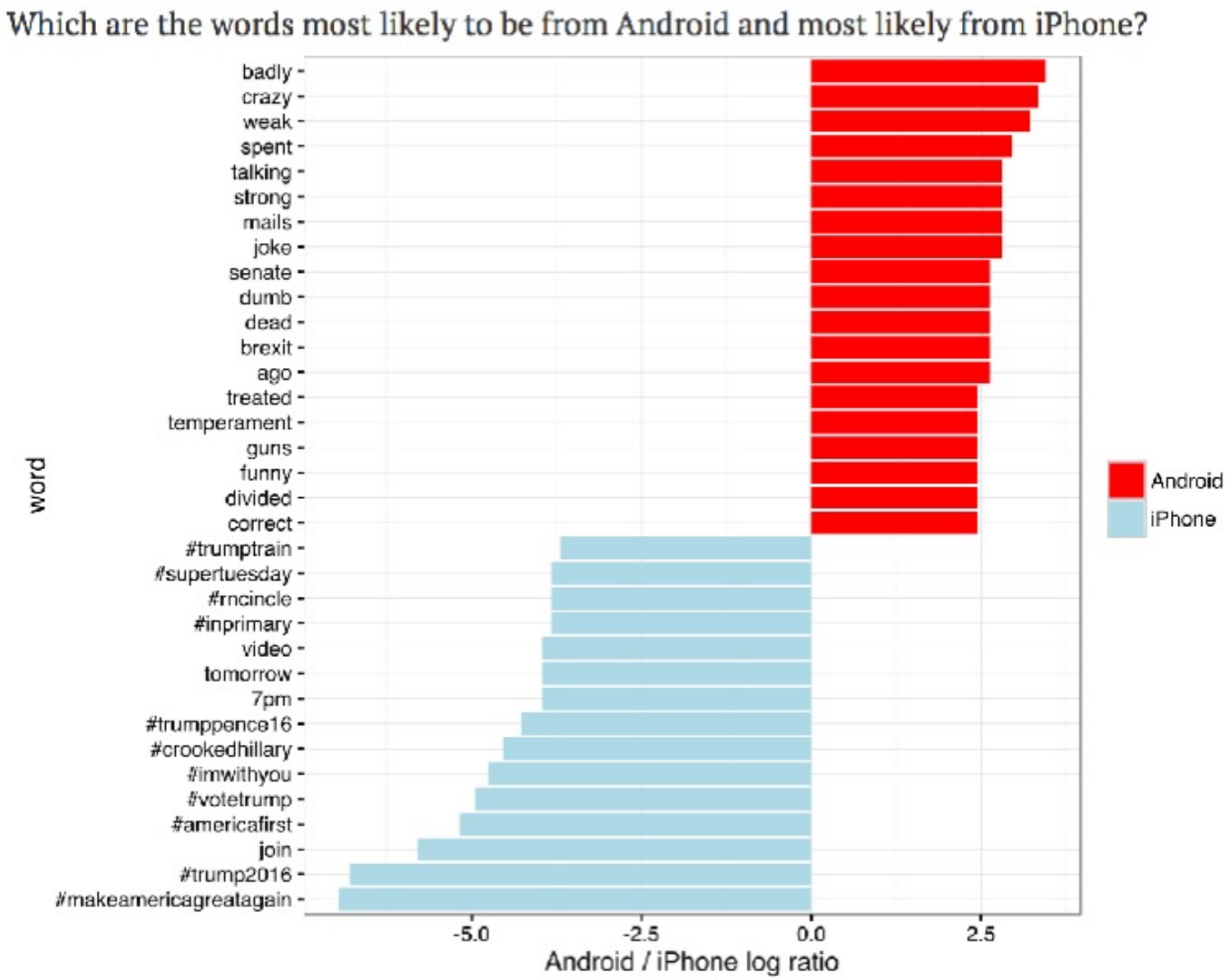
Gentzkow and Shapiro (2006), “What Drives Media Slant? Evidence from U.S. Daily Newspapers,” Econometrica



Schwartz et al. (2013), "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach



https://twitter.com/tvaziri/status/762005541388378112/photo/1 - https://twitter.com/tvaziri/status/762005541388378112/photo/2



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

Distinctive terms

- Finding distinctive terms is useful:
  - As a pre-processing step of feature selection.
  - As a data exploration exercise to understand larger trends in individual word differences).

