

# Computational text analysis: A brief introduction



© 2019

**TextXD**, Dec. 3, 2019

*Jaren Haber*

With special thanks to Ben Gebre-Medhin, Laura Nelson,  
Geoff Bacon, and Caroline Le Pennec-Caldichoury



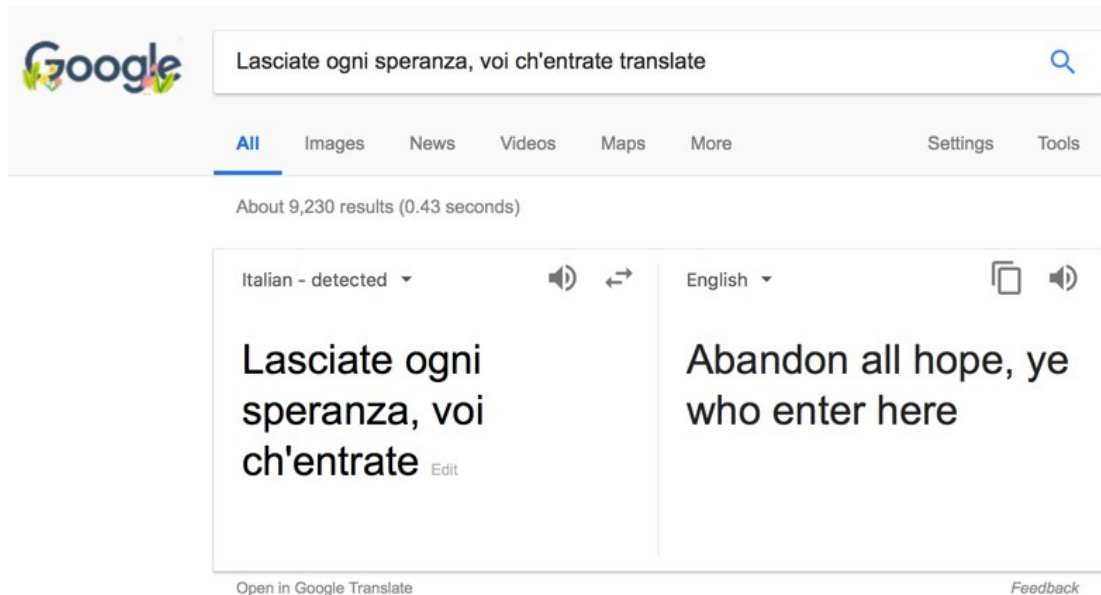
# Introductions

- Instructor
- Participants
  - Name
  - Affiliation
  - Any immediate projects or use cases? (Briefly)
  - Familiarity with Python or R? (Beginner, Intermediate, Advanced)
  
- If you do have immediate projects, bring them in!

# Goals of this workshop

- Provide a general roadmap of computational text analysis (CTA)
- Build intuitions about using text as data
- Gain practice with preprocessing and more
- Understand at a high-level
  - how a few primary CTA methods work
  - what kinds of questions they answer
  - how to design and implement a CTA project

# Machine translation



The image shows a screenshot of the Google Translate web interface. At the top left is the Google logo. To its right is a search bar containing the text "Lasciate ogni speranza, voi ch'entrate translate". Below the search bar are navigation links: "All" (underlined), "Images", "News", "Videos", "Maps", "More", "Settings", and "Tools". Below these links, it says "About 9,230 results (0.43 seconds)". The main translation area is divided into two columns. The left column is labeled "Italian - detected" and contains the text "Lasciate ogni speranza, voi ch'entrate" with an "Edit" link. The right column is labeled "English" and contains the translation "Abandon all hope, ye who enter here". At the bottom left, there is a link "Open in Google Translate", and at the bottom right, there is a "Feedback" link.

