

Enter into browser: https://bit.ly/nlp-python-2020

GitHub repository: http://bit.ly/nlp-repo-2020

Jaren Haber, PhD October 19th, 2020 Massive Data Institute Georgetown University

Agenda today

- Introductions (~5 mins)
- Slides: Introduction to NLP (~40 mins)
- Break (5 mins)
- Notebook: Practice with preprocessing (~50 mins)

Introductions

- Instructor
- Participants (breakouts)
 - Name
 - Affiliation
 - Any immediate projects or use cases? (Briefly)
 - Familiarity with Python? (Beginner, Intermediate, Advanced)

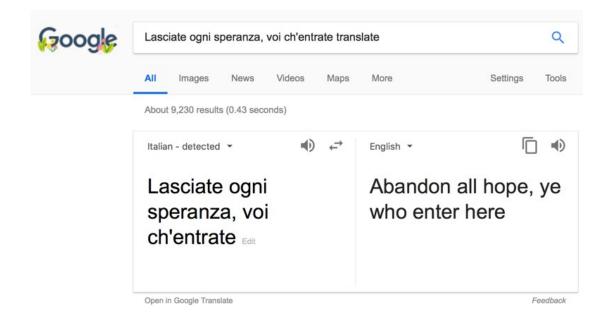
Goals of this workshop

- Build intuitions about using text as data
- 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
- Gain practice with:
 - preprocessing text data
 - dictionary methods
 - NLTK and Scikit-learn
- Acquire resources for further learning





Machine translation



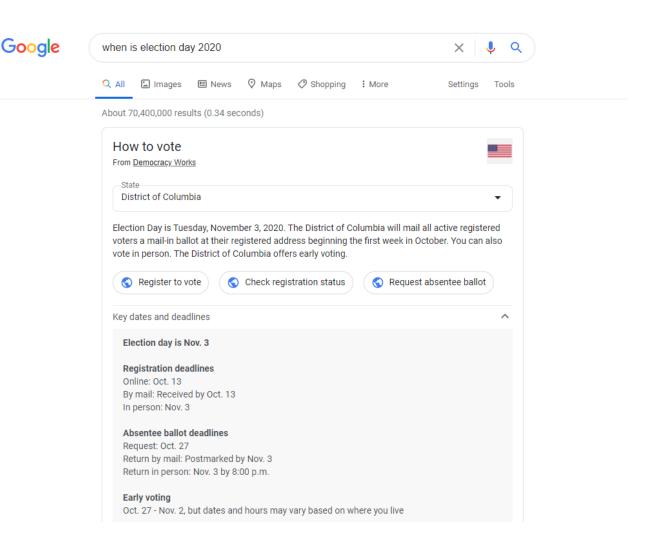


Speech recognition

"Alexa, do you believe in ghosts?"



Question answering





Software/Libraries















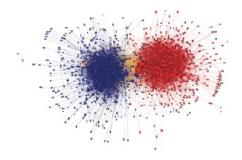


NLP is interdisciplinary

- 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



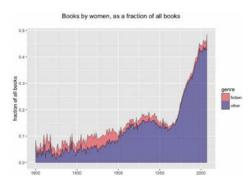
Adamic and Glance 2005

Computational Journalism



Covid in the U.S.: Latest Map and Case Count, NY Times (Oct 16, 2020)

Computational Humanities



Underwood et al. 2018

Text as data



march 17, 1926 with and are word first profession of a march of the profession of the control of

Extle said that of while the a fairy of an arthuris of constitution of the started off.

The stay was close, for the most boot, with large should be the start with large should be the start with last afternoon these last in the afternoon these



Exhibit Feedback

1. Please explain below:

What did you think overall?

What would you improve?



NOTES NOTES

Types of languages

- Natural languages

Time flies like an arrow. Fruit flies like abanana.

- Artificial languages

```
Y = f3_0 + f3_1X_1 + f3_2X_2 + f3_3X_3 + ... + f3_nX_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?

...as President, I can't be locked in a room someplace for the next year and just stay and do nothing. And every time I go into a crowd, I was with the parents of our fallen heroes... And they came up to me and they would hug me... and I'm not to not let them do it, to be honest with you.

Donald Trump (Oct. 15, 2020)



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 $\hat{v_i}$ can then be used for descriptive or causal analysis



Movie revenues

Input: text of movie review

Output: box office revenue



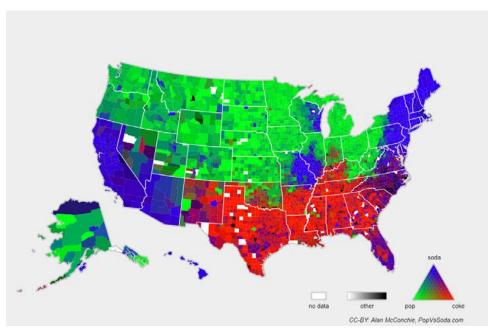


Geographical location

Input: tweet

Output: latitude, longitude

POP vs SODA



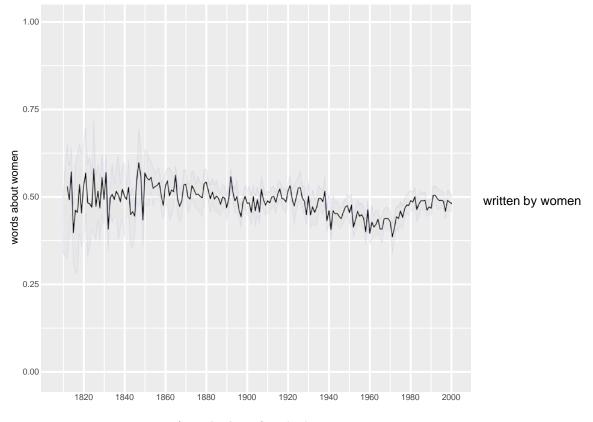
http://popvssoda.com





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

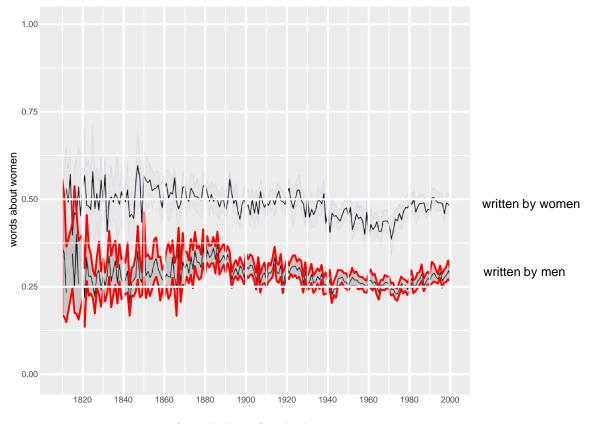




Fraction of words about female characters

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





Fraction of words about female characters

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

CTA lifecycle



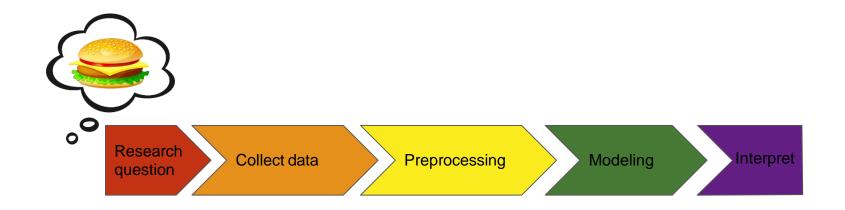


NLP research design





NLP research design





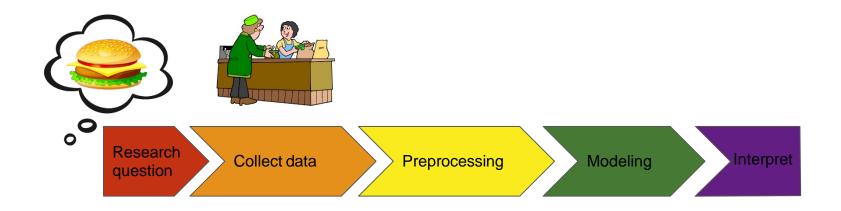
Research question

- 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?



NLP research design





Getting data

What do we want?

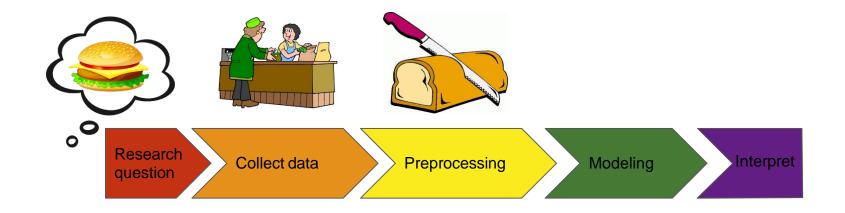
- Plain, machine-readable text

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



NLP research design



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.

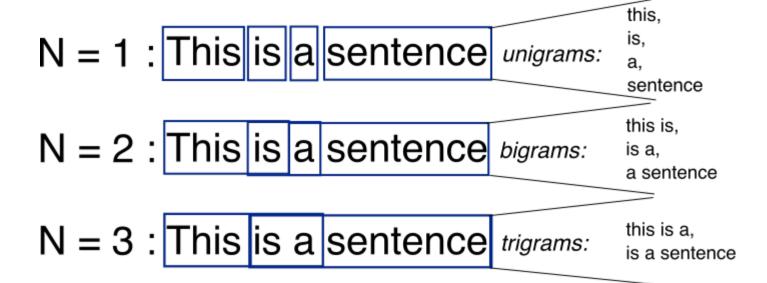
Preprocessing

Tokenization

```
<sentence>
<word>Friends</word</pre>
<word>don't</word>
<word>let</word>
<word>friends</word>
<word>make</word>
<word>word</word>
<word>clouds</word>
</sentence>
```



What are N-grams?





Document-term matrix

	Hamlet	Macbeth	Ro me o & Juli et	Richard III	J uli us Ca es ar	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.

Document vector

Vector representation of the document; vector size = V

Hamlet
1
2
17
64
75

King Lear	
2	
12	
17	
48	
44	

Word vector

knife	1	1	4	2	2	2
swor d	17	2	7	12	2	17

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

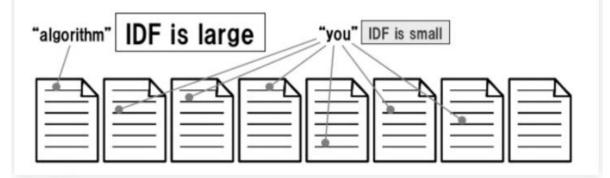


TF-IDF = Term Frequency...

Inverse Document Frequency (IDF)

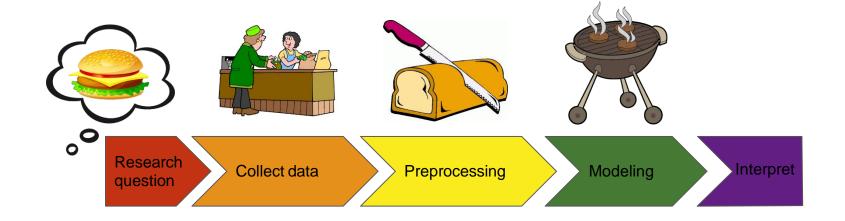
Give more weight to a term occurring in less documents

$$IDF(t) = \log \frac{|D|}{df(t)}$$
 $t : Term$
 $df(t) : Document frequency of t$
 $|D| : Number of documents in D$





NLP research design





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, ...

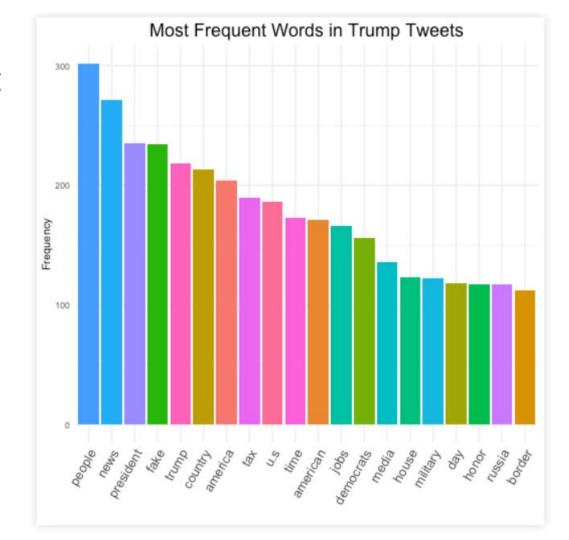


I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy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





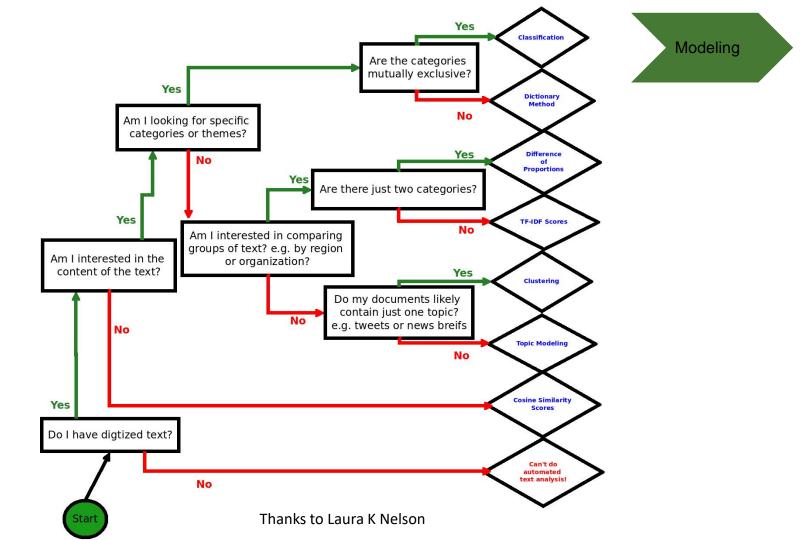
Frequent words



Common NLP modeling methods

- Dictionary methods
 - Lists of positive/negative words (sentiment analysis)
- Distinctive words
 - Through difference of proportions, Chi-square test, classification, etc.
- Classification
 - Use text features to sort into categories, e.g. spam/not spam
- Topic modeling
 - Mixture of words over topics and topics over documents
- Word embeddings
 - Each word represented by numerical vector
- Clustering
 - Of DTM or TF-IDF, or topics, or word embeddings





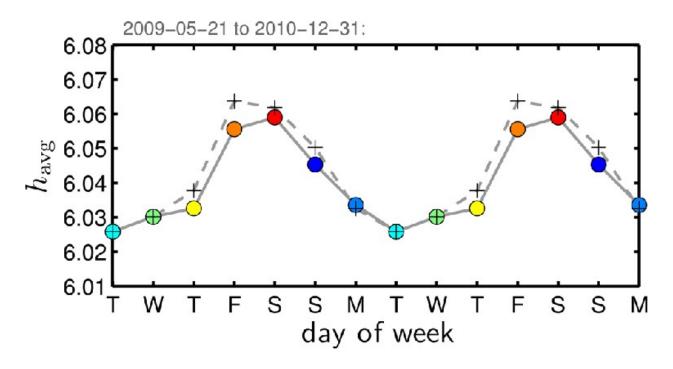


Dictionary method





Dictionary method



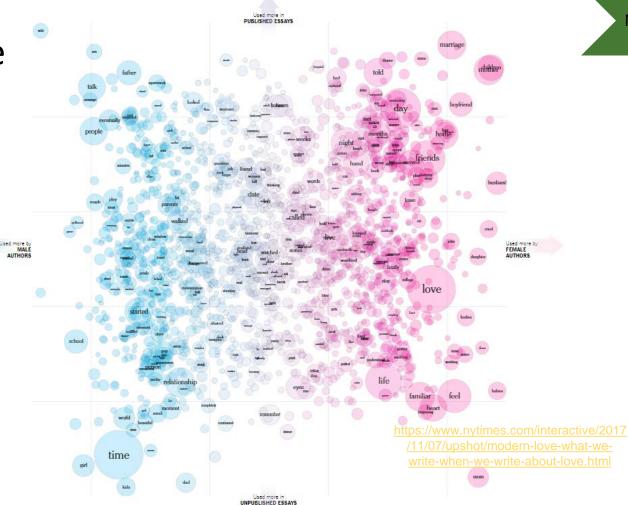
Dodds et al. (2011), "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter" (PLoS One)

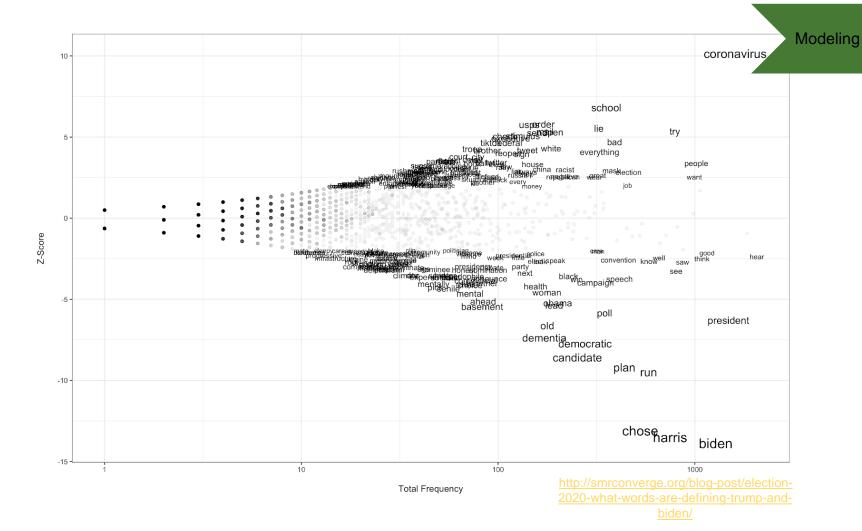
	LIWC		LIWC Cont.			
Category	Example	T-statistics	Category	Example	T-statistics	Mode
Linguistics Processes			Negative emotion	hurt, ugly, nasty	6.49***	Wiod
Words > 6 letters		-3.41**	Anxiety	fearful, nervous	2.37	
Dictionary words		9.60****	Anger	hate, kill, annoy	5.30***	
Total function words		8.98****	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*	
Spoken Categories	, 1		Space	down, in, thin	8.87****	
Assent	agree, OK, yes	7.05****	Time	end, until	5.87***	
Nonfluency	er, hm, umm	1.41	Personal Concerns	•		
Filters	blah, imean		Work	job, majors	0.05	
Psychological	•		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		V /		

How Certain We Are That A Word Was ...

Modeling

Distinctive words







Classification



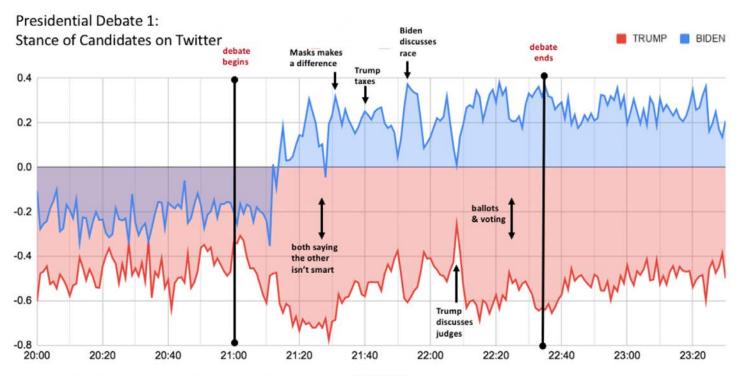


Spam

Not spam



Classification



This analysis was conducted using 1.3 million tweets that contained one of the debate hashtags. We determined if the tweet shows support, opposition, or neither for each candidate. For each minute, we compute an aggregate stance score: Stance Score = (# Support - # Oppose) / (# of tweets that minute having a stance)

Time (EST)

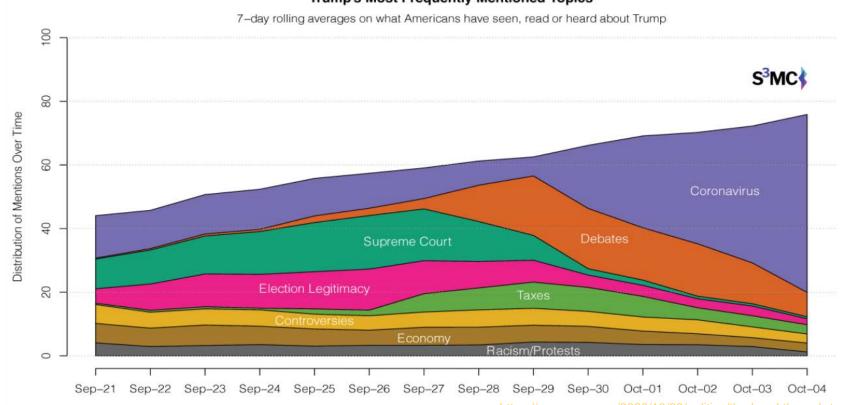
https://mccourt.georgetown.edu/news/pre sidential-debate-candidate-stanceanalysis/





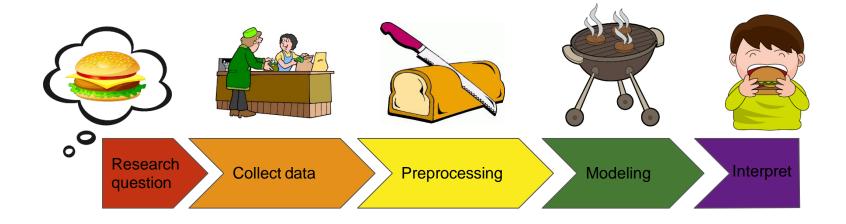
Topic modeling

Trump's Most Frequently Mentioned Topics





NLP research design





What you've been after...

- 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 morelikely 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.



To review...



Questions?



Now to get our hands dirty

http://bit.ly/nlp-repo-2020

Further resources

- Introduction to Jupyter Notebooks (Real Python)
- Quick Python intro (a Jupyter Notebook)
- Great book on Python (with exercises): "Python for Everybody" (Charles Severance)
- Official Python Tutorial
- NLP course & scripts, for social scientists & digital humanists (Laura Nelson)
- NLP textbook (Jurafsky & Martin @ Stanford)
- Book on NLTK (NLTK team)
- Datasets for NLP (Hugging Face)
- Intro to SpaCy and NLP concepts (Allison Parrish)
- Workshops on NLTK and SpaCy (Geoff Bacon @ D-Lab)

Further examples of CTA applications

- <u>Estimate political ideology from Twitter</u>
- 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
- ...