- When the two datasets are **A** and **¬A**, these terms also provide insight into what A is about.

- Many methods for finding these terms!
  (Developed in NLP, corpus linguistics, political science, etc.)

Difference in proportions

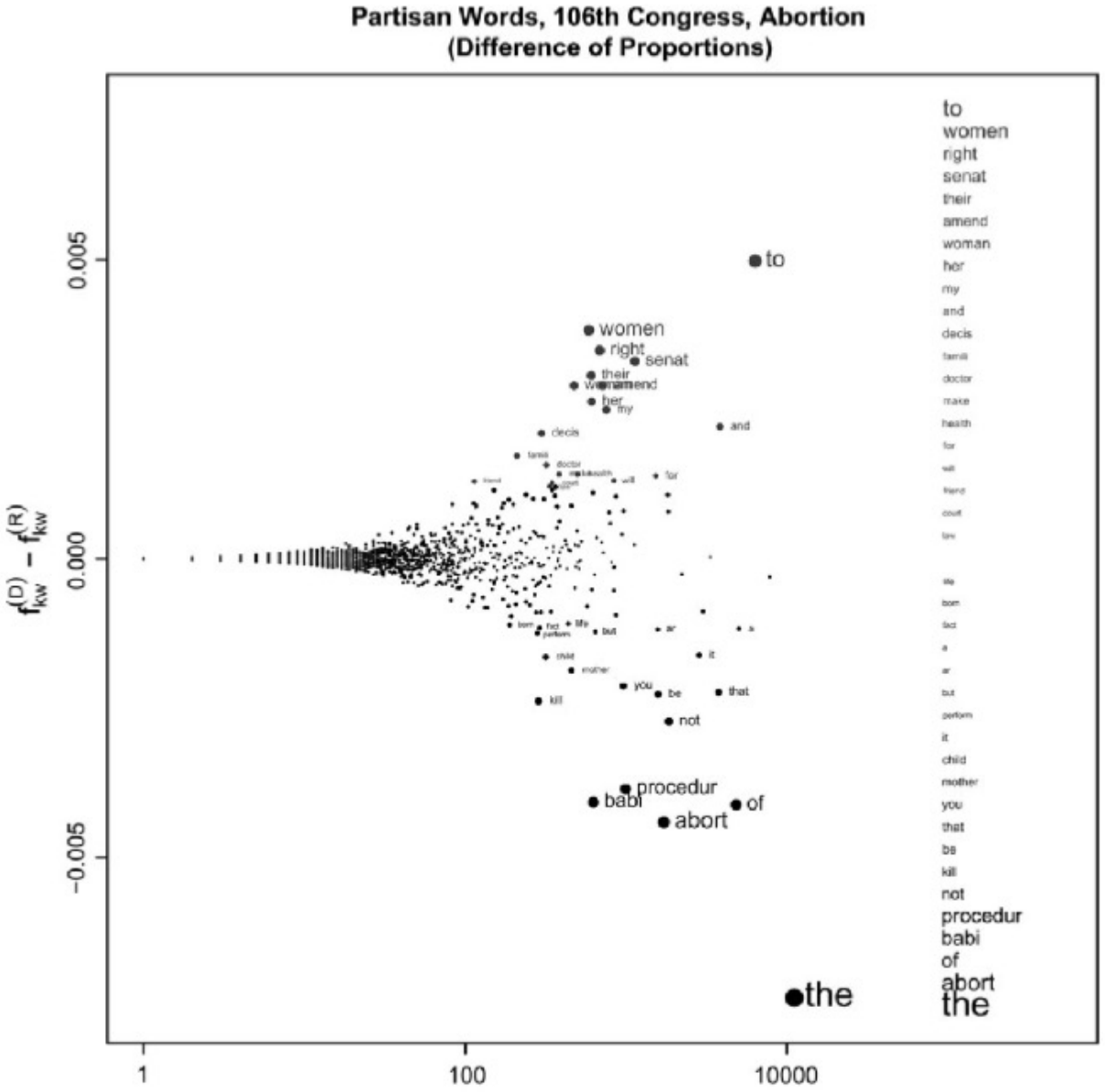
- For word w written by author with label k (e.g., {democrat, republican}), define the frequency to be the normalized count of that word

$$f_{w,k} = \frac{C(w,k)}{\sum_{w'} C(w',k)}$$

count of word w in group k

count of all words in group k

$$f_{w,k=dem} - f_{w,k=repub}$$



Monroe et al. (2009), “Fightin’ Words”