Google

Lasciate ogni speranza, voi ch'entrate translate

All Images News Videos Maps More Settings Tools

About 9,230 results (0.43 seconds)

Italian - detected

Lasciate ogni speranza, voi ch'entrate [Edit](#)

English

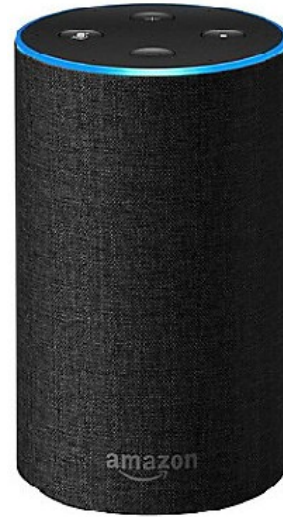
Abandon all hope, ye who enter here

[Open in Google Translate](#) [Feedback](#)



# Speech Recognition

“Alexa, how many cups are  
in a quart?”



# Question Answering



when was the last total eclipse in the united states



All

News

Images

Videos

Shopping

More

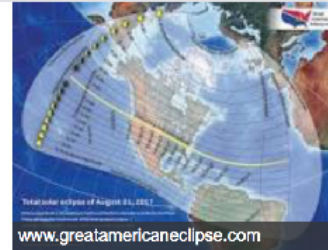
Settings

Tools

About 63,600,000 results (0.72 seconds)

## August 21, 2017

See more photos of the **August 21** eclipse. Bottom line: After the **August 21, 2017**, eclipse, the next total solar eclipse visible from North America will be **April 8, 2024**. Jul 5, 2018



[When's the next total solar eclipse for North America? | Astronomy ...](https://earthsky.org/astronomy-essentials/whens-the-next-total-solar-eclipse-in-the-us)

<https://earthsky.org/astronomy-essentials/whens-the-next-total-solar-eclipse-in-the-us>



# Software/Libraries



spaCy



NLTK





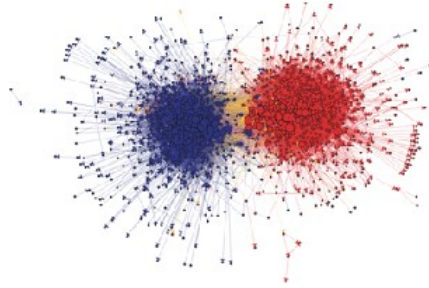
# NLP is interdisciplinary

- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)





## Computational Social Science



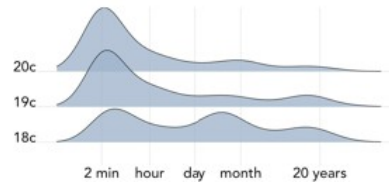
Adamic and Glance 2005

## Computational Journalism



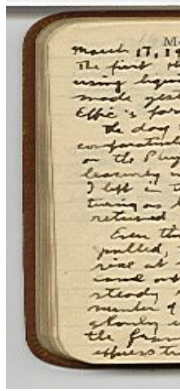
Change in insured Americans under the ACA,  
NY Times (Oct 29, 2014)

## Computational Humanities



Underwood 2018

# Text as data

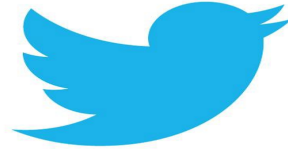


**Exhibit Feedback**

**1. Please explain below:**

What did you think overall?

What would you improve?



# Types of languages

- Natural languages

*Time flies like an arrow. Fruit flies like a banana.*

- Artificial languages

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + E$$

```
import scipy
from scipy import sparse

n = 200000
matrix = scipy.sparse.rand(n,
n, density=.001)  print(matrix)
```

# How do humans analyze texts?

*We need to steer clear of this poverty of ambition, where people want to drive fancy cars and wear nice clothes and live in nice apartments but don't want to work hard to accomplish these things. Everyone should try to realize their full potential.*

- Barack Obama



Close reading

# The promise of distant reading

- Scale/speed
- Reproducibility
- Does not have human biases
- Has other (at times unknown) biases
- Consistent

# A simple representation of text

Corpus of *documents*

Objective: map raw text of each document  $i$  to some attribute  $v_i$

The estimated attribute  $\hat{v}$  can then be used for descriptive or causal analysis:

- ▶ causal effect of racial animus predicted from Google search data on vote for Obama (Stephens-Davidowitz (2014))

