Before the Workshop:

go.unc.edu/topmod

A GENTLE INTRODUCTION TO

Text Analysis

Lorin Bruckner

RESEARCH HUB
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

What is text analysis and why do we use it?

What is Text Analysis

A Tool for Answering Research Questions

- What challenges do speech-language pathologists experience when working in pediatric palliative care?
- What perceptions of vaccination are being communicated on Twitter?
- How can we identify Jim Crow laws in North Carolina?
- How do posts on social media affect the early impact of research articles?
- What important pieces of information can be found in historical plague outbreak reports?

What is Text Analysis

Text analysis is the process by which *meaningful information* is extracted from *unstructured* text data.

What is Text Analysis

Text Analysis is related to **Content Analysis**

- Used in **Social Sciences and Humanities** fields
- Technique for **systematically identifying patterns** in different types of communication
- **Difficult and time consuming** to do **manually** without the help of digital tools
- **Manifest content** – what is **said**
  quantitative techniques work better
- **Latent content** – what is **meant**
  qualitative techniques work better

Quantitative Text Analysis

Quantitative Text Analysis

Quantitative Text Analysis Techniques

- Bag of Words
- Classification
- Named Entity Recognition
- Sentiment Analysis

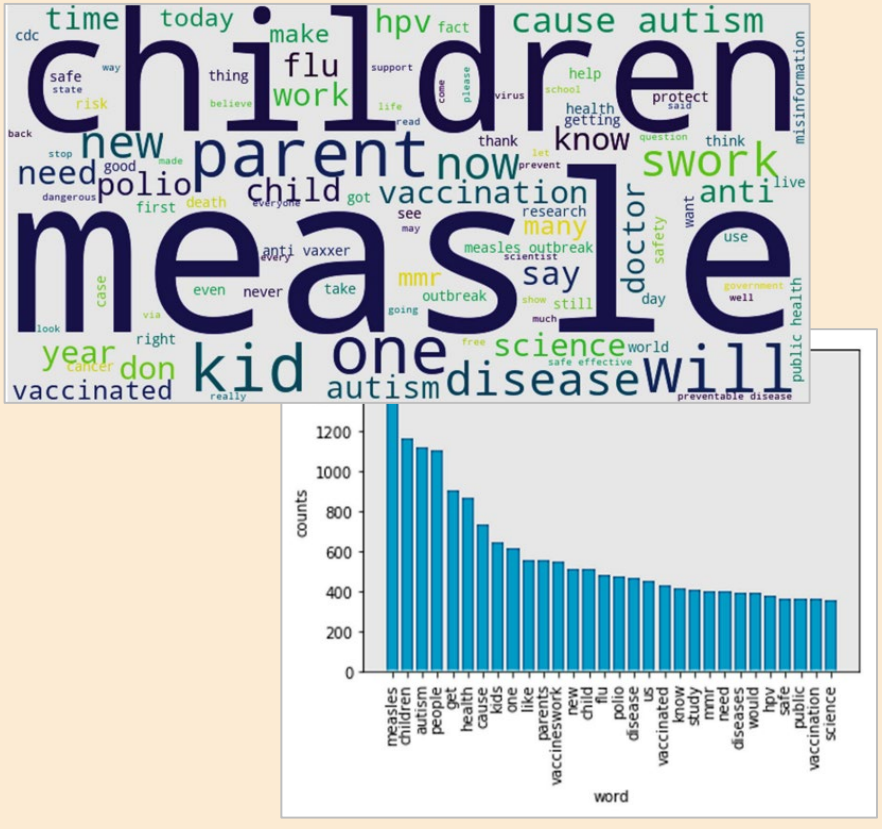
Bag of Words

Bag of Words

EXAMPLE

Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets

- Examined **vaccine-related tweets** from a period in 2019 when a measles outbreak occurred in the U.S.
- Used **bag of words methods** that focus on **how often words occur** in the text, a.k.a. **term frequency**



Raghupathi, V., Ren, J., & Raghupathi, W. (2020). Studying public perception about vaccination: A sentiment analysis of tweets. *International Journal of Environmental Research and Public Health, 17*(10), 3464. doi:https://doi.org/10.3390/ijerph17103464

Bag of Words

What does “Bag of Words” Mean?

Nearly **ALL innate structure is removed** from the text including punctuation, white space and paragraphs, the order in which words occur, etc., so we end up with an **unstructured** “bag” full of jumbled-up words.

Corpus
(e.g., collection of tweets)

Document
(e.g. single tweet)

Bag of Words

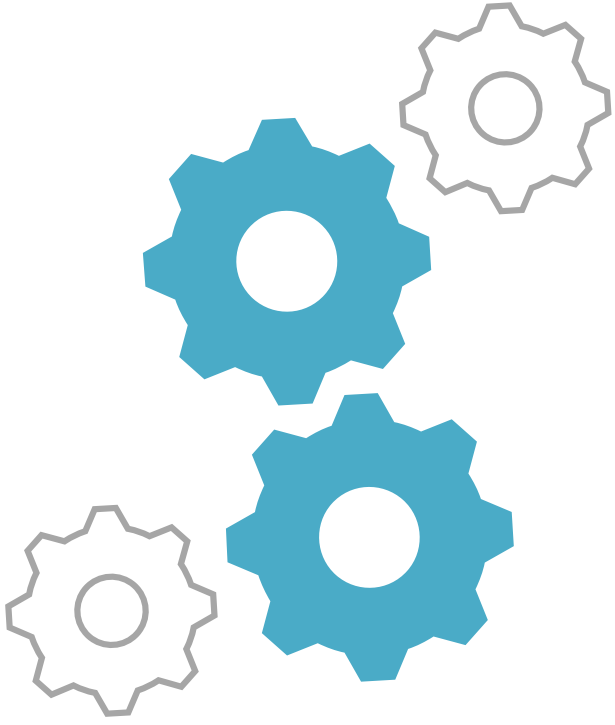
Bag of Words

Preprocessing

- Refers to **transformations** that must be performed on the text **before analysis** can occur

Common Preprocessing Steps

- Tokenization
- Removal of stop words
- Stemming/Lemmatization



Bag of Words

Tokenization

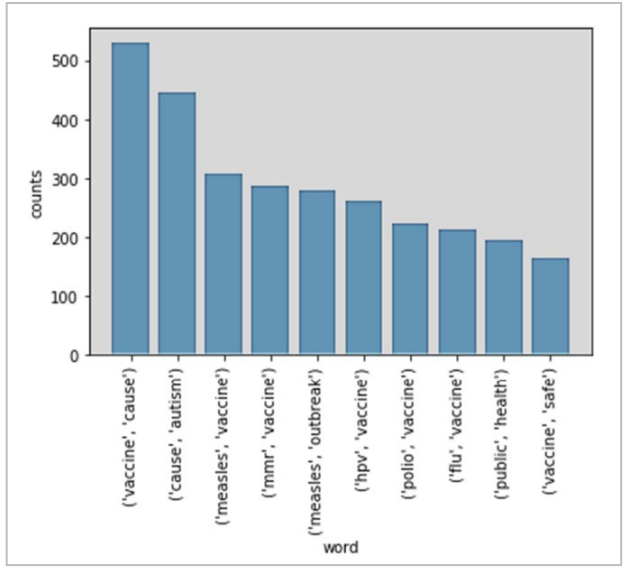
- Breaks a document or corpus into **separate clusters of text** that become the **unit of analysis**

Types of Tokenization

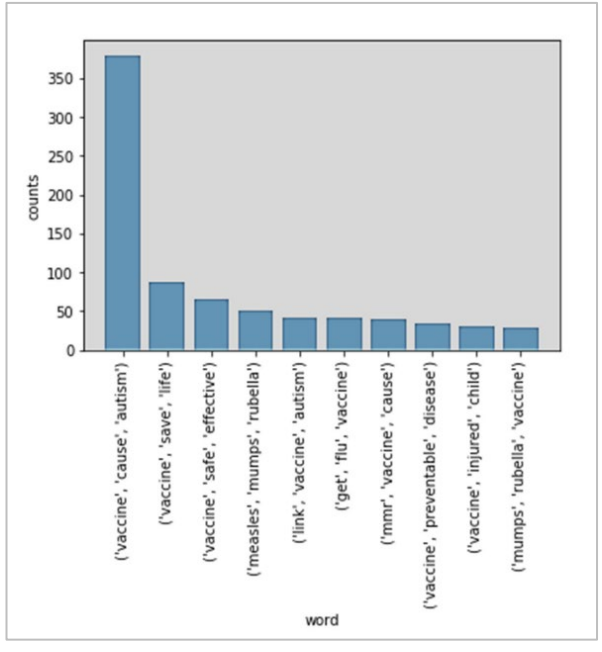
- Word
- N-gram

This is a sentence. → this | is | a | sentence

bigrams



trigrams