Difference in proportions

- The difference in proportions is a conceptually simple measure and easily interpretable.
- Drawback: tends to emphasize words with high frequency (where even comparatively small differences in word usage between groups is amplified).
- Also, no measure whether a difference is statistically meaningful. We have **uncertainty** about the what the true proportion is for any group.

X^2

- X^2 (chi-square) is a statistical test of dependence – here, dependence between the two variables of word identity and corpus identity.
- For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

	word	¬word
corpus 1	7	104023
corpus 2	104	251093

X^2

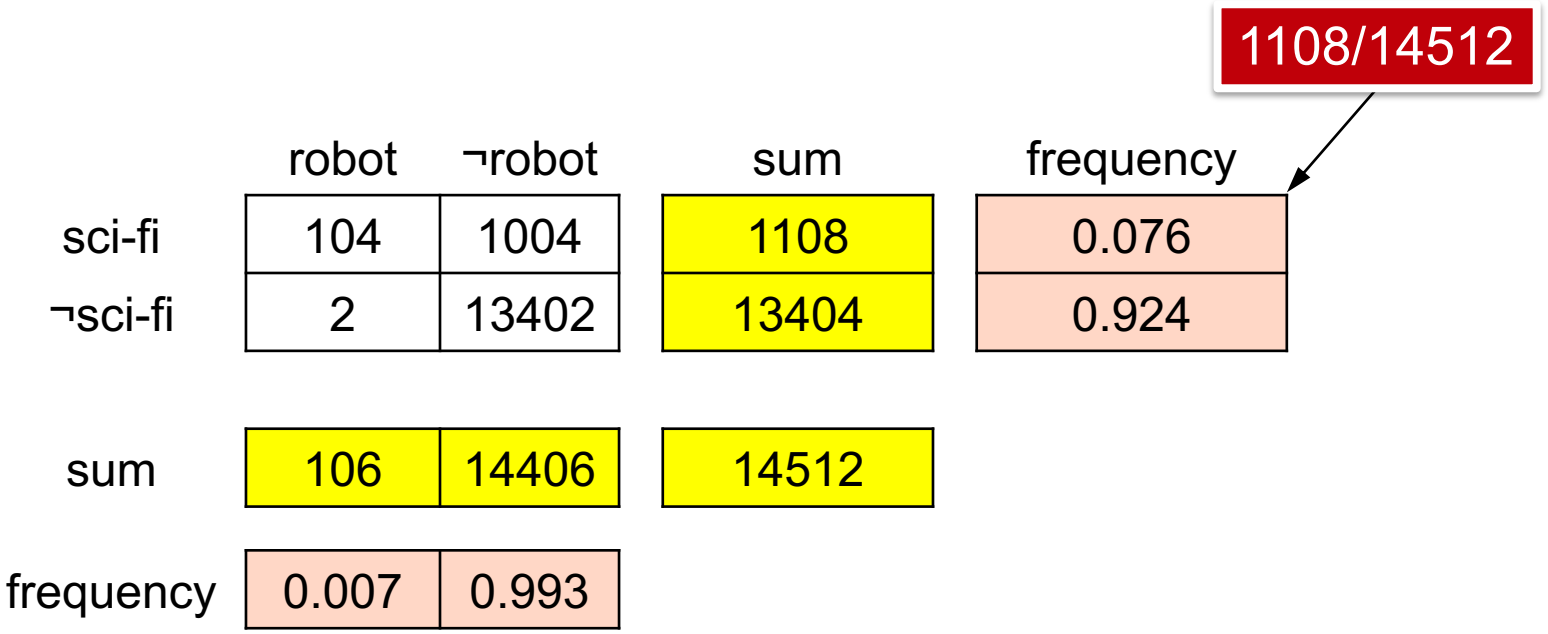
- Does the word *robot* occur **significantly** more frequently in science fiction?

	robot	¬robot	
sci-fi	104	1004	= 10.3%
¬sci-fi	2	13402	= 0.015%

X^2

- For each cell in contingency table, sum the squared difference between **observed value (O_{ij})** in cell and the **expected value (E_{ij})** assuming independence.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$


1108/14512
robot
¬robot
sum
frequency
sci-fi
104
1004
1108
0.076
¬sci-fi
2
13402
13404
0.924
sum
106
14406
14512
frequency
0.007
0.993

Assuming independence

- $P(\text{robot}, \text{scifi}) = P(\text{robot}) \times P(\text{scifi})$
  $= 0.007 \times 0.076$
  $= 0.00053$

- Among 14,512 words, we would expect to see 7.69 occurrences of *robot* in sci-fi texts.

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

	frequency
P(sci-fi)	0.076
P(¬sci-fi)	0.924

	P(robot)	*P*(¬robot)
frequency	0.007	0.993

X^2

- What X^2 is asking is:
  How different are the observed counts from the counts we would expect given complete independence?

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

X^2

- With algebraic manipulation, simpler form for 2x2 table O (cf. Manning and Schütze 1999)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

- $N = O_{11} + O_{12} + O_{21} + O_{22}$

	robot	¬robot
sci-fi	O_{11}	O_{12}
¬sci-fi	O_{21}	O_{22}

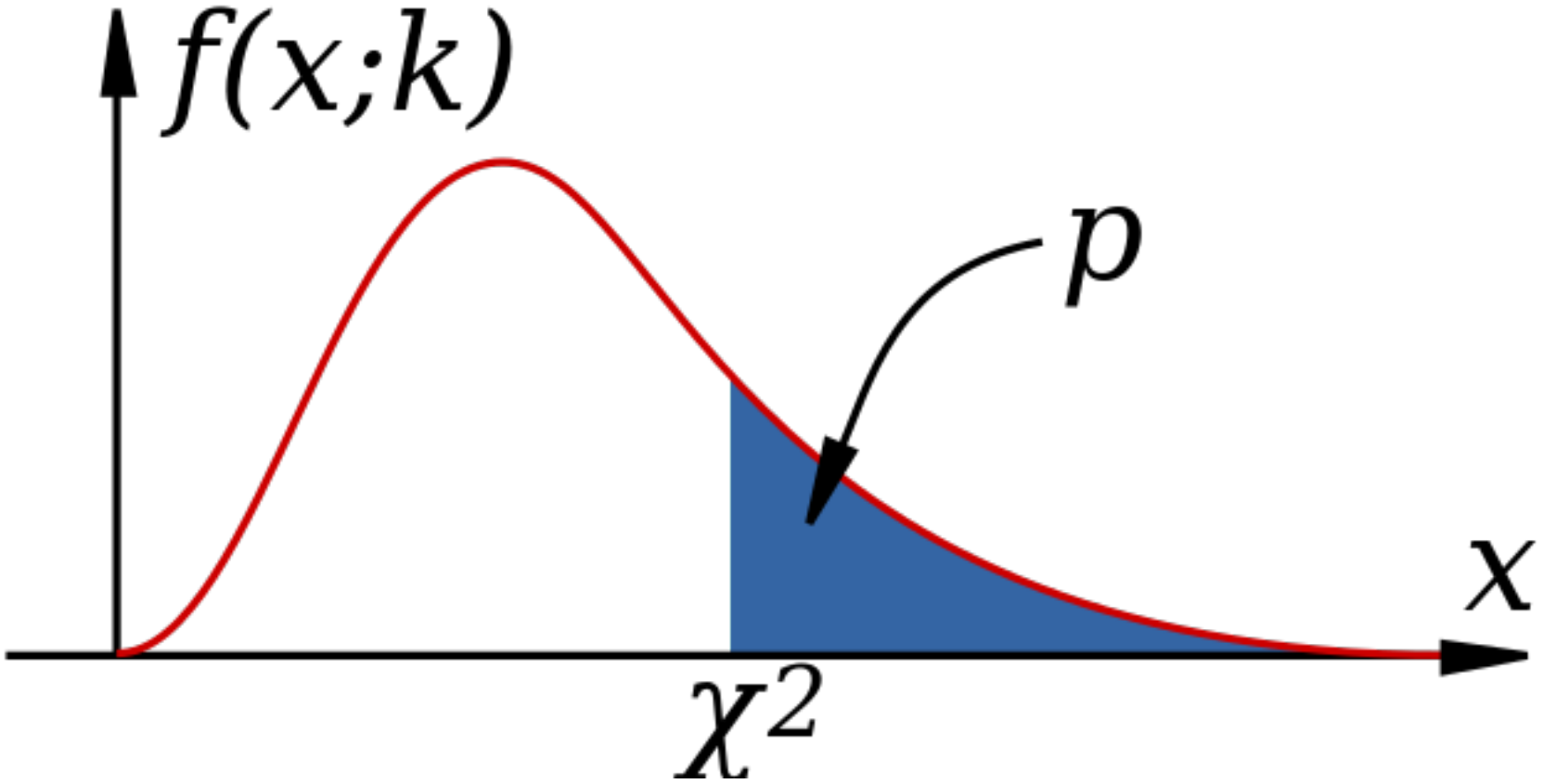
- The chi-square statistic X^2 is **1239.5312**.

X^2

- The X^2 value is a statistic of dependence with a probability governed by a **X^2 distribution**
- If this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.
- k = degrees of freedom (#rows-1)(#cols-1)

k	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	**3.841**	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

X^2 distribution



Inductiveload, Public domain, via Wikimedia Commons

X^2

- Chi-square is **ubiquitous** in corpus linguistics (and in NLP as a measure of collocations).

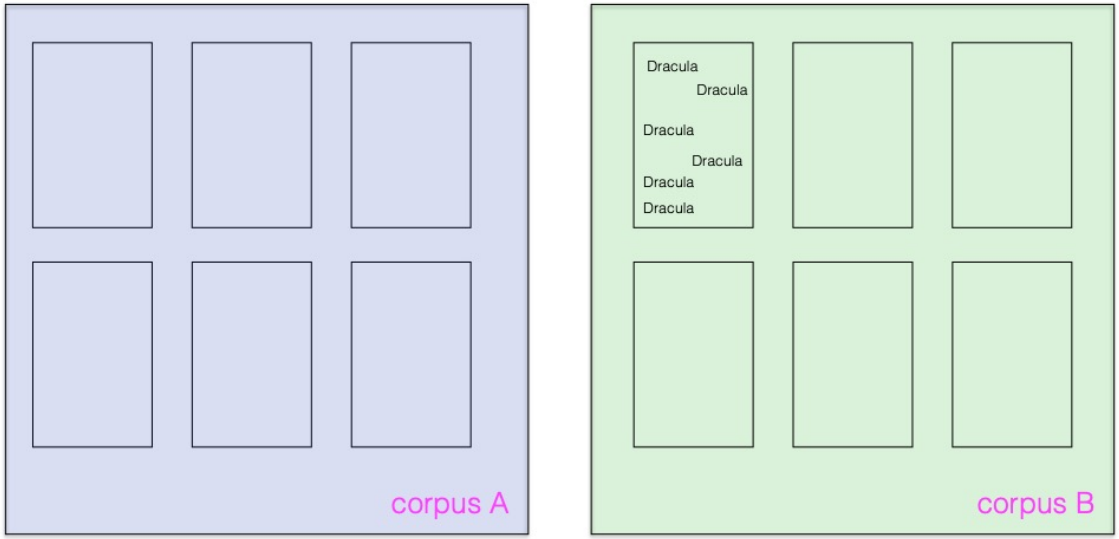
- A few caveats for its use:
  - Each cell should have an *expected* count of at least 5
  - Each observation is independent

- A drawback, however is due to the burstiness of language: the tendency for the same words to clump together in texts.

- Chi-square is testing for independence of two variables (word identity and corpus identity), but it **assumes** each mention of the word is independent from the others.

X^2

- Is Dracula really a word that distinguishes these two corpora?
- It distinguishes one text, but otherwise doesn't appear in the corpus at all.



- Alternatives to X^2: Mann-Whitney or G-Test

Mental homework

- Hypothesize terms that will be different between groups in the ESUPOL dataset (male vs. female, CDU vs. SPD, etc.)
- Execute chi-square to find terms that are different
- Online X^2 https://www.socscistatistics.com/tests/chisquare/default2.aspx