# Movie revenues

Input: text of movie  
review

Output: box office  
revenue



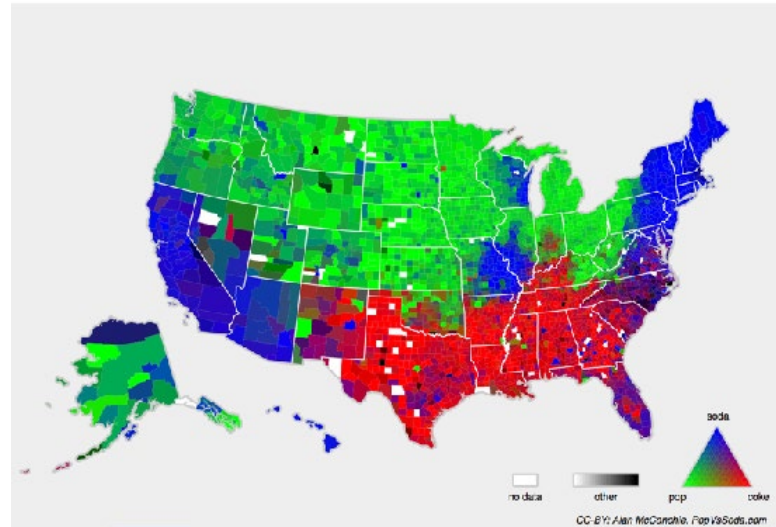


# Geographical location

## POP vs SODA

Input: tweet

Output: latitude,  
longitude



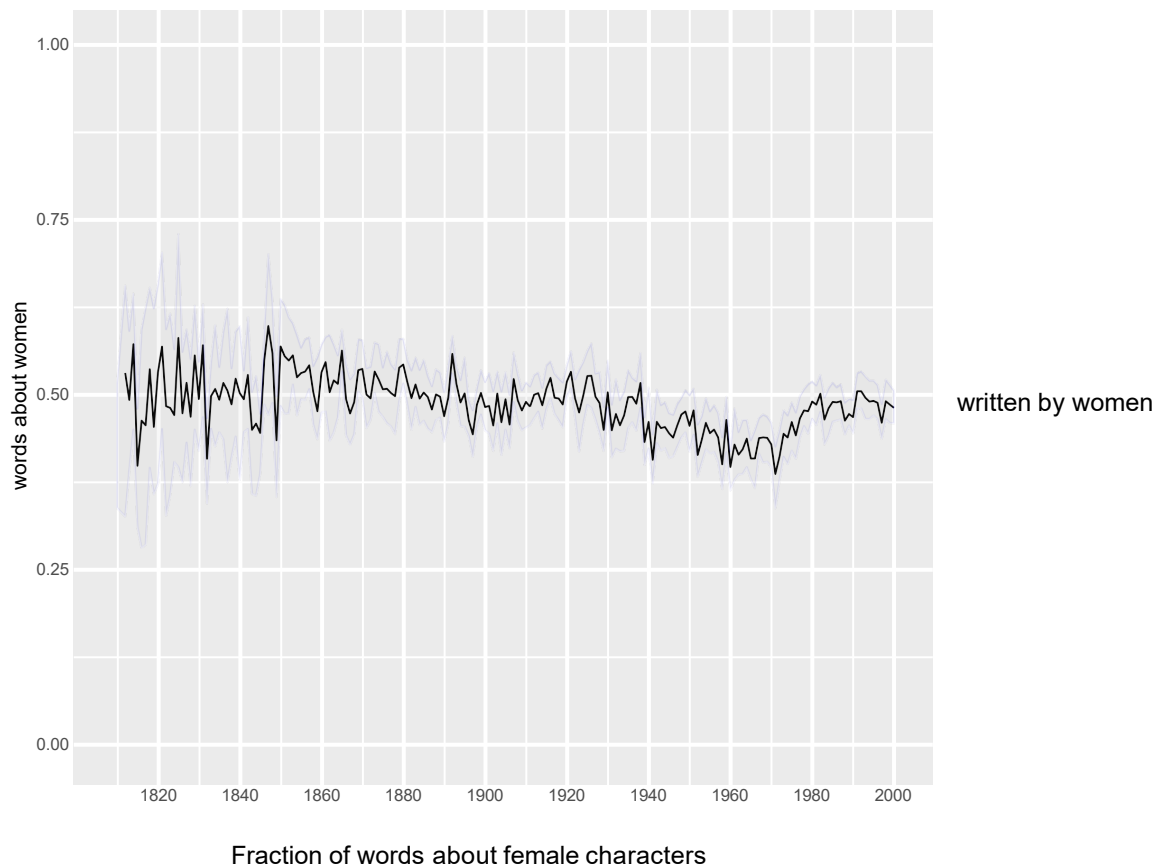
<http://popvssoda.com>

Wing and Baldrige (2011), "Simple supervised document geolocation with geodesic grids" (ACL)

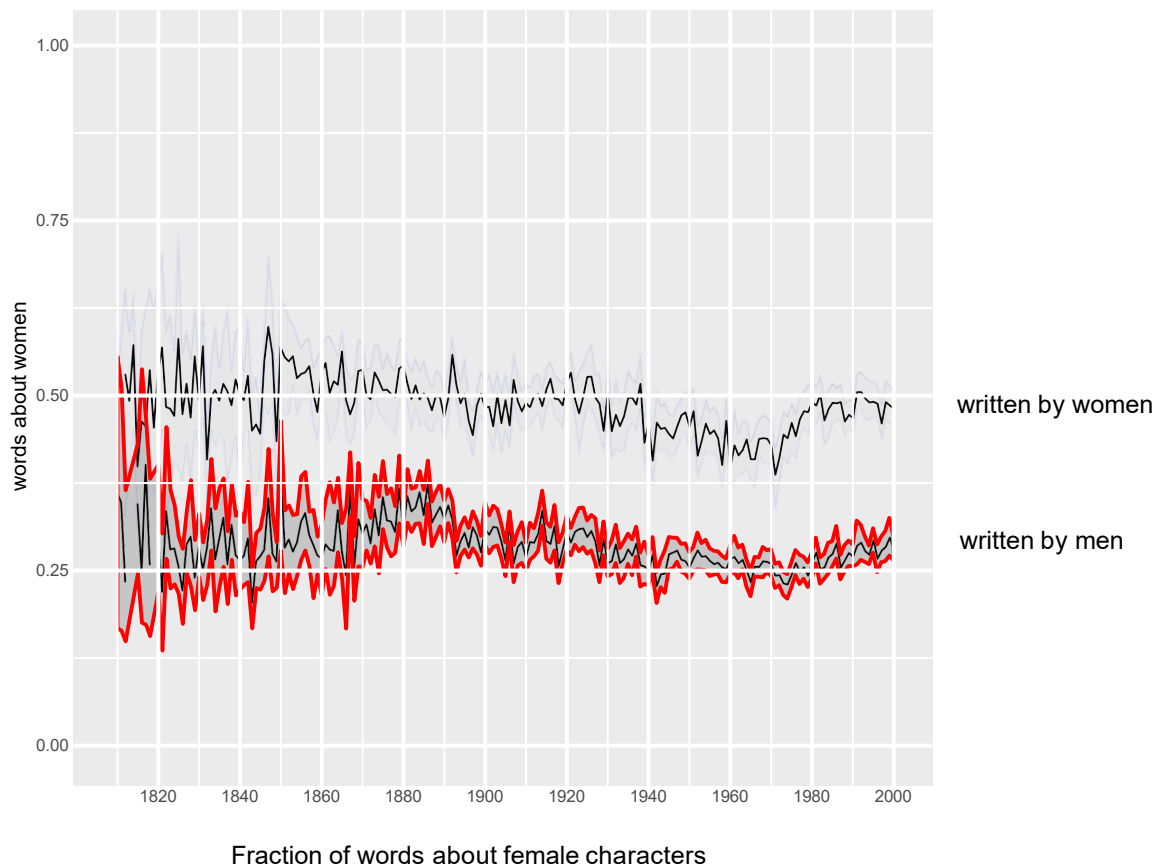




- Data: Random acts of pizza (subreddit)
- Response: Is a request successful in getting a pizza?



Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)



Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)

# CTA lifecycle



# CTA lifecycle



- Cooking analogy

# Research question

- Domain specific
  - Possible answer is encoded in text
- 
- E.g. *What are the early warning symptoms of depression?*
  - *How did different European nations react to the election of Trump?*
  - *Do Twitter users react differently to mass shootings based on the ethnicity of the perpetrator?*
  - *What distinguishes different styles of hip-hop?*
  - *Have any of these essays been plagiarized?*

# Getting data

What do we want?

- Plain, machine-readable text

How do we get it?

- You (your collaborator or a stranger on the Internet) already have it\*
- Web scraping
- API
- OCR
- \*Still important to know how the data was collected

# What is preprocessing?

- Tokenization = separating running text into words
- Sentence segmentation/tokenization = separating words into sentences
- Text normalization = dealing with upper/lower case, spelling mistakes, removing special characters, replacing URLs, numbers, etc.
- Remove “stop words”
- Stemming/lemmatization = removing morphological affixes
- POS tagging = assigning a part-of-speech category to each word
- Syntactic parsing = assigning a (normally graph or tree) structure to a sentence
- Chunking = shallow version of syntactic parsing
- Named entity recognition = identifying the proper nouns in a text
- ...



# Why do we do preprocessing?

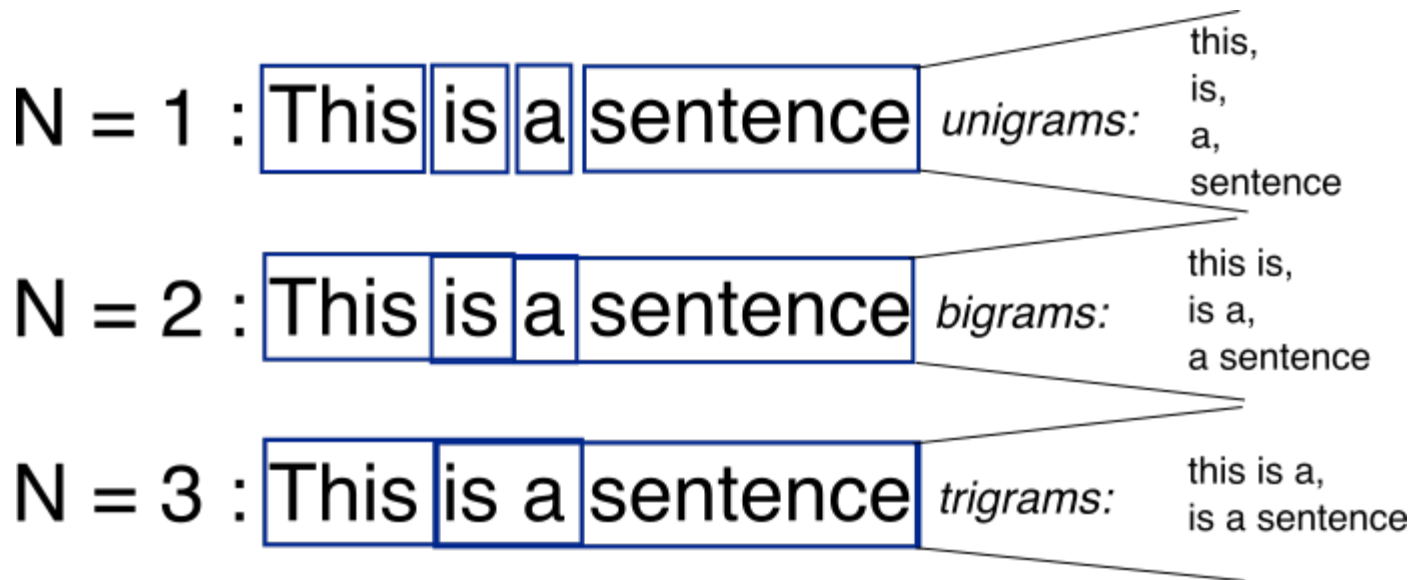
- Because later methods require preprocessed data as input
  - Counting words requires having already identified the words of a text
  - Knowing the POS of a word might help us in knowing whether it is important (modal *can* vs noun *can*)
- Because we gain intuition about our data
  - It forces us to look at the data
  - Often coupled with exploratory data analysis (EDA)
  - We might find out that all the reviews are exactly 500 characters long, which suggests some have been truncated.

# Inputs to modeling

- Topic modeling: input = many different texts, output = what each text is about
  - Newspaper articles
  - Emails
- Classification: input = many different texts and hand-labeled categories, output = something that can take in unlabeled texts and predict the category.
  - Hand label a bunch of documents, train a computer to mimic your hand coding
  - Spam /ham
  - positive/negative reviews
- The big difference is the need for labeled data (unsupervised vs supervised)
- Text = document, could be a review, newspaper article, journal article, whole book, tweet, ...



# What are N-grams?





# Term-document matrix

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

Context = appearing in the same document.

# Vector

Vector  
representation of  
the **document**;  
vector size =  $V$

Hamlet
1
2
17
64
75

King Lear
2
12
17
48
44

# Vectors

knife	1	1	4	2		2		2
-------	---	---	---	---	--	---	--	---

sword	17	2	7	12		2		17
-------	----	---	---	----	--	---	--	----

Vector representation of the  
**term**; vector size = number  
of documents

# The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



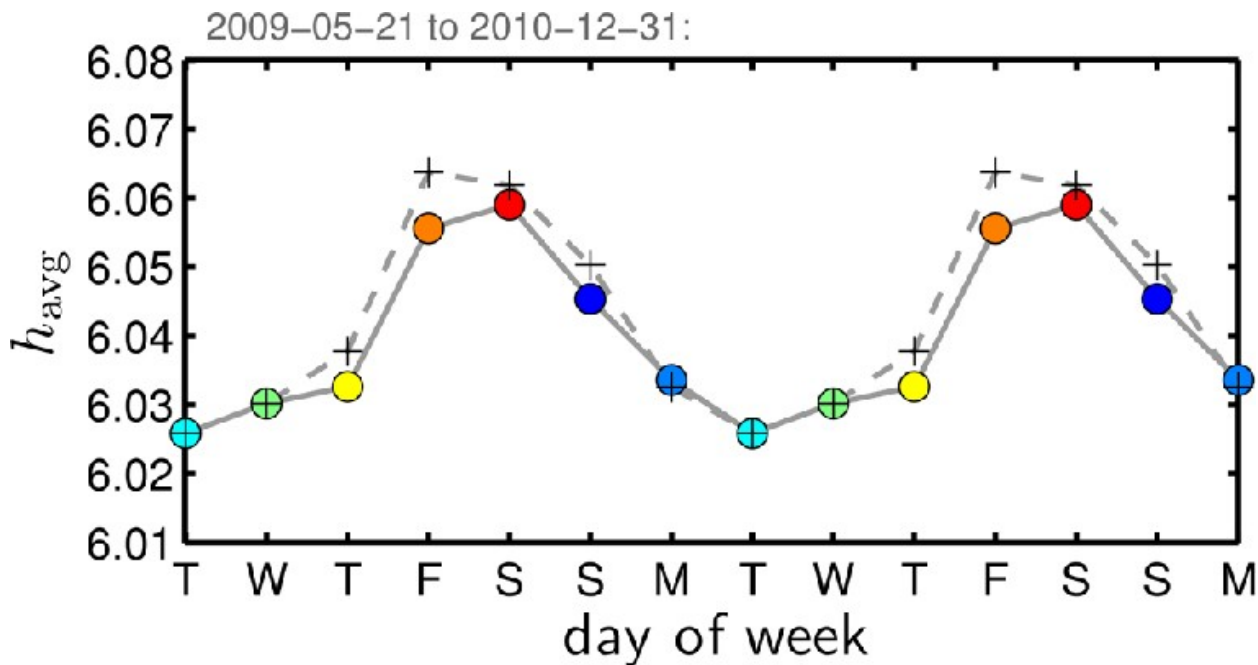
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

# But there's more out there

- Dictionary methods
  - Lists of positive/negative words
- Document-Term Matrix (DTM)
  - Words are rows, documents are columns, entries are number of times word appears in document.
- Term Frequency-Inverse Document Frequency (TF-IDF)
  - Modification of DTM
  - Entries are scaled by how common a word is across the whole corpus
- Distinctive words
  - Through difference of proportions, Chi-square test, classification, etc.
- Clustering
  - Of DTM or TF-IDF



# Dictionaries



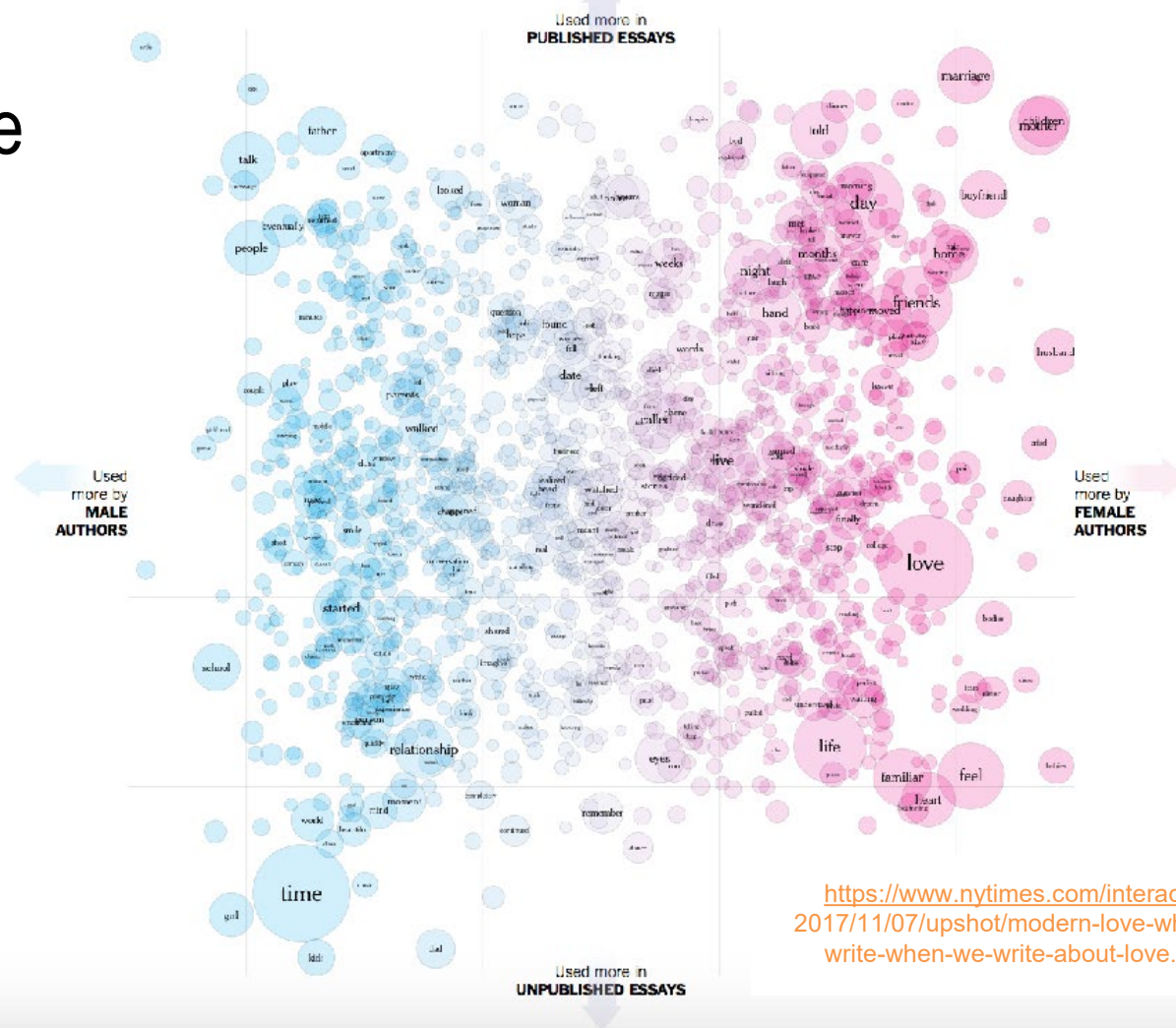
Dodds et al. (2011), "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter" (PLoS One)

LIWC			LIWC Cont.		
Category	Example	T-statistics	Category	Example	T-statistics
<b>Linguistics Processes</b>			Negative emotion	hurt, ugly, nasty	6.49***
Words > 6 letters		-3.41**	Anxiety	fearful, nervous	2.37
Dictionary words		9.60****	Anger	hate, kill, annoy	5.30***
Total function words		8.98****	Sadness	cry, grief, sad	3.54***
Personal pron.	I, them, her	7.07****	Cognitive process	cause, ought	6.09***
1st pers singular	I, me, mine	9.83****	Insight	think, know	0.11
1st pers plural	we, us, our	-2.38	Causation	effect, hence	0.93
2nd person	you, your, thou	-0.91	Discrepancy	should, would	5.53***
3rd pers singular	she, her, him	3.63**	Tentative	maybe, perhaps	5.95***
3rd pers plural	their, they'd	2.47	Certainty	always, never	4.02***
Impersonal pron.	it, it's, those	7.07****	Inhibition	block, constrain	0.32
Articles	a, an, the	4.13***	Inclusive	with, include	4.74 ***
Common verbs	walk, went, see	6.27***	Exclusive	but, without	7.53 ****
Auxiliary verbs	am, will, have	5.76***	Perceptual process		1.93
Past tense	went, ran, had	8.70****	See	view, saw, seen	1.68
Present tense	is, does, hear	4.00***	Hear	listen, hearing	-0.88
Future tense	will, gonna	5.84***	Feel	feels, touch	1.94
Adverbs	very, really	7.92****	Biological process		4.22***
Prepositions	to, with, above	7.62****	Body	cheek, spit	5.02***
Conjunctions	and, whereas	4.59***	Health	clinic, flu, pill	1.51
Negations	no, not, never	1.71	Sexual	horny, incest	-0.61
Quantifiers	few, many, much	2.98*	Ingestion	dish, eat, pizza	4.37***
Numbers	second, thousand	-3.68**	Relativity	area, bend, exit	9.52 ****
Swear words	damn, piss, fuck	5.53***	Motion	arrive, car	3.07*
<b>Spoken Categories</b>			Space	down, in, thin	8.87****
Assent	agree, OK, yes	7.05****	Time	end, until	5.87***
Nonfluency	er, hm, umm	1.41	<b>Personal Concerns</b>		
Filters	blah, imean		Work	job, majors	0.05
<b>Psychological</b>			Leisure	chat, movie	2.97*
Social process	mate, talk, child	0.10	Achievement	earn, win	-1.22
Family	son, mom, aunt	2.24	Home	family, kitchen	3.37**
Friends	buddy, neighbor	2.10	Money	audit, cash	0.23
Humans	adult, baby, boy	0.89	Religion	church, altar	-0.77
Affective process	happy, cry	3.55**	Death	bury, coffin	0.49
Positive emotion	love, nice, sweet	0.08			

Table 1. Two-sample T-test statistics of linguistic variables between geo-locator and non-locators. Significant differences of each LIWC attribute are indicated in the third column. (\*p < 0.01, \*\*p < 0.001, \*\*\*p < 0.0001, \*\*\*\*p < 1e-10)

Modeling

# Distinctive words



Modeling

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

# Now what?

- Use the insight from the preprocessing and modeling stage to understand the initial question
- In classification, this could be looking at words with high coefficients.
  - E.g. *People with depressive symptoms are more likely to talk about abstract concepts and use more negation.*
- In topic modeling, this could be qualitative inspection of the topics.
  - E.g. *In Germany the main themes centered around taxes and immigration, while in France people were more concerned about racism.*

# CTA lifecycle



Questions?





# Now to get our hands dirty

<http://bit.ly/intro-textxd19>

# Further resources

- D-Lab: <http://dlab.berkeley.edu/>
  - Regular workshops on CTA, Python, R, etc.
  - Consulting
- CTAWG: <http://dlabctawg.github.io/>
- [Lectures from Stanford's NLP class](#)
- [Tutorials on NLTK and SpaCy](#)
- Jurafsky and Martin [textbook](#)



# Further examples of CTA applications

- Estimate political ideology from Twitter
- Uncover government censorship
- Relate the stock market to sentiment in the media
- Study why particular papers get cited
- Detect impending disease epidemics
- Determine who actually wrote something
- ...