

Introduction to Natural Language Processing in the Social Sciences

LEARNING OUTCOMES

At the other end of this class, you should:

1. Be familiar with important terminology and ideas in NLP
2. Know what text analysis method you can use to answer your research question.
3. Know what steps are involved in preprocessing data for NLP
4. Know some common NLP methods
5. Know pros/cons with different ways of doing NLP



THIS WILL BE IN THE WORKSHOP!

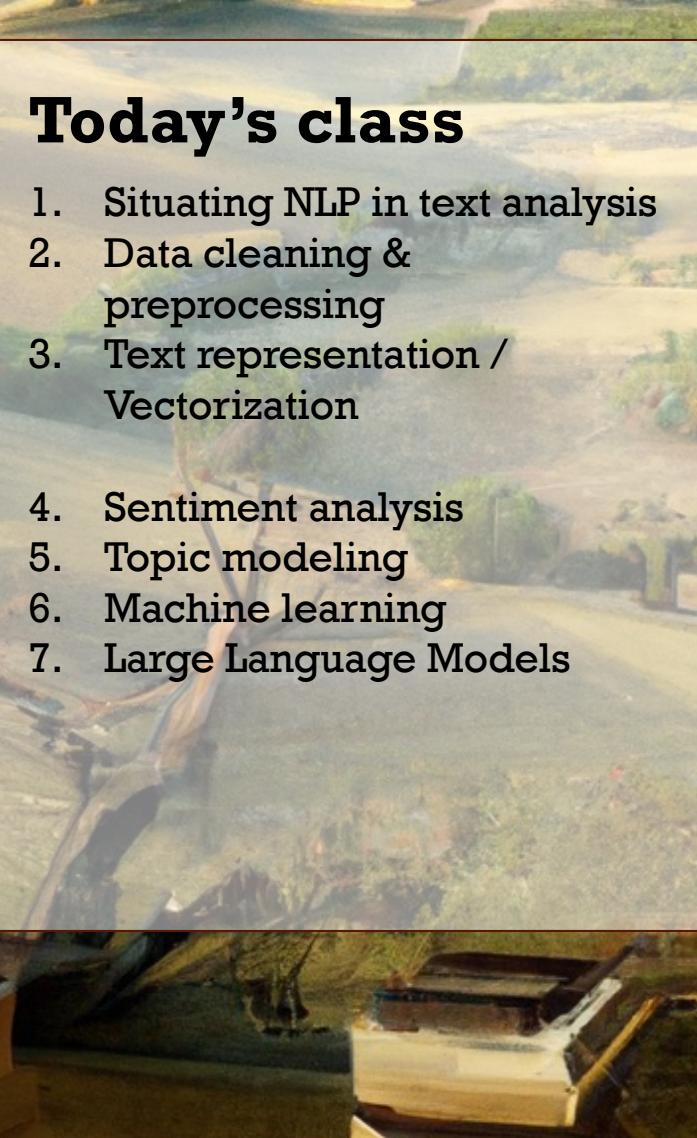
- I will present various data analyses of **Trump's tweets** in this lecture
- But I haven't done proper preprocessing
- Your job in the workshop will be to fix the preprocessing!

	id	text	isRetweet	isDeleted	device	favorites	retweets	date
0	98454970654916608	Republicans and Democrats have both created ou...	f	f	TweetDeck	49	255	2011-08-02 18:07:48
1	1234653427789070336	I was thrilled to be back in the Great city of...	f	f	Twitter for iPhone	73748	17404	2020-03-03 01:34:50
3	1304875170860015617	The Unsolicited Mail In Ballot Scam is a major...	f	f	Twitter for iPhone	80527	23502	2020-09-12 20:10:58
6	1223640662689689602	Getting a little exercise this morning	f	f	Twitter for iPhone	285863	30209	2020-02-01 16:14:02
7	1319501865625784320		f	f	Twitter for iPhone	130822	19127	2020-10-23 04:52:14

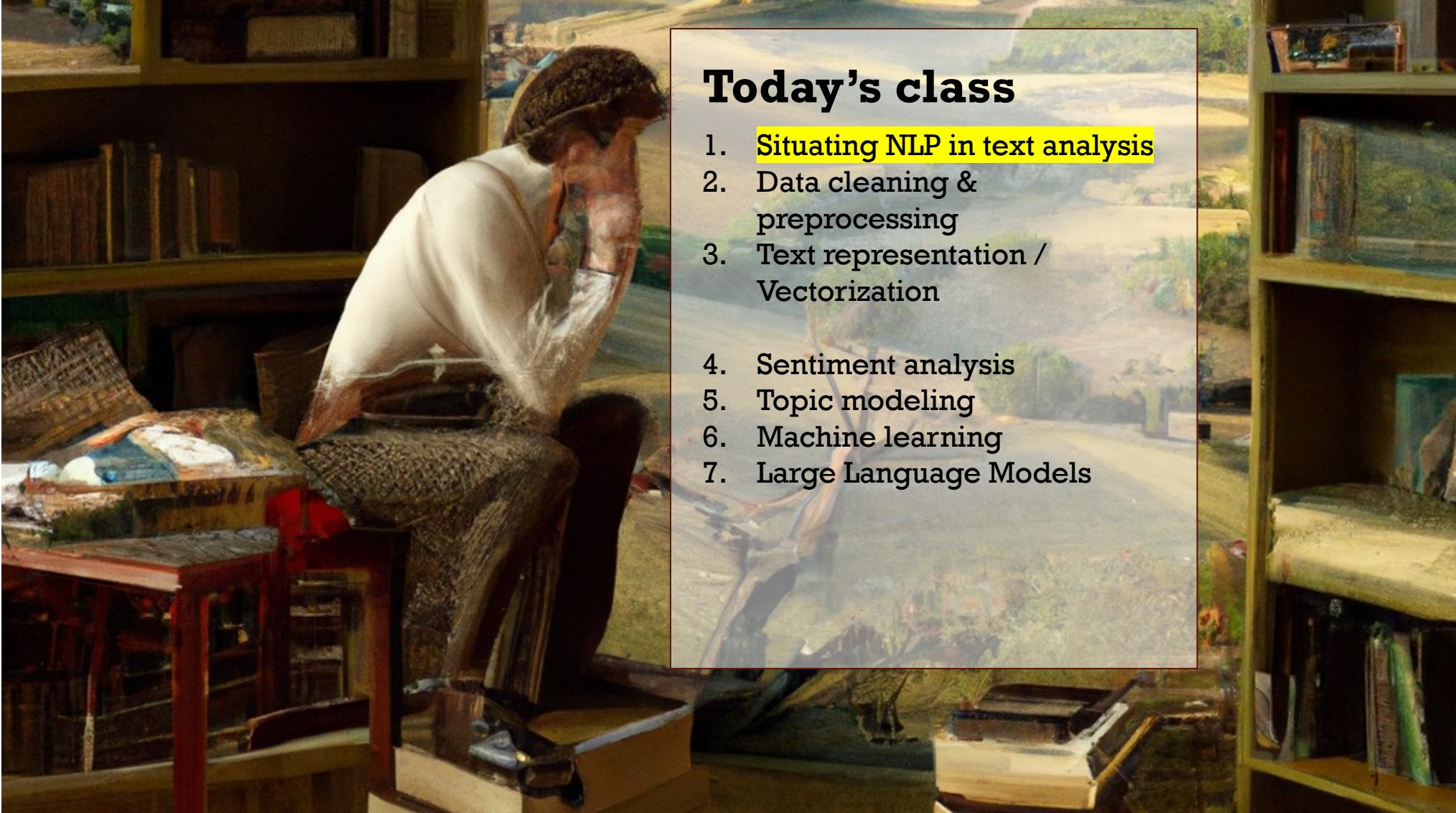




Today's class

1. Situating NLP in text analysis
 2. Data cleaning & preprocessing
 3. Text representation / Vectorization
 4. Sentiment analysis
 5. Topic modeling
 6. Machine learning
 7. Large Language Models
- 





Today's class

1. Situating NLP in text analysis
2. Data cleaning & preprocessing
3. Text representation / Vectorization
4. Sentiment analysis
5. Topic modeling
6. Machine learning
7. Large Language Models

SITUATING NLP

There are different traditions of text analysis in the social sciences.

1. Qualitative text analysis
2. Quantitative text analysis
3. Natural Language Processing



QUALITATIVE TEXT ANALYSIS

e.g. critical discourse analysis

“Read attentively and draw conclusions based on your theoretical knowledge”

- Michel Foucault: Power/knowledge through language.
 - Erving Goffman: Everyday conversation and "frame analysis".
 - Teun A. van Dijk & Norman Fairclough: Critical discourse analysis focusing on power and ideology.
-
- **Pro:** High reliability. Deep, theoretically grounded analysis.
 - **Con:** Low validity. Can't handle large data. Lack of rigor. Cherry-picking.



Examples qualitative research questions:

1. How did Donald Trump's Twitter discourse construct and represent his political identity and that of his adversaries during his presidency?
2. How did Donald Trump's Twitter messages utilize rhetorical strategies to frame discussions of immigration policy? How does his characterization of migrant groups fit into longer histories of *othering* in the US context?
3. How did Donald Trump's construct a separation between in-group and out-group among his political followers?
4. What discursive strategies did Donald Trump employ in relation to media and the press?



QUANTITATIVE METHODS

e.g. Corpus Linguistics

“Turn language into numbers and do statistics!”

Counting and Measuring:

- Conversion of qualitative data to quantitative metrics
- E.g., word frequency, sentence length, readability, word co-occurrence
- Chi-Square Tests. Z-score.
- Deductive: make hypotheses and test

Pro: High validity. Can be used on large text data.

Con: Low reliability, “shallow” analysis.

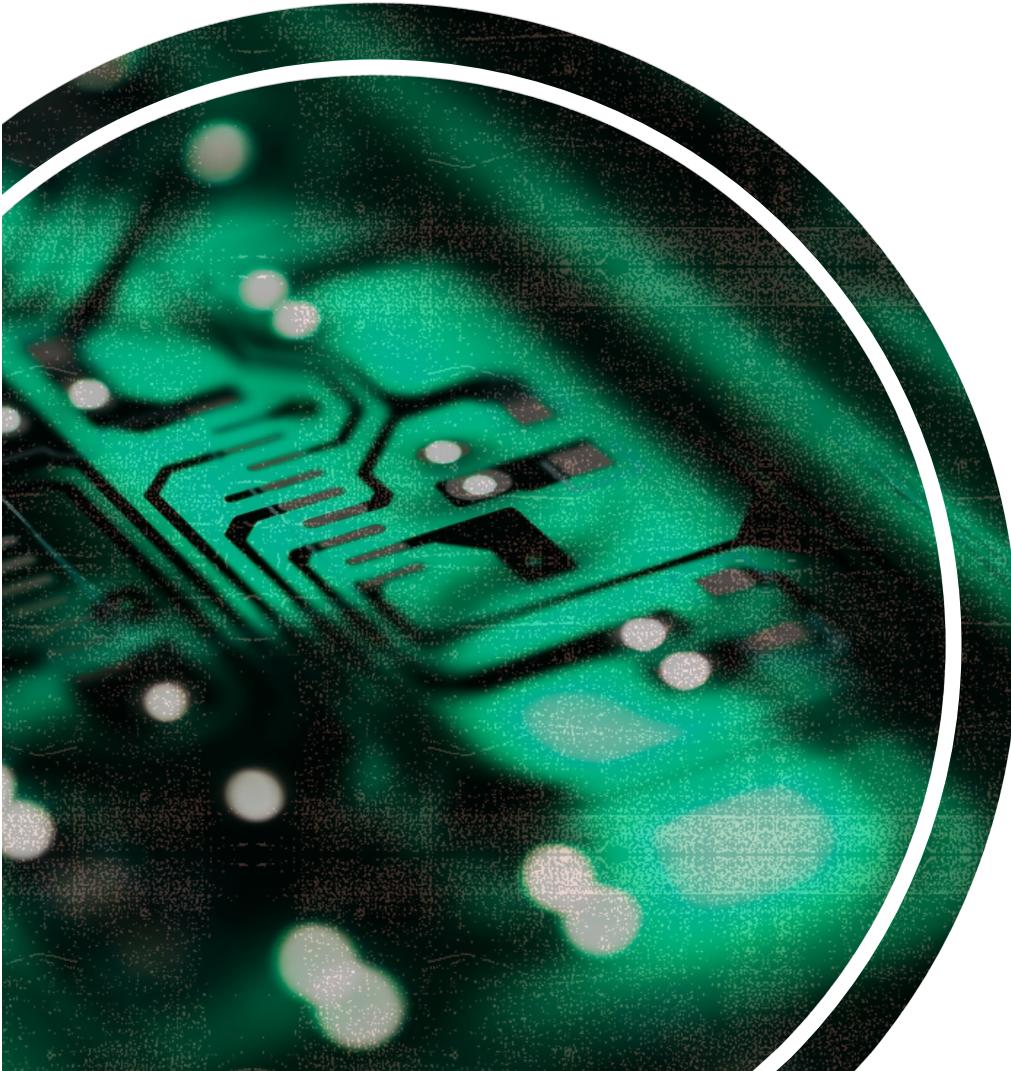




- **Example quantitative research questions:**

- How has the level of similarity between British English and American English changed over time?
- Has the use of gender-neutral pronouns in English language newspapers increased over the last decade?
- Are there differences in the words used by men and women in job applications?
- What are the predominant grammatical structures used in spoken versus written English?
- How does the use of passive voice in scientific writing compare across different disciplines?
- Do people use more markers of politeness ("please", "thank you") in the US South than in the North?





NATURAL LANGUAGE PROCESSING

- New field at the intersection of computer science, artificial intelligence, and linguistics
- One of the most important method of Computational Social Science
- Often doesn't fit into the classic quant-qual divide
- NLP can be used to 'predict', but can also be used as part of a quantitative or qualitative approach
- Predict, classify, cluster
- In-depth and large scale



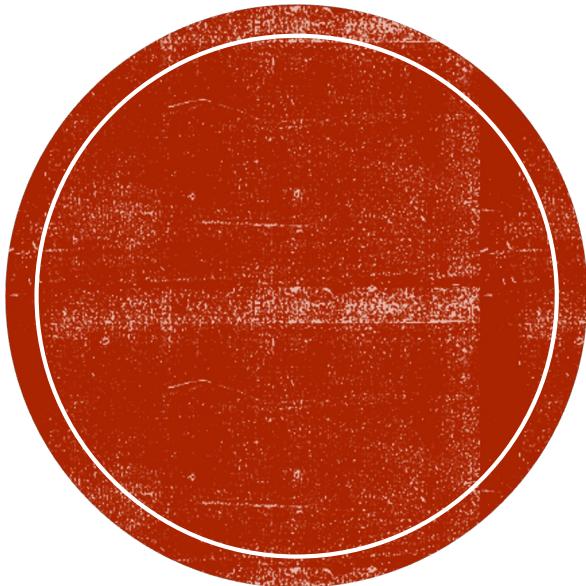


Examples of NLP methods

- **Sentiment analysis:** Identify the feelings expressed in a given text.
- **Topic modeling:** Unsupervised method to identify themes from textual data.
- **Machine Translation:** Translate from one language to another
- **Named Entity Recognition (NER):** Identify all the people and objects in a text.
- **Machine Learning:** Make predictions, automatically annotate data.

NLP often involves a combination of several techniques to carry out some specific task.





Examples research questions:

- Can we predict how stocks will go up or down on the stock market based on what Donald Trump writes on Twitter?
- Did Donald Trump's tweets become more emotionally negative over time?
- Can we predict who will win an election based on what people write on Twitter?
- Did Trump's Twitter messages result in a spread of misinformation among his followers?
- Does a high level of toxicity and incivility in the language of Trump's tweets result in more engagement from his followers?
- Did Trump focus on specific topics before elections, such as migration?

NLP can be combined with qual or quant approach!

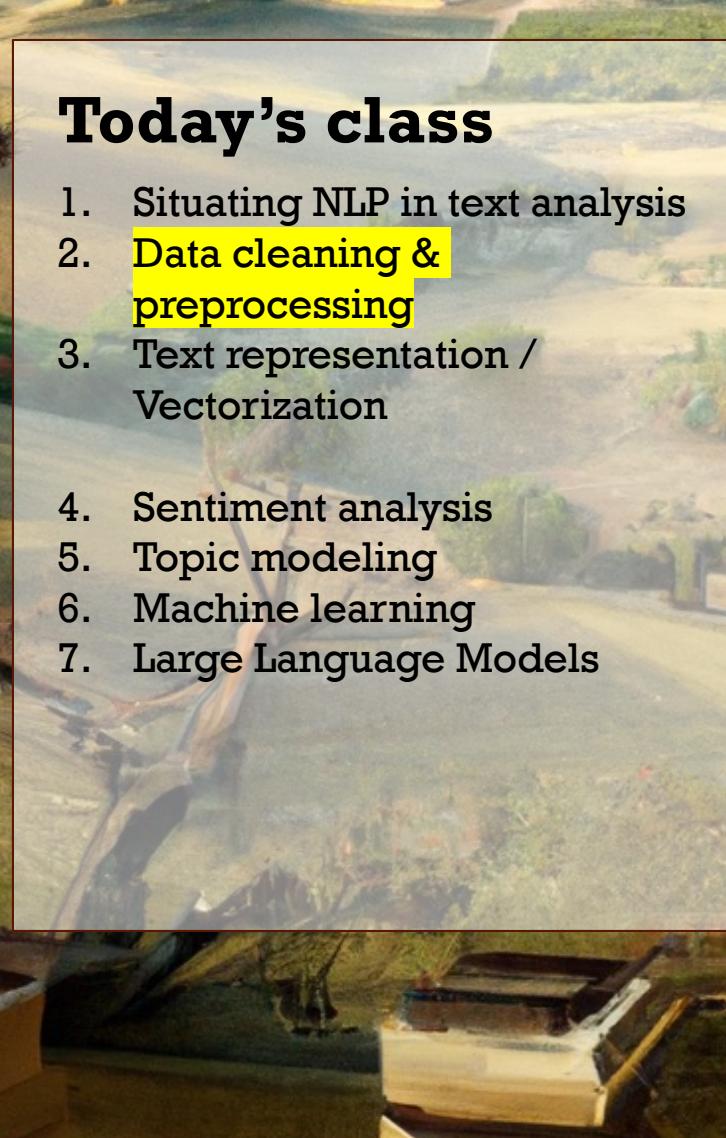
Discuss with your neighbor:

1. Did Trump focus on specific issues as a strategy for dealing with negative news during his presidency?
3. What fraction of Donald Trump's tweets contains attacks of political opponents?
3. How did Donald Trump construct his public identity through his Twitter messages?





Today's class

1. Situating NLP in text analysis
 2. Data cleaning & preprocessing
 3. Text representation / Vectorization
 4. Sentiment analysis
 5. Topic modeling
 6. Machine learning
 7. Large Language Models
- 



Step 0. Data cleaning

- Remove duplicates
- Remove rows with missing data
- Fill in missing data
- Remove outliers
- Harmonize data
- Fix data types



Step 1. Text cleaning: Remove HTML tags, URLs, numbers, etc.

- The first step is to clean the text. We remove any irrelevant items like HTML tags, URLs, and codes when dealing with web data. We also get rid of special characters, numbers, or punctuation that might not be necessary for analysis.
- We might want to drop rows with retweets, posts with URL, mentions, short messages!

“So now the Fake News @nytimes is tracing the CoronaVirus origins back to Europe, NOT China. This is a first!
#fakenews
I wonder what the Failing New York Times got for this one? Are there any NAMED sources? They were recently thrown out of China like dogs, and obviously want back in. Sad! <https://t.co/4qwCKQOjOw>”

```
import re

def clean_text(text):
    # Remove URLs
    text = re.sub(r'http\S+', '', text)
    # Remove hashtags (but keep the word following the hashtag)
    text = re.sub(r'#', '', text)
    # Remove special characters and numbers
    text = re.sub(r'[^A-Za-z\s]', '', text)
    # Remove additional white spaces
    text = re.sub(r'\s+', ' ', text).strip()
    return text
```



“So now the Fake News nytimes is tracing the CoronaVirus origins back to Europe NOT China This is a first fakenews I wonder what the Failing New York Times got for this one Are there any NAMED sources They were recently thrown out of China like dogs and obviously want back in Sad”



Step 2: Case normalization

Next, we normalize the case by converting all the text to lower case. This ensures that words like 'House', 'house', and 'HOUSE' are all treated as the same word, preventing the model from treating them as different entities.

```
def case_normalization(text):
    return text.lower()
```

“So now the Fake News nytimes is tracing the CoronaVirus origins back to Europe NOT China This is a first fakenews I wonder what the Failing New York Times got for this one Are there any NAMED sources They were recently thrown out of China like dogs and obviously want back in Sad”



“so now the fake news nytimes is tracing the coronavirus origins back to europe not china this is a first fakenews i wonder what the failing new york times got for this one are there any named sources they were recently thrown out of china like dogs and obviously want back in sad”



Step 3: Tokenization

Tokenization is where we break down the text into pieces - **t**okens.

Tokens can be words, phrases, or even sentences. In English, this might be as simple as splitting by spaces, but it can get complicated with languages that don't use spaces or have complex morphology.

Using **n-grams**, we can keep words like "United States" together: `United_States`.

"so now the fake news nytimes is tracing the coronavirus origins back to europe not china this is a first fakenews i wonder what the failing new york times got for this one are there any named sources they were recently thrown out of china like dogs and obviously want back in sad"



```
def tokenization(text):  
    return text.split()
```

['so', 'now', 'the', 'fake', 'news', 'nytimes', 'is', 'tracing', 'the', 'coronavirus', 'origins', 'back', 'to', 'europe', 'not', 'china', 'this', 'is', 'a', 'first', 'fakenews', 'i', 'wonder', 'what', 'the', 'failing', 'new', 'york', 'times', 'got', 'for', 'this', 'one', 'are', 'there', 'any', 'named', 'sources', 'they', 'were', 'recently', 'thrown', 'out', 'of', 'china', 'like', 'dogs', 'and', 'obviously', 'want', 'back', 'in', 'sad']

Step 4: Stop Words Removal

After tokenization, we often remove stop words.

These are common words like 'is', 'and', 'the', which appear frequently in the text but usually don't carry significant meaning for the analysis.

```
stopwordlist = set(['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
def remove_stopwords(tokens):
    return [token for token in tokens if token not in stopwordlist]
```

['so', 'now', 'the', 'fake', 'news', 'nytimes',
'is', 'tracing', 'the', 'coronavirus', 'origins',
'back', 'to', 'europe', 'not', 'china', 'this', 'is',
'a', 'first', 'fakenews', 'i', 'wonder', 'what',
'the', 'failing', 'new', 'york', 'times', 'got',
'for', 'this', 'one', 'are', 'there', 'any',
'named', 'sources', 'they', 'were',
'recently', 'thrown', 'out', 'of', 'china', 'like',
'dogs', 'and', 'obviously', 'want', 'back', 'in',
'sad']



['fake', 'news', 'nytimes', 'tracing', 'coronavirus',
'origins', 'back', 'europe', 'china', 'first',
'fakenews', 'wonder', 'failing', 'new', 'york',
'times', 'got', 'one', 'named', 'sources', 'recently',
'thrown', 'china', 'like', 'dogs', 'obviously', 'want',
'back', 'sad']



Step 5: Stemming or Lemmatization

Stemming and lemmatization are two forms of word normalization that are often performed in the fields of natural language processing and text mining.

Stemming typically involves removing the ends of words in the hope of achieving this goal more accurately and quickly, often with the aid of heuristics.

Lemmatization, uses vocabulary and morphological analysis to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the "lemma".

```
['fake', 'news', 'nytimes', 'tracing',
'coronavirus', 'origins', 'back', 'europe',
'china', 'first', 'fakenews', 'wonder',
'failing', 'new', 'york', 'times', 'got', 'one',
'named', 'sources', 'recently', 'thrown',
'china', 'like', 'dogs', 'obviously', 'want',
'back', 'sad']
```



```
import nltk
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize

# Initialize stemmer and lemmatizer
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

# Perform stemming
stemmed_words = [stemmer.stem(token) for token in tokens]

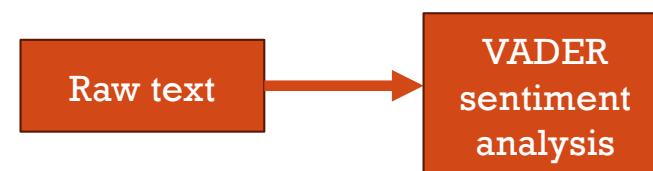
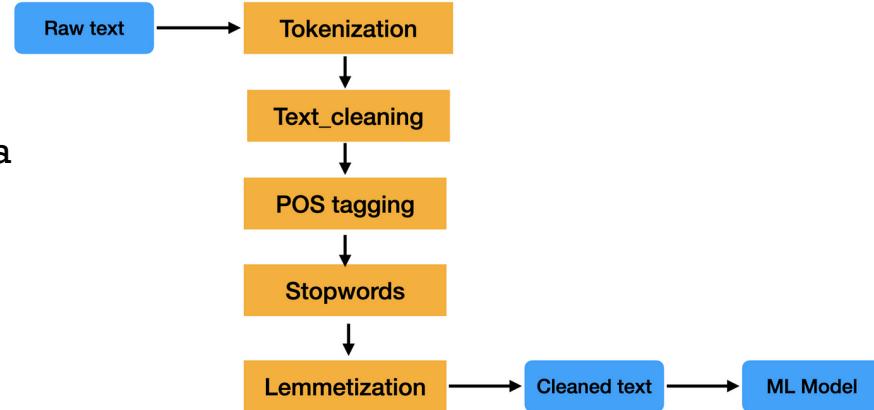
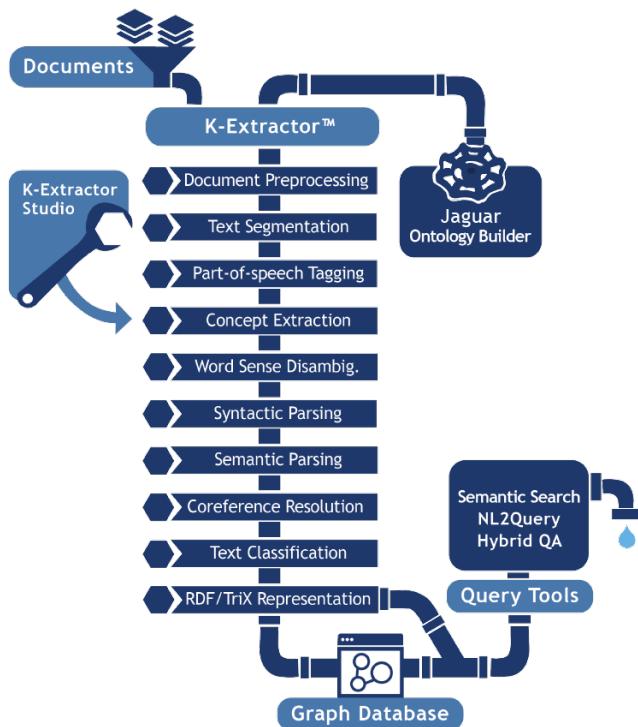
# Perform lemmatization
lemmatized_words = [lemmatizer.lemmatize(token) for token in tokens]
```

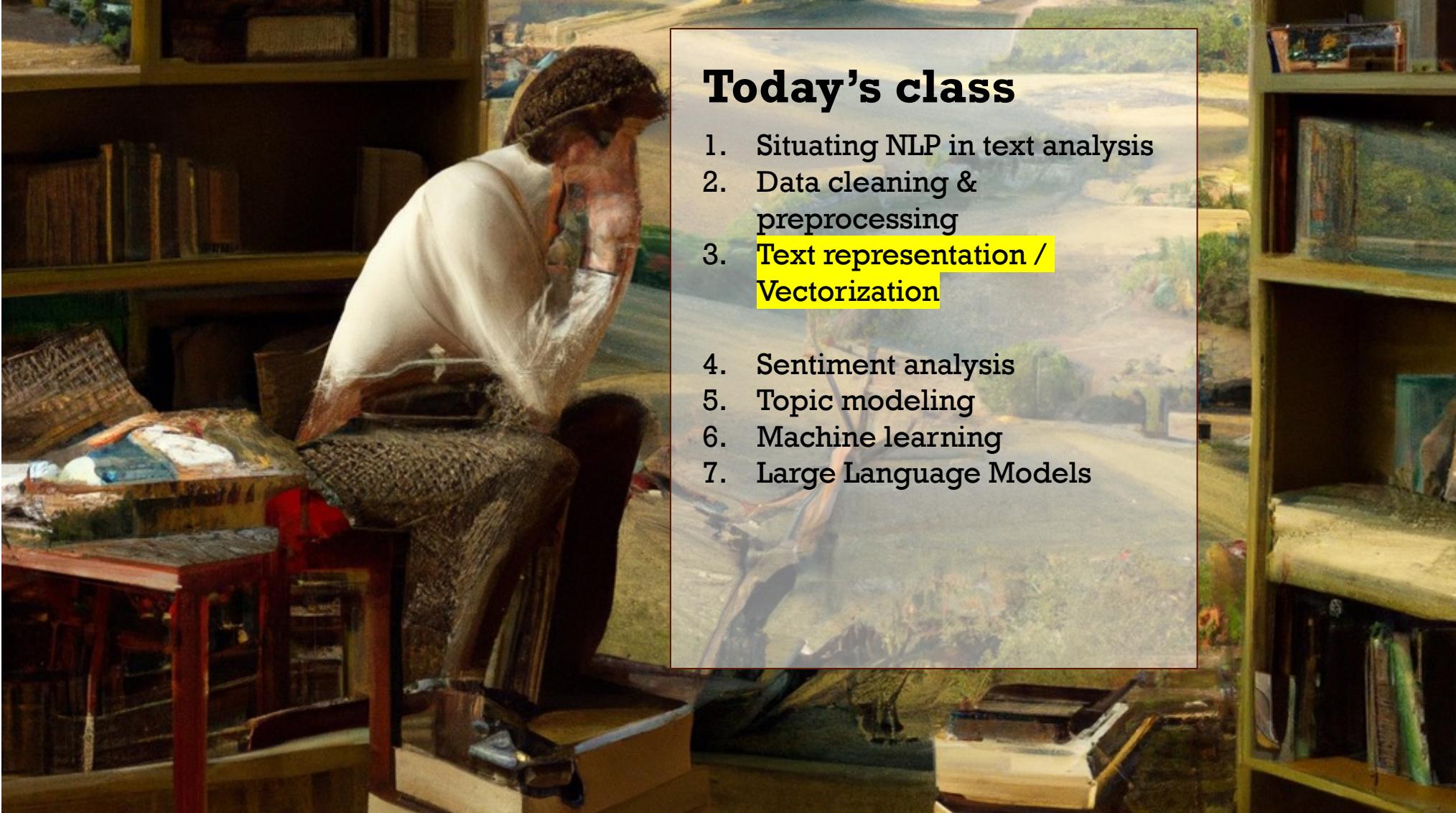
Stemmed : ['fake', 'news', 'nytim', 'trace',
'coronaviru', 'origin', 'back', 'europ', 'china',
'first', 'fakenew', 'wonder', 'fail', 'new', 'york',
'time', 'got', 'one', 'name', 'sourc', 'recent',
'thrown', 'china', 'like', 'dog', 'obvious', 'want',
'back', 'sad']

Lemmatized: ['fake', 'news', 'nytimes', 'tracing',
'coronavirus', 'origin', 'back', 'europe', 'china',
'first', 'fakenews', 'wonder', 'failing', 'new', 'york',
'time', 'got', 'one', 'named', 'source', 'recently',
'thrown', 'china', 'like', 'dog', 'obviously', 'want',
'back', 'sad']

NLP pipelines

A data analysis pipeline refers to a series of systematically arranged processes or steps that data goes through from collection to the generation of insights or models.





Today's class

1. Situating NLP in text analysis
2. Data cleaning & preprocessing
3. Text representation / Vectorization
4. Sentiment analysis
5. Topic modeling
6. Machine learning
7. Large Language Models

Text representation / Vectorization

- Transforms text into a format that machines can understand and process: numbers!

Common methods include:

1. Bag-of-words
2. TF-IDF
3. Word embeddings



1. Bag-of-words

- Just count the occurrence of each word.
- The order of words is ignored.
- Simplicity and ease of understanding.
- E.g., word frequency analysis

```
from collections import Counter
def bag_of_words(tokens):
    return Counter(tokens)

bag_of_words(tokens)
```

['fake', 'news', 'nytimes', 'tracing', 'coronavirus',
'origin', 'back', 'europe', 'china', 'first',
'fakenews', 'wonder', 'failing', 'new', 'york',
'time', 'got', 'one', 'named', 'source', 'recently',
'thrown', 'china', 'like', 'dog', 'obviously', 'want',
'back', 'sad']



{"back": 2, "china": 2, "fake": 1, "news": 1, "nytimes": 1,
"tracing": 1, "coronavirus": 1, "origin": 1, "europe": 1,
"first": 1, "fakenews": 1, "wonder": 1, "failing": 1,
"new": 1, "york": 1, "time": 1, "got": 1, "one": 1,
"named": 1, "source": 1, "recently": 1, "thrown": 1,
"like": 1, "dog": 1, "obviously": 1, "want": 1, "sad": 1}



2. TF-IDF (Term Frequency Inverse Document Frequency)

- TF-IDF takes into account how unique the words are to the documents.
- The **TF** (Term Frequency) part measures how often a term occurs in a document.
- The **IDF** (Inverse Document Frequency): what fraction of documents does the word occur in?
- Multiplying TF by IDF results in the TF-IDF score of a word for a particular document, with higher scores indicating greater importance.

1. "The quick brown fox jumps over the lazy dog",
2. "Never jump over the lazy dog quickly",
3. "Bright vixens jump; dozy fowl quack",
4. "Quick wafting zephyrs vex bold Jim"



```
[{'the': 1.3862943611198906,  
 'quick': 0.6931471805599453,  
 'brown': 1.3862943611198906,  
 'fox': 1.3862943611198906,  
 'jumps': 1.3862943611198906,  
 'over': 0.6931471805599453,  
 'lazy': 0.6931471805599453,  
 'dog': 0.6931471805599453},  
 {'never': 1.3862943611198906,  
 'jump': 1.3862943611198906,  
 'over': 0.6931471805599453,  
 'the': 0.6931471805599453,  
 'lazy': 0.6931471805599453},  
 ....]
```



2. TF-IDF of Trump; month-by-month

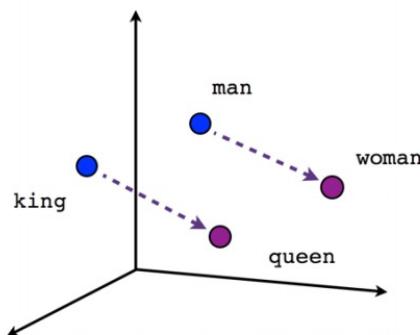
- While simple, TF-IDF is often enough to give an overview of a set of documents.
- What words are most typical of a certain document?
- For instance, what Trump talked about month by month:

```
2009 ['keychain', 'commercialfree', 'bookkeychain', 'bahamas', 'hysterical', 'jareds', 'bountiful', 'habitat']
2010 ['wahlberg', 'internationals', 'casting', 'soho', 'bret', 'telethon', 'contest', 'wwwyoutubeusermattresserta', 'tune', 'episode', 'episode', 'farewell']
2011 ['cavuto', 'desk', 'trumproast', 'danmcisaac', 'vegasusa', 'campanella', 'barackobama', 'barackobama', 'barackobama', 'barackobama', 'timetogettough']
2012 ['barackobama', 'barackobama', 'barackobama', 'sweepstweet', 'barackobama', 'barackobama', 'barackobama', 'embassies', 'mittromney', 'sandy', 'sugar']
2013 ['billmaher', 'thanks', 'celebapprentice', 'celebapprentice', 'dannyzuker', 'snowden', 'realdonaldtrump', 'realdonaldtrump', 'syria', 'website', 'website', 'mandela']
2014 ['saintanselm', 'buzzfeed', 'oscars', 'afphq', 'mayweather', 'pinehurst', 'vera', 'ebola', 'joan', 'ebola', 'ferguson', 'sony']
2015 ['via', 'entrepreneurs', 'entrepreneurs', 'jonahnro', 'entrepreneurs', 'ncgopcon', 'via', 'jeb', 'asktrump', 'carson', 'crippled', 'jebbush']
2016 ['cruz', 'cruz', 'votetrump', 'nyprimary', 'goofy', 'hillary', 'rncircle', 'imwithyou', 'imwithyou', 'bigleaguetruth', 'draintheswamp', 'thankyoutour']
2017 ['hacking', 'fake', 'caucus', 'ossoff', 'fake', 'kateslaw', 'uswomensopen', 'hurricaneharvey', 'luther', 'corker', 'apec', 'pensacola']
2018 ['daca', 'daca', 'agony', 'memos', 'hunt', 'barriers', 'nato', 'strzok', 'florence', 'jobsnotmobs', 'broward', 'border']
2019 ['border', 'kim', 'collusion', 'mueller', 'mueller', 'tariffs', 'elijah', 'dayton', 'fake', 'schiff', 'impeachment', 'impeachment']
2020 ['impeachment', 'mini', 'coronavirus', 'ventilators', 'transit', 'usdot', 'biden', 'portland', 'biden', 'biden', 'biden', 'briankempga']
2021 ['stopthesteal']
```

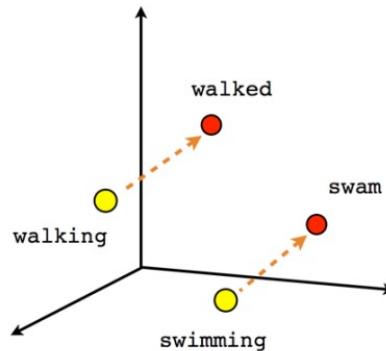


3. Word embeddings: Math with words

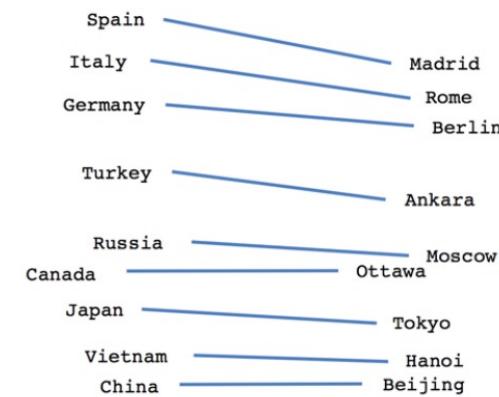
- More advanced technique. Words are mapped to vectors in a high-dimensional space so similar-used words are close to each other.
- Synonyms are near each other.
- 'king' and 'queen' will be closer in the vector space than 'king' and 'apple'.
- Allows doing math with language! $\text{King} + \text{woman} = \text{queen}$
- Capture the semantic meaning. Excellent for tasks requiring a deep understanding of language, like sentiment analysis or machine translation. The downside is that they require a lot of data to train.
- Techniques include **Word2Vec** or **GloVe**.



Male-Female

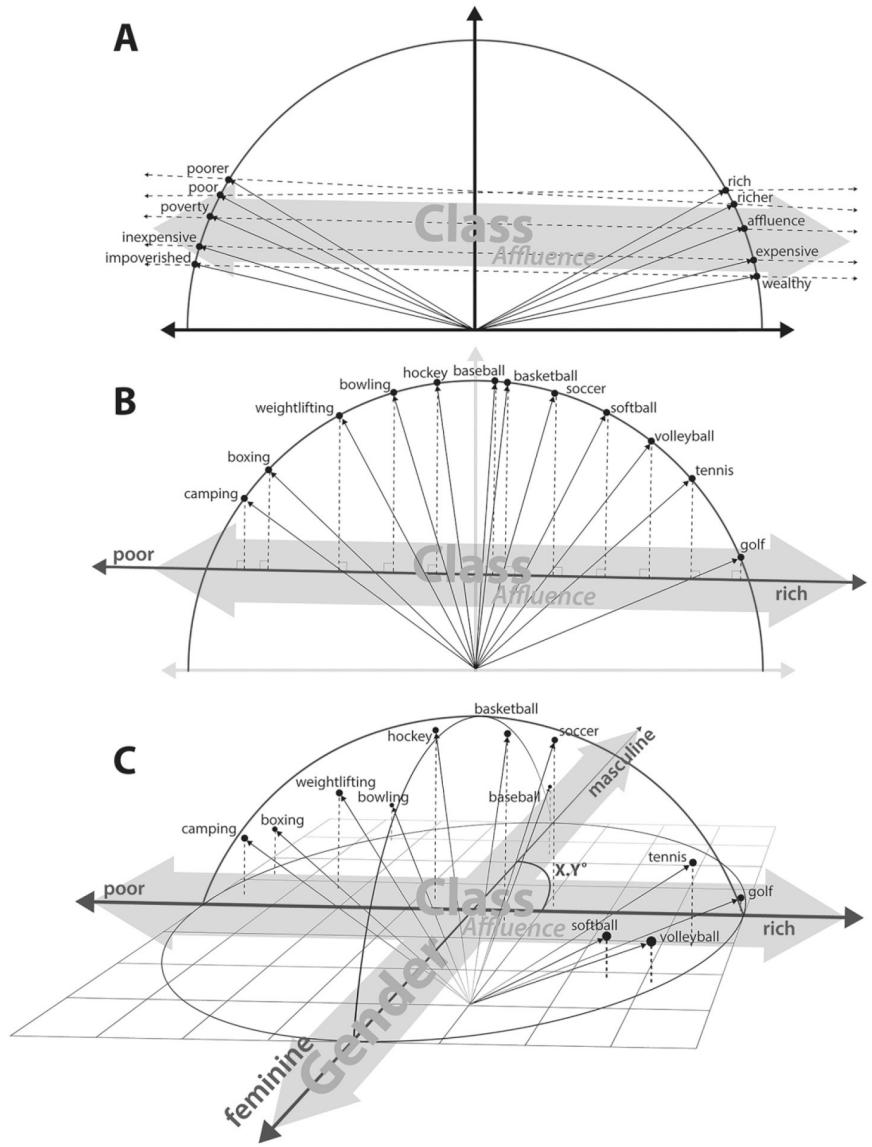


Verb tense

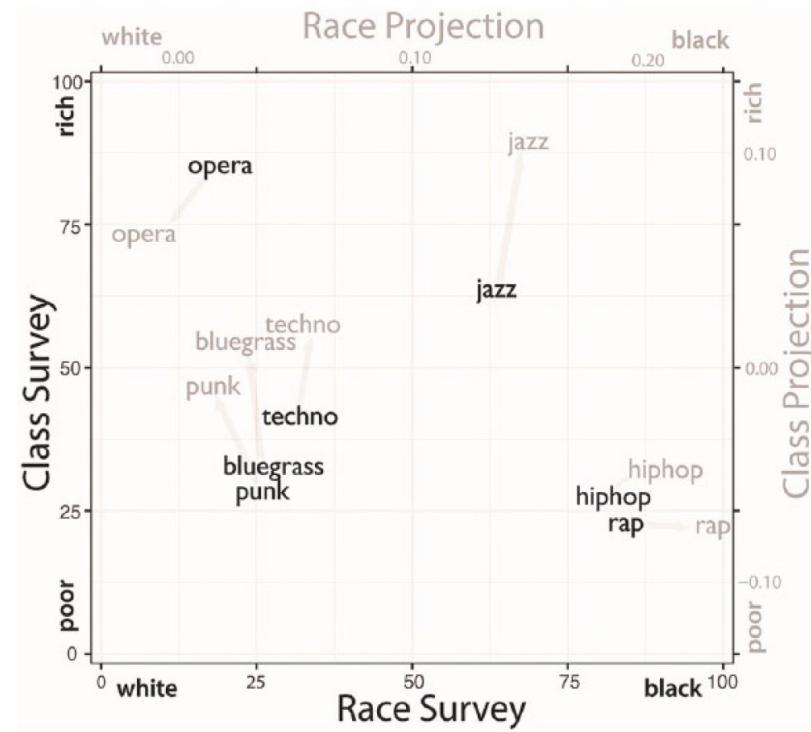


Country-Capital



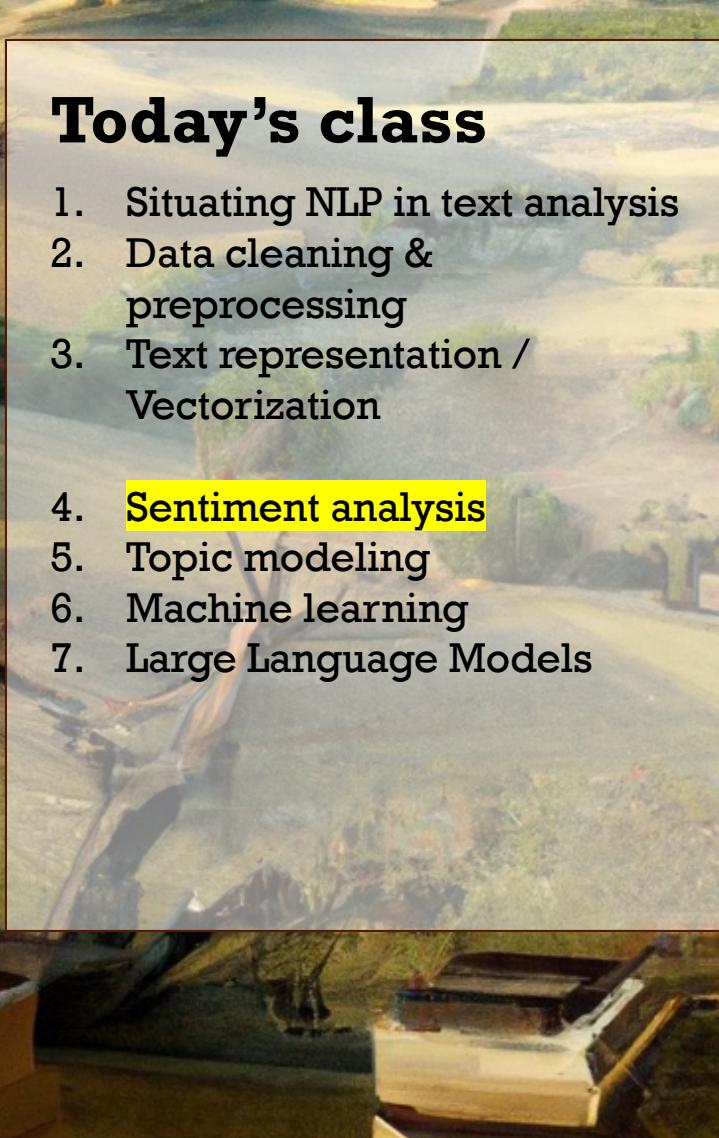


Can be used to capture stereotypes and to measure cultural representations!





Today's class

1. Situating NLP in text analysis
 2. Data cleaning & preprocessing
 3. Text representation / Vectorization
 4. **Sentiment analysis**
 5. Topic modeling
 6. Machine learning
 7. Large Language Models
- 



Sentiment analysis

Find emotional tone, aimed at understanding the attitudes, opinions, and emotions expressed within an online mention or text.

Approaches for sentiment analysis:

- Lexicon-based
- Rule-based
- Machine learning-based
- Deep learning

Sentiment analysis tools and libraries:

Vader, nltk, textblob, Stanford, Hugging Face transformers.

Fine-tuning pre-trained models on specific datasets

Rule-based approach

Includes grammar rules and heuristics.

E.g., VADER (Valence Aware Dictionary and sEntiment Reasoner)

- **Intensity modifiers:** These amplify or reduce the sentiment score (e.g., "extremely good" vs. "good").
- **Conjunctions:** VADER adjusts the sentiment score based on conjunctions like "but" which can change the sentiment of a sentence (e.g., "The product is great, but the service is terrible").
- **Punctuation:** The presence of an exclamation mark can increase the sentiment intensity.
- **Capitalization:** If a word is in all caps, it may indicate increased sentiment intensity (e.g., "GREAT" vs. "great").
- **Emojis and Emoticons:** 😊
- **Negations:** VADER effectively handles negations, flipping the sentiment of the words that follow (e.g., "not good").

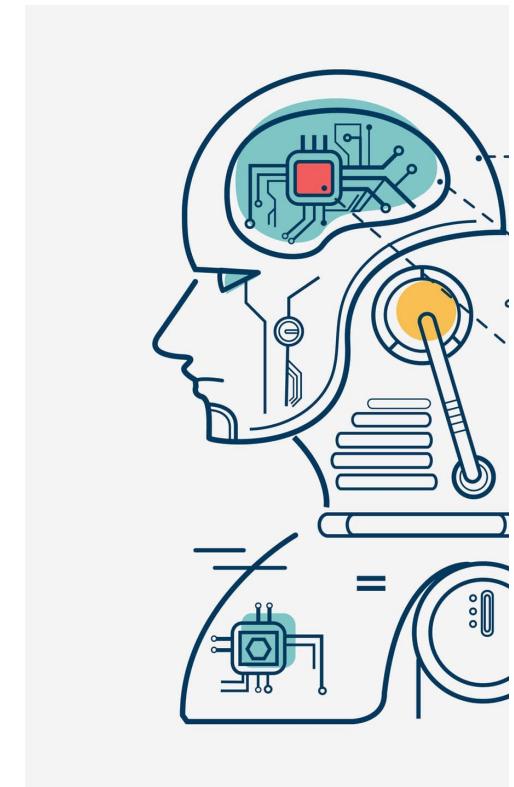
VADER works pretty well... at least for being so simple.



Machine Learning Approach

The machine learning approach to sentiment analysis involves training classifiers using labeled data.

- **Upsides:**
 - Accurate and adapted to specific corpus
 - Fast and scalable
- **Downsides:**
 - You need high-quality training data
 - Some models can be quite complex and may require extensive computational resources.



How do you choose?

- Overall, the choice between these approaches often comes down to a **trade-off between accuracy, interpretability, and computational efficiency**.
- Lexicon-based methods are fast and interpretable but less accurate.
- Machine learning strikes a balance but requires careful feature engineering and sufficient labeled data.
- Deep learning offers high accuracy and can handle a wide range of complex linguistic features but at the cost of computational resources and often interpretability.





```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

#Create an instance of the VADER sentiment analyzer
analyzer = SentimentIntensityAnalyzer()

#Define a function to perform sentiment analysis using VADER
def get_sentiment(tokens):
    sentiment = analyzer.polarity_scores(tokens)
    compound_score = sentiment['compound']
    return compound_score

# Apply the function to the 'tokens' column of the DataFrame
df['sentiment'] = df['text'].apply(get_sentiment)
```

Most Negative

[According to press reports, Iran may be planning an assassination, or other attack, against the United States in retaliation for the killing of terrorist leader Soleimani, which was carried out for his planning a future attack, murdering U.S. Troops, and the death & suffering...],

'It is outrageous that Poisonous Synthetic Heroin Fentanyl comes pouring into the U.S. Postal System from China. We can, and must, END THIS NOW! The Senate should pass the STOP ACT – and firmly STOP this poison from killing our children and destroying our country. No more delay!',

'Our prayers are with the people of Vienna after yet another vile act of terrorism in Europe. These evil attacks against innocent people must stop. The U.S. stands with Austria, France, and all of Europe in the fight against terrorists, including radical Islamic terrorists.]

Most positive

[Congratulations to Patrick Reed on his great and courageous MASTERS win! When Patrick had his amazing win at Doral 5 years ago, people saw his great talent, and a bright future ahead. Now he is the Masters Champion!]',

'Today, it was my great honor to proudly accept the endorsement of the @NYCPBA! I have deeply and profoundly admired the brave men and women of the #NYPD for my entire life. New York's Finest are truly the best of the best — I will NEVER let you down! #MAGA <https://t.co/o115kAx34v>',

'I am a big fan of Drew Brees. I think he's truly one of the greatest quarterbacks, but he should not have taken back his original stance on honoring our magnificent American Flag. OLD GLORY is to be revered, cherished, and flown high...']

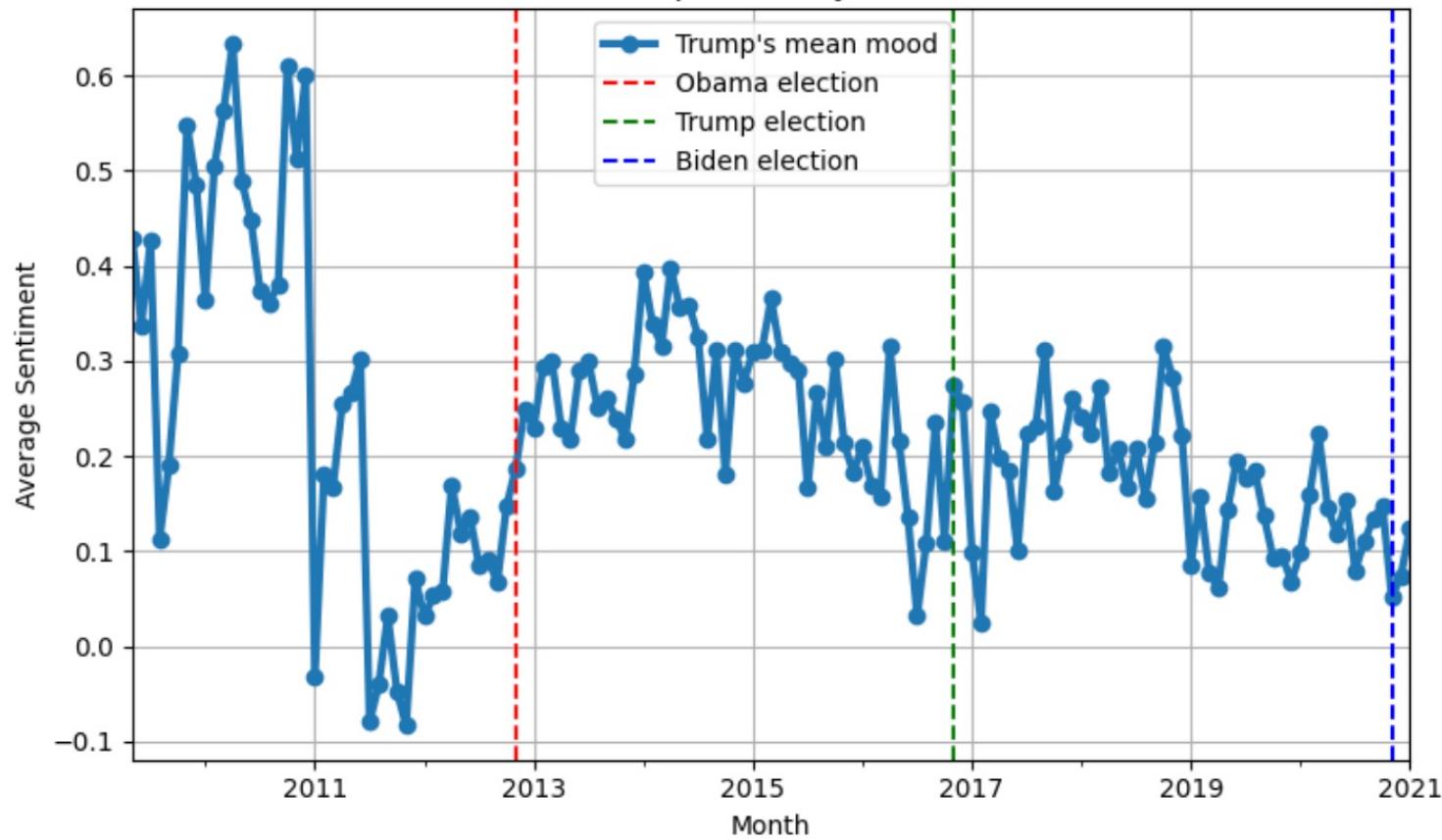


BE CONSCIOUS OF WHAT YOU ACTUALLY MEASURE

- Vader captures general sentiment of the phrase
- Often, you may want to capture the person's view *on something*. Or what feeling they are expressing.
- Make sure you're measuring what you intend!



Trump's Monthly Mood

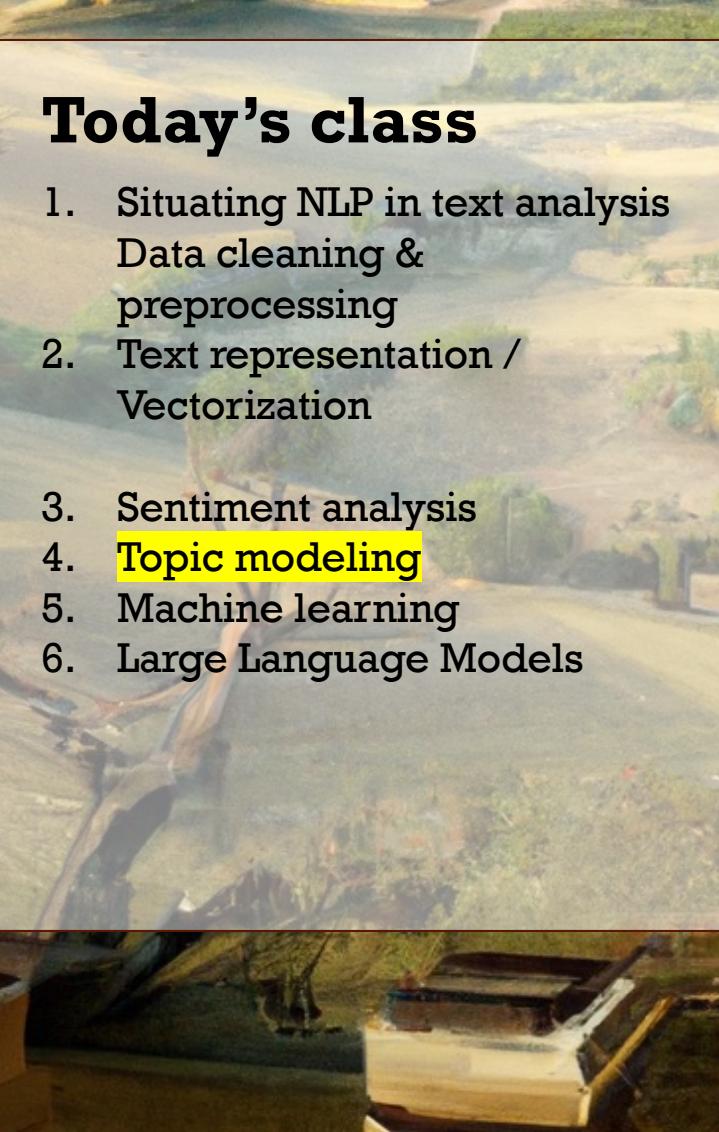


Can you think of a research question that you could use sentiment analysis to answer?





Today's class

1. Situating NLP in text analysis
Data cleaning &
preprocessing
 2. Text representation /
Vectorization
 3. Sentiment analysis
 4. Topic modeling
 5. Machine learning
 6. Large Language Models
- 



TOPIC MODELING

- By analyzing the co-occurrence patterns of words in documents, Topic Modeling identifies themes or topics in large number of documents.
- Useful for summary and overview, and for discourse analysis
- You can use it to feed a quantitative model.
- Or you can combine it with a qualitative approach, such as Critical Discourse Analysis

Törnberg, A. Törnberg, P. 2016. "Combining CDA and Topic Modeling"



TYPES OF TOPIC MODELING

Latent Dirichlet Allocation (LDA): *Oldie but goldie.* Assumes documents are a mixture of topics and that each word in the document is attributable to the document's topics.

Correlated Topic Model (CTM): An extension of LDA that allows topics to be correlated.

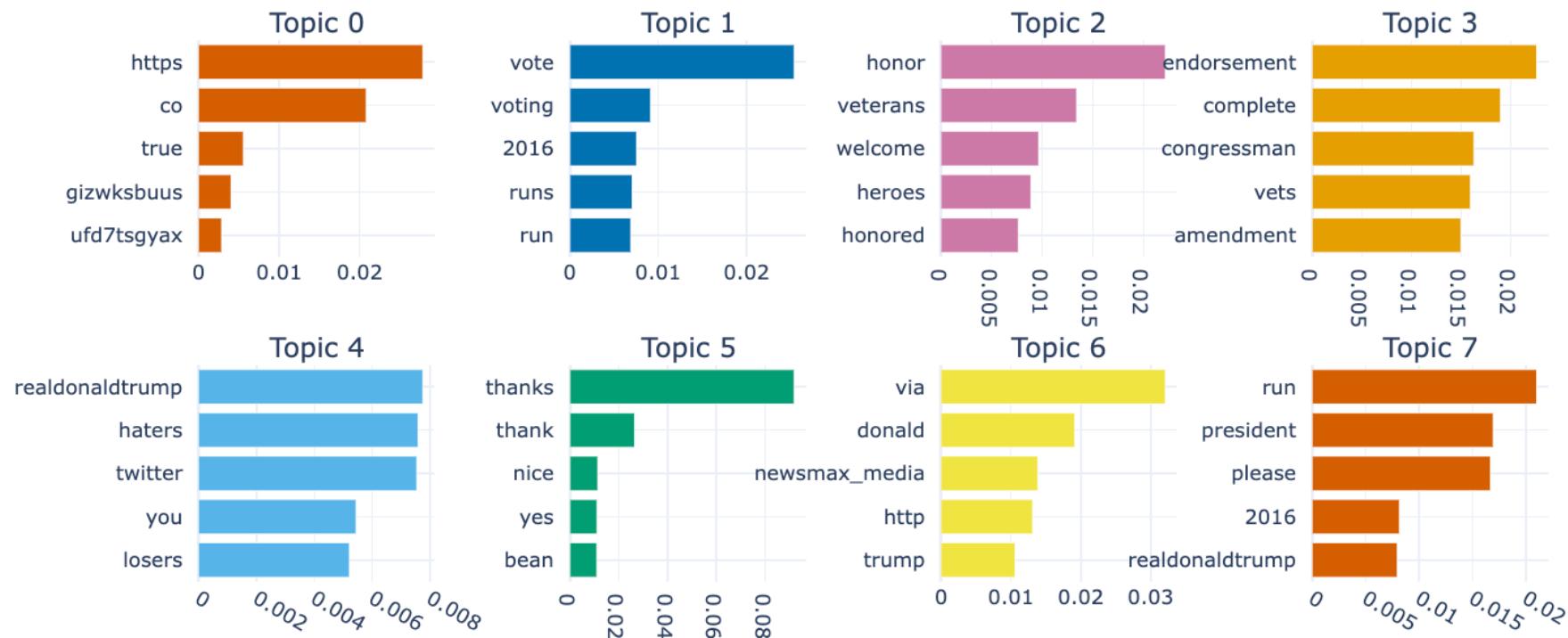
Dynamic Topic Model (DTM): Captures the evolution of topics over time and is useful for analyzing temporal trends in a corpus.

BERTopic uses the semantical understanding of words. So “man” and “boy” are treated as more similar than “man” and “apple.”

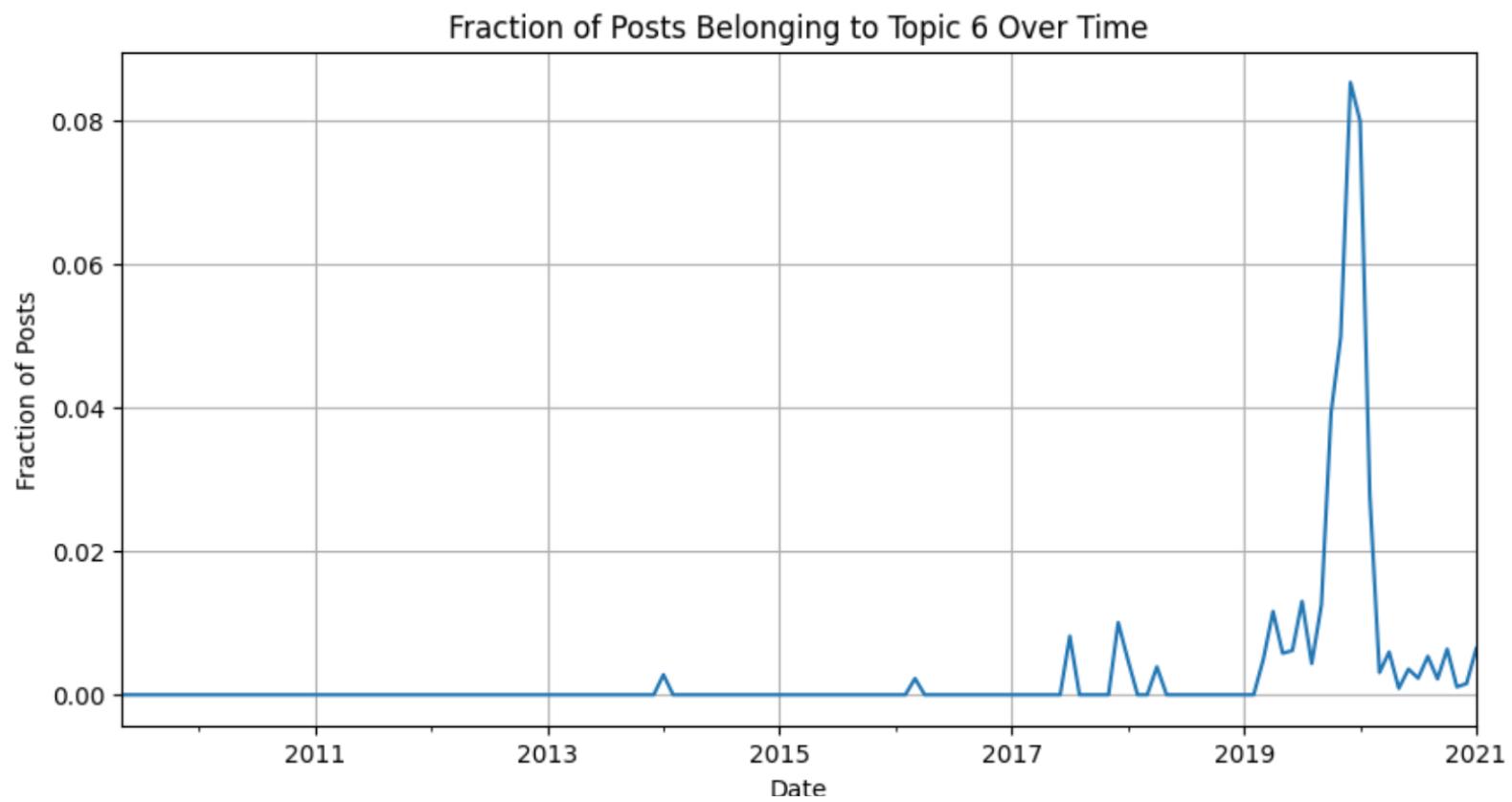


Topic	Count	Name	Representation	Representative_Docs
0	-1	25457 -1_trump_rt_realdonaldtrump_the	[trump, rt, realdonaldtrump, the, president, t...]	[Andrew McCarthy - "I said this could never ha...
1	0	1138 0_https_co_true_gizwksbuus	[https, co, true, gizwksbuus, ufd7tsgyax, jxzm...]	[I am not a fan of Bitcoin and other Cryptocur...
2	1	626 1_vote_voting_2016_runs	[vote, voting, 2016, runs, run, my, ran, reald...	["""@OGRumple: @realDonaldTrump has my first p...
3	2	564 2_honor_veterans_welcome_heroes	[honor, veterans, welcome, heroes, honored, to...	[It was my great honor to host a @WhiteHouse C...
4	3	533 3_endorsement_complete_congressman_vets	[endorsement, complete, congressman, vets, ame...	[Congressman @MichaelCBurgess has done a Great...
5	4	522 4_realdonaldtrump_hatters_twitter_you	[realdonaldtrump, hatters, twitter, you, losers...	["""@utkarsh_chheda: @realDonaldTrump @Elgiann...
6	5	501 5_thanks_thank_nice_yes	[thanks, thank, nice, yes, bean, hellcat1963, ...]	[@ErolYalim Thanks and yes., @cloverc8 Thank...
7	6	457 6_via_donald_newsmax_media_http	[via, donald, newsmax_media, http, trump, cpac...	[Via @Newsmax_Media: "Trump to Speak at CPAC" ...
8	7	436 7_run_president_please_2016	[run, president, please, 2016, realdonaldtrump...	["""@RowleyRhonda: I feel our country is falli...
9	8	342 8_china_chinese_currency_tariffs	[china, chinese, currency, tariffs, trade, pro...	[..There will be nobody left in China to do bu...
10	9	323 9_barackobama_debt_budget_cont	[barackobama, debt, budget, cont, solyndra, de...	[@BarackObama trillion \$ stimulus was supposed...
11	10	311 10_crowd_rally_carolina_iowa	[crowd, rally, carolina, iowa, heading, hampsh...	[Iowa was amazing today. Great crowd, great pe...
12	11	256 11_rt_stoolpresidente_https_realdonaldtrump	[rt, stoolpresidente, https, realdonaldtrump, ...]	[RT @stoolpresidente: (PART 3) - @realDonaldTrump...
13	12	254 12_iran_nuclear_iranian_sanctions	[iran, nuclear, iranian, sanctions, deal, iran...	[Iran's quest for nuclear weapons is a major t...
14	13	250 13_makeamericagreatagain_trump2016_trump2016ht...	[makeamericagreatagain, trump2016, trump2016ht...	[RT @TimMurtaugh: You got it.\n\n#MakeAmericaG...
15	14	248 14_thank_https_you_co	[thank, https, you, co, ari, working, z8dgzqxk...	[Thank you Ari! https://t.co/ghPg3w2jmc, Thank...
16	15	246 15_obamacare_repeal_website_premiums	[obamacare, repeal, website, premiums, healthc...	[It's Monday. How much will premiums rise toda...
17	16	243 16_nancy_pelosi_she_schumer	[nancy, pelosi, she, schumer, chuck, cryin, ne...	[Nancy Pelosi will go down as the absolute wor...
18	17	233 17_fake_media_news_dishonest	[fake, media, news, dishonest, enemy, sources,...]	[FIGHT THE CORRUPT FAKE NEWS MEDIA. VOTE!, So ...
19	18	198 18_coronavirus_task_briefing_force	[coronavirus, task, briefing, force, press, re...	[RT @WhiteHouse: LIVE: Press Briefing with Cor...

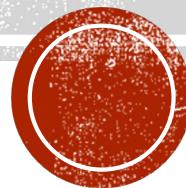
Topic Word Scores

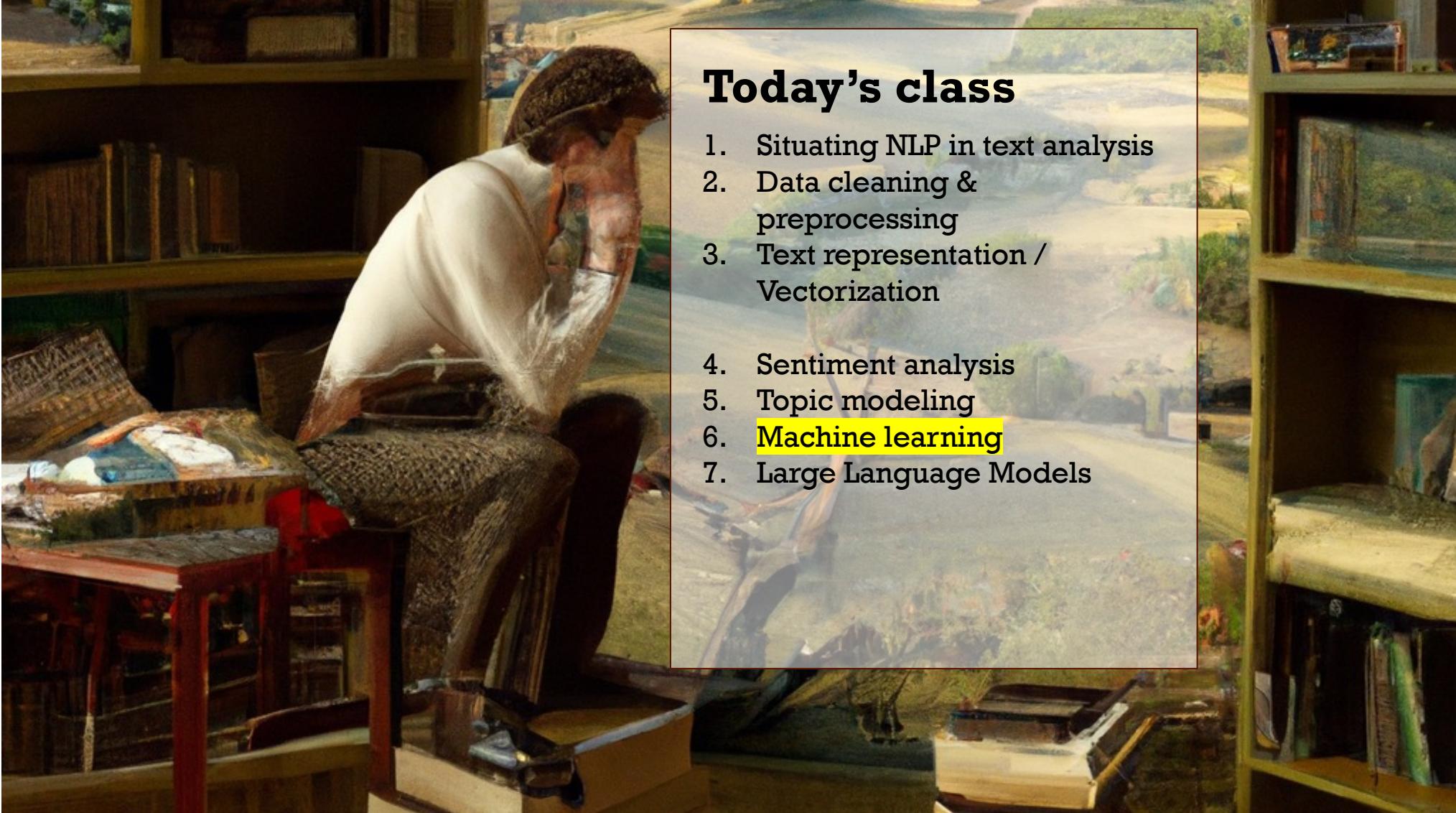


IMPEACHMENT TOPIC OVER TIME



Can you think of a research question that you could use topic modeling to answer?





Today's class

1. Situating NLP in text analysis
2. Data cleaning & preprocessing
3. Text representation / Vectorization
4. Sentiment analysis
5. Topic modeling
6. Machine learning
7. Large Language Models

Machine Learning

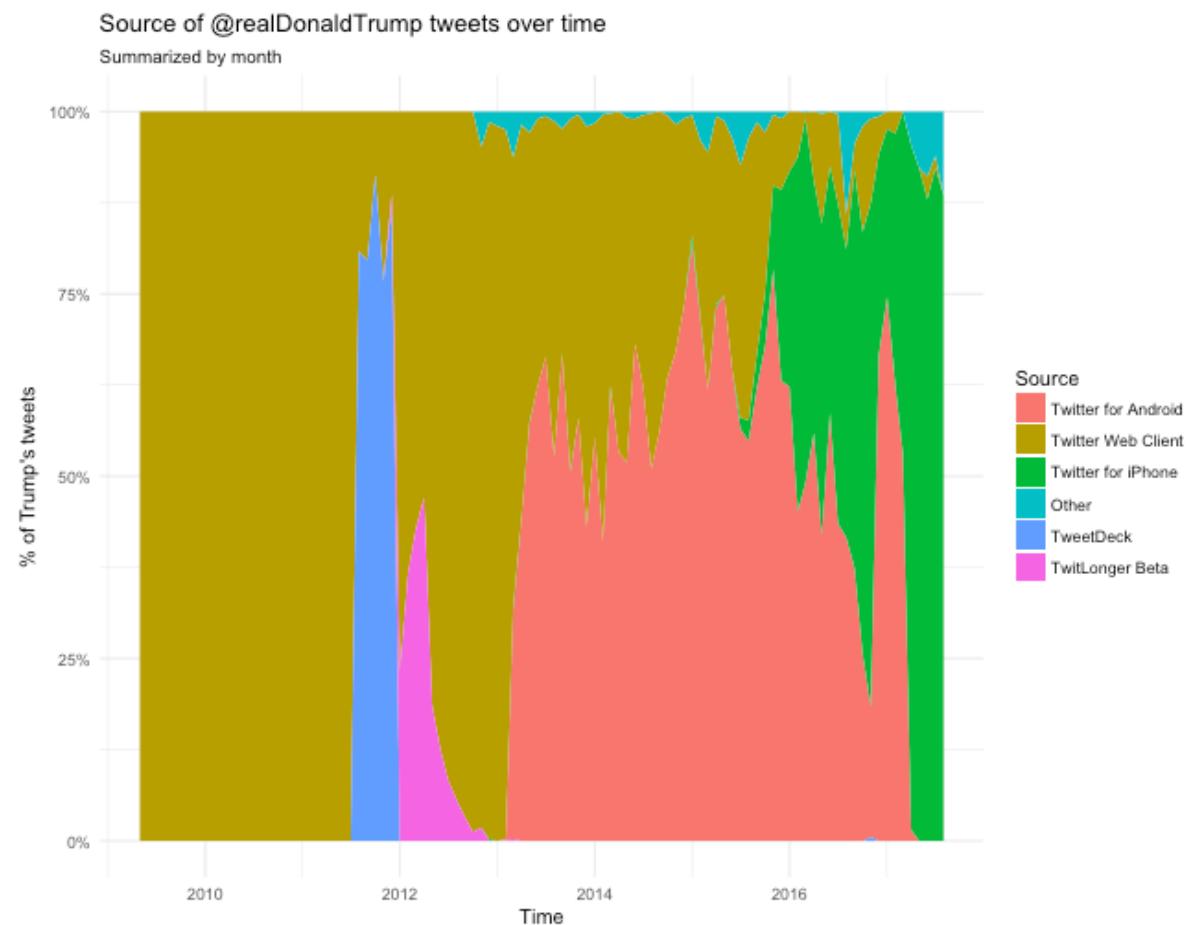
- Machine learning is highly versatile: can do a whole range of prediction, and annotation tasks.
- You usually manually classify a lot of tweets, and then train the ML on this data.
- By learning from examples, these models can process and analyze large volumes of text data to support decision-making, generate recommendations, or understand human language nuances.

For example:

- How toxic is the language?
- Is the tweet focused on electoral politics?



Example: Donald Trump's devices



	precision	recall	f1-score	support
Twitter for Android	0.78	0.37	0.50	1411
Twitter for iPhone	0.85	0.97	0.91	5373
accuracy			0.85	6784
macro avg	0.82	0.67	0.71	6784
weighted avg	0.84	0.85	0.82	6784

We can guess surprisingly well
from which device he sent the
message



Most important words for the two groups

Twitter for iPhone

Word: obama, Importance: 0.007366149190531945
Word: hillary, Importance: 0.007255819145786383
Word: amp, Importance: 0.005783251336030852
Word: thank, Importance: 0.004910221418174412
Word: great, Importance: 0.004896118493564506
Word: democrats, Importance: 0.004808223974237916
Word: fake, Importance: 0.003668706258841094
Word: amazing, Importance: 0.003632951769746381
Word: interviewed, Importance: 0.0032869110471028164
Word: news, Importance: 0.003257693138817188
Word: people, Importance: 0.002967642742242895
Word: just, Importance: 0.0028917800312117724
Word: crooked, Importance: 0.0027413681526823846
Word: trump, Importance: 0.0026793189318149173
Word: fox, Importance: 0.002650839703158417
Word: enjoy, Importance: 0.002618503988235729
Word: big, Importance: 0.0026130824224925857
Word: make, Importance: 0.0023194059991433493
Word: really, Importance: 0.0022721572470865226
Word: cnn, Importance: 0.0022697066017065763
Word: tonight, Importance: 0.0022639762109963745
Word: foxnews, Importance: 0.002197461214519149
Word: new, Importance: 0.002082472837855381

Twitter for Android

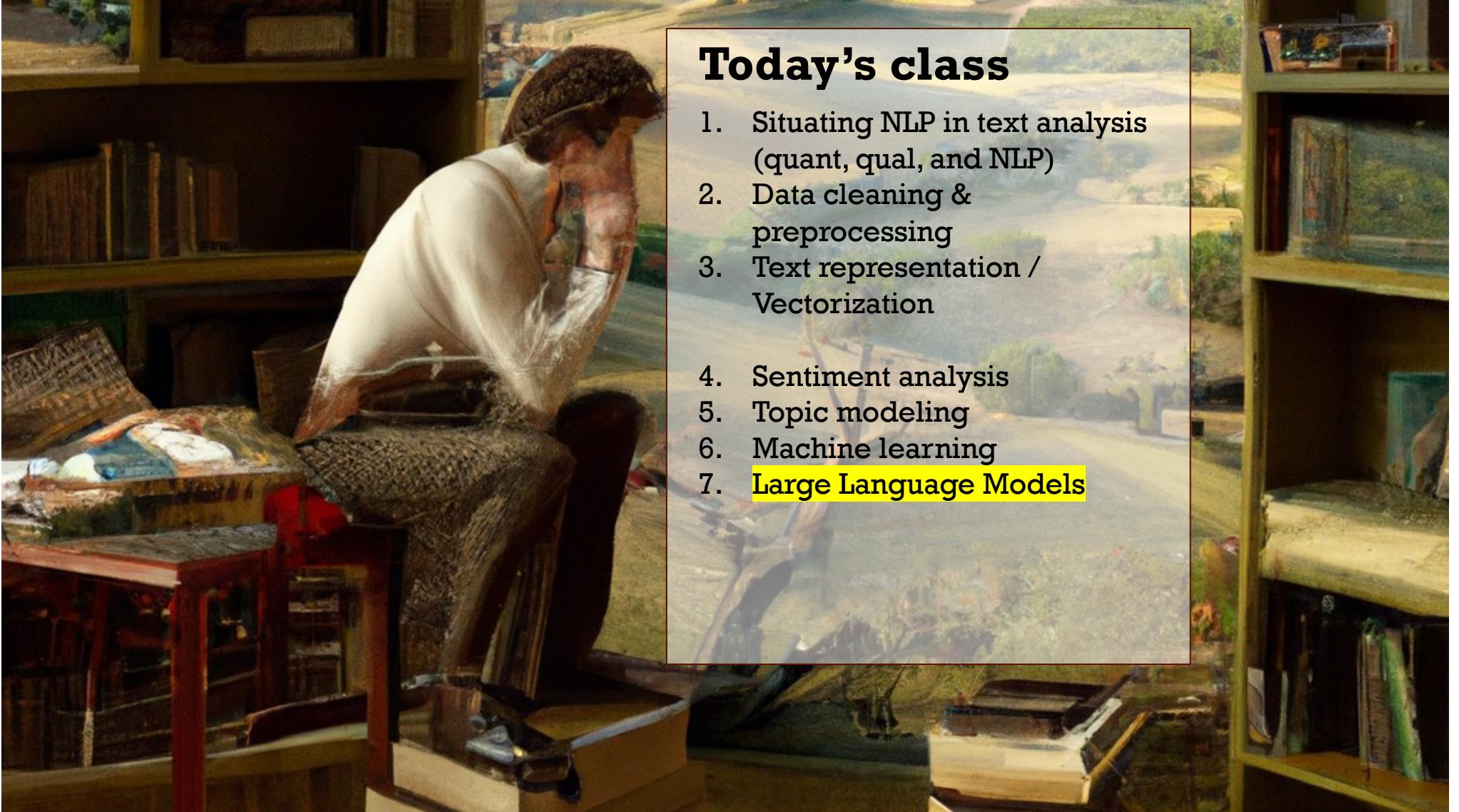
Word: thanks, Importance: 0.022453657168298197
Word: apprentice, Importance: 0.005399930970775972
Word: cruz, Importance: 0.004122119780227118
Word: dumb, Importance: 0.00375014445206056
Word: stupid, Importance: 0.0037256964287739123
Word: ebola, Importance: 0.0032234555243585887
Word: warming, Importance: 0.003209821578552305
Word: fun, Importance: 0.003147397722625146
Word: celebrity, Importance: 0.0030954056039482537
Word: megynkelly, Importance: 0.0027963367966790313
Word: ted, Importance: 0.002764278317993042
Word: snowden, Importance: 0.0023561051231059576
Word: jeb, Importance: 0.002317080142100785
Word: magazine, Importance: 0.002160031429753843
Word: global, Importance: 0.002102557331083346
Word: losers, Importance: 0.00209228982379177
Word: dannyzuker, Importance: 0.002057652979181853

Which account is Trump, and which is his aids?



Can you think of a research question that you could use machine learning to answer?





Today's class

1. Situating NLP in text analysis
(quant, qual, and NLP)
2. Data cleaning &
preprocessing
3. Text representation /
Vectorization
4. Sentiment analysis
5. Topic modeling
6. Machine learning
7. Large Language Models



LARGE LANGUAGE
MODELS ARE TAKING
OVER!



How does ChatGPT compare to human coders?

- **Task:** Guess the political affiliation of a poster based on a single tweet.
- Tweets are collected from a database of politicians from 12 countries.
- Comparison between ChatGPT-4, experts, MTurk workers, and leading-edge finetuned models

ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning

Peter Thirion^{a*}

^aArmenian Institute of National Rebirth (AINT), Yerevan, Armenia

Received: April 10, 2023; revised: April 10, 2023

Accepted: April 10, 2023; published: April 13, 2023

Editorial handling: Peter Thirion (✉)

© The Author(s). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holders. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Published online: April 13, 2023

ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning



Get the paper
↑



**“A friend, a neighbor, a mom, who could soon
serve on the highest court in the land”**



“A friend, a neighbor, a mom, who could soon serve on the highest court in the land”

“This tweet is likely referring to the nomination of Amy Coney Barrett to the Supreme Court, which took place during the specified time frame. Barrett was nominated by then-President Donald Trump, a Republican, and her nomination was supported by the Republican party. The positive tone of the tweet suggests that the poster is in favor of her nomination, which leads me to believe that the poster is a **Republican**.”



“Down Syndrome Awareness Month helps raise awareness for what it means to have Down syndrome and how individuals with Down syndrome play a vital role in our lives and communities.”



“Down Syndrome Awareness Month helps raise awareness for what it means to have Down syndrome and how individuals with Down syndrome play a vital role in our lives and communities.”

“[...] Republicans often emphasize the importance of raising awareness for individuals with disabilities, such as Down syndrome, and their contributions to society. This is sometimes connected to their pro-life stance [...]”

2/3 experts and 8/10 Mturkers incorrectly guessed that this was written by a Democrat.





ChatGPT outperformed everyone, even human experts

- Across national contexts and languages, LLMs outperform everyone.
- They are better than humans even at “distinctly human capacities”, such as subjective interpretation, contextual inferences, implicit meaning...

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Model	Accuracy	F1 Score
Zero-shot GPT4	<u>0.894</u>	<u>0.898</u>
Naïve Bayes	0.766	0.741
BERT	0.828	0.822
MTurk average	0.804	0.820
MTurk plurality	0.854	0.866
Experts average	0.830	0.837
Experts plurality	0.860	0.855



LLMS CAN CARRY OUT NEARLY ANY TEXT ANALYSIS TASK!

- Easy to use. Simply call the API with a text analysis request:

User: "What is the emotional sentiment expressed in the following Twitter message. Respond only with the name of the feeling.

'I feel a profound warmth in my heart when I see you, a gentle stirring, it courses through my veins, filling every fiber with a radiant light that only the heart can understand and only the soul can express.'"

ChatGPT: "Love"



PARADIGM SHIFT IN THE MAKING?

- An explosion of use. The ease-of-use, high accuracy, and relatively low costs of LLMs have meant that they have become hugely popular.
- While computational methods have before made social science more and more technically challenging, LLMs have had the opposite result: empowering qualitative scholars.
- Every underfunded ECR can now afford coding large datasets!



DIS/ADVANTAGES WITH LLMS



1. Powerful, versatile, no training data required
2. Very easy to use using API: simple as getting the weather!
3. Bridges quant and qual by enabling computational interpretation
4. Works across languages and contexts!



1. Expensive and much slower than supervised models:
can't be used on large datasets
2. Success depends on task (Rathje et al., 2023; Pangakis et al. 2023)
3. Black-boxes





Wild West of
LLM-based
text annotation



The 8 commandments of LLM text annotation

1. Choose an appropriate model
2. Follow a structured coding process
3. Develop a Prompt Codebook
4. Validate your model
5. Use structured prompts
6. Consider ethical and legal implications
7. Prompt stability analysis
8. Consider that your data may be in the training data.



COMBINING LLMS AND MACHINE LEARNING

- We can use LLMs to provide training data to a supervised model (e.g., BERT)

Have the cake and eat it too!

- This is a versatile approach that can deal with nearly any annotation task!
- Fast, cheap, scalable

- We will do this in Week 5 (the Machine Learning week)
- This week, we will just test ChatGPT through the web interface



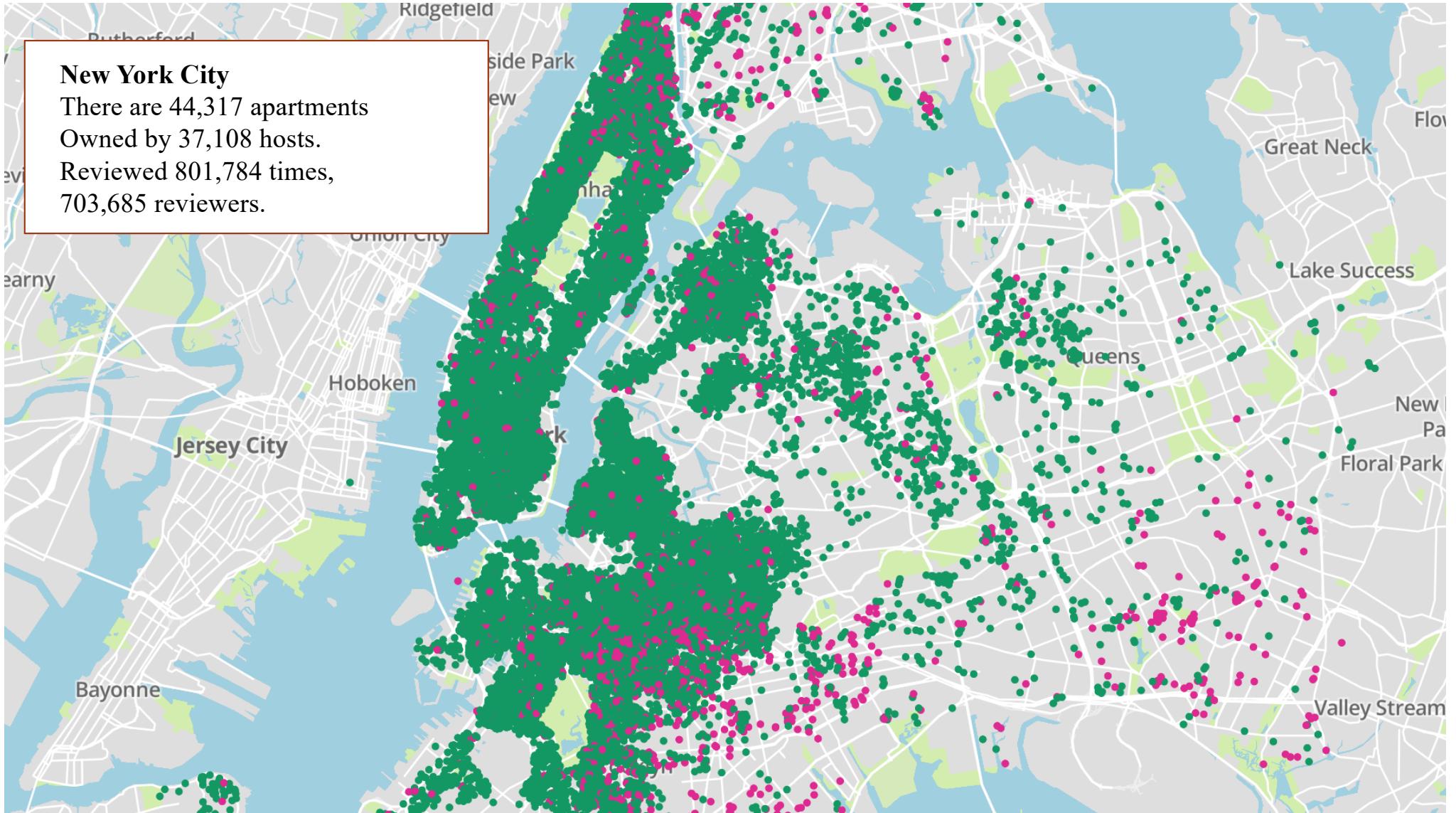
Can you think of a research question that you could use LLMs to answer?



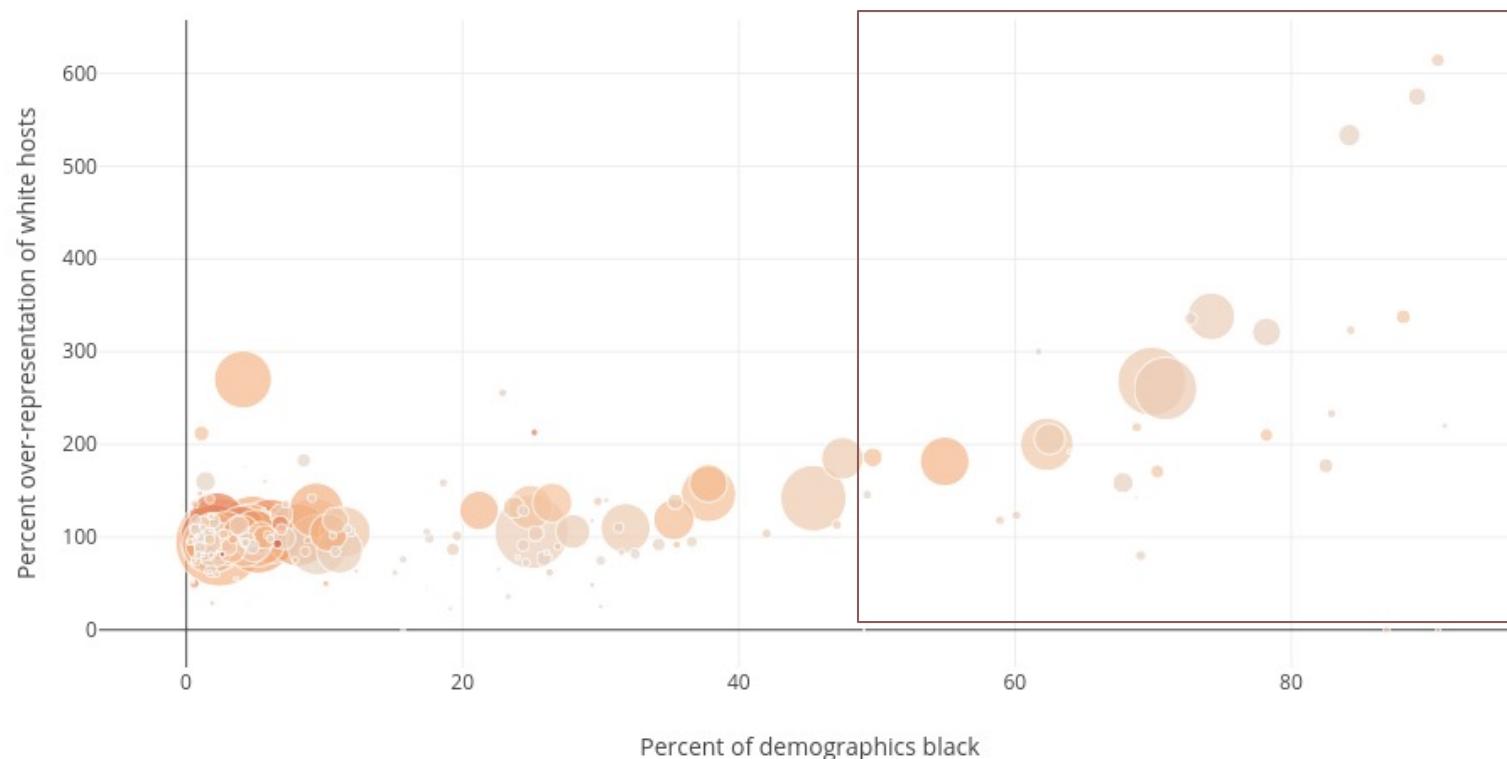


Today's class

1. Situating NLP in text analysis
(quant, qual, and NLP)
2. Data cleaning &
preprocessing
3. Text representation /
Vectorization
4. Sentiment analysis
5. Topic modeling
6. Machine learning
7. Large Language Models
8. Extra: Case study 1: *Race & Airbnb in New York City*



White hosts are over-represented in many black majority neighborhoods



Hosts framing neighborhoods



The host canceled this reservation 7 days before arrival. This is an automated posting.

Hosted by Jordan

New York, United States · Joined in January 2017

7 Reviews Verified



Contact host

Always communicate through Airbnb · To protect your payment, never transfer money or communicate outside of the Airbnb website or app.
[Learn more](#)

347 kr SEK per night

★★★★★ 6

Dates

Check In → Check Out

Guests

1 guest

Book

You won't be charged yet

Report this listing

The neighborhood

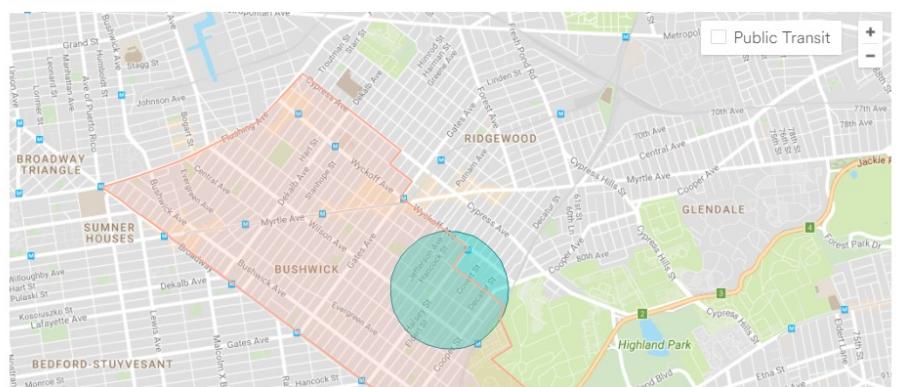
Features · artsy · gritty · latin · hipster · industrial · accessible

Jordan's home is located in Brooklyn, New York, United States.

The Neighborhood boasts a growing number of local bars restaurants, and clubs. Within a couple minute walk you will find 'Nowadays' Beer Garden host of the 'Mr Sunday' daytime dance party every Sunday. 'Industry' across the street has a Karaoke every Tuesday night. New venues and events are popping up all the time, so Explore! Within just a few stops on the L many more options open up including the House of YES (who host unique parties almost every night of the week), a number of great vintage shops, bars, and restaurants. My favorite being Los Hermanos, a tortillaria that distributes all over the world and serves up a mean quesadilla for \$4 (BYO).

[Read more about the neighborhood](#) ▾

Things to do in New York



Most overrepresented words

WHITE HOSTS

lakeside institutions
racetrack
vegetable
black-suit keyfood
saturdays football lee
single used supper
designed parkside
hipster read
blessings colonists
flats quirky
bizarre
fabulous
sound sweet
increasingly support
predominantly roberta
smorgasburg
predominantly
writers native
ultra-hip avenues

BLACK HOSTS

Topics as a function of host race

White	
	brooklyn park prospect restaurants bars museum neighborhood great coffee shops
	15 park restaurants neighborhood central walk great away bars minute manhattan
	18 busy general block day native trains possibility away may times
	19 min away walk restaurants park neighborhood central airport many manhattan
	21 book harlem neighborhood well renowned restaurants five white shops busses
	22 park blocks central restaurants area harlem away neighborhood new bars
	22 johns forest diversity simple st commons part several new borough area brooklyn lenox neighborhood restaurant views avenue dinner located museum
	32 neighborhood also police located around rd quiet block accessible middle miles restaurants min park th distance walking green street st
	34 acres security india drinks doorman city town rest door via park away central shopping transportation jamaica casino close neighborhood
	40 walking
	47 48 neighborhood stores safe working kinds restaurants long drug street jamaican
	51 jfk minutes airport away mall stores shopping several supermarkets green
	57 parking salons jamaica neighborhood hair vehicles class safe free years

% black

Black



White hosts

Airbnb as a Platform

“this neighborhood is a **diamond in the rough**, especially in the Summer months you will get that Brooklyn **authenticity** of people chilling, hanging out, barbecuing and enjoying the weather until the small hours of the morning. **If you can't 'walk like you belong', then this isn't the location for you . . .**”
Ocean Hill, Brooklyn

“This area is well-suited for travelers who want the authentic New York experience. Perfect for any **palette- Jamaican bakeries, hipster coffee shops, asian restaurants, 24/7 bodegas, bars, and organic grocery stores are all within walking distance**”

Prospect Lefferts Gardens-Wingate, Brooklyn

“**Bushwick has been recently named as the new Soho.** A neighborhood full of **graffiti art, alternative art galleries, artist studios, bars, coffee places and restaurants.** Close to Williamsburg, **a vibrant hipster neighborhood with restaurants, bars, flee markets, concert venues.**”

Stuyvesant Heights, Brooklyn

“**What I love about my neighborhood is that, although there is a coffee shop and a wine store just down the street, the hipsters haven't yet fully taken over, and the neighborhood still has more of an authentic Brooklyn feel to it.** There are always people hanging out on their stoops, and everyone you pass greets you.”

Stuyvesant Heights, Brooklyn

“Bedford-Stuyvesant ('Bed-Stuy') is a vibrant, constantly growing community in central Brooklyn, New York. **In the last five years, this middle-class neighborhood has exploded with a diverse mix of students, artists, professionals and hipsters of all races, mixing in with the existing neighborly, predominantly African American community.**”

Stuyvesant Heights, Brooklyn

Black hosts

'The feeling of **safety** as you walk outside and realize that a **police station** is directly across the street gives me an extra sense of **security**.', *Stuyvesant Heights, Brooklyn*

"The neighborhood is **very safe**, well-populated, highly residential, family-oriented and is routinely **patrolled by local NYPD**. Various transportation options are nearby.", *Central Harlem South, Manhattan*

"Relatively safe", *Baisley Park, Queens*

"Family oriented neighborhood. **Close access to a Grocery Store, Family Dollar, and on street parking.**", *Stuyvesant Heights, Brooklyn*

"If you are not opposed the ethnic and racial diversity my property of in fact New York is not for you.", *Stuyvesant Heights, Brooklyn*

"*** TO BE CLEAR, **this is a Black neighborhood. We are Black people.** We wouldn't mention this if it wasn't an issue with some renters in the past. **Bigots need not apply :)**", *Prospect Lefferts Gardens-Wingate, Brooklyn*

NLP can be used as part of a critical discourse analysis approach

- Edward Said's concept of **colonial discourse**, as presented in his seminal work "Orientalism," posits that the way the West talks about the East (the Orient) is not an objective description but a 'discourse' shaped by political forces. This discourse reflects the power dynamics of colonialism, where the West constructed knowledge about the East that reinforced its own superiority and justified colonial rule. Said argues that this discourse is not only a reflection of power but also a means of exerting it, as it shapes perceptions and understanding of Eastern societies, often depicting them as exotic, backward, and in need of Western intervention or control.
- Our understanding of the Other is shaped by the economic and power interests of the powerful.
- Gentrified neighborhoods are the urban stages of colonial expansion. Blackness is marketed as a consumption experience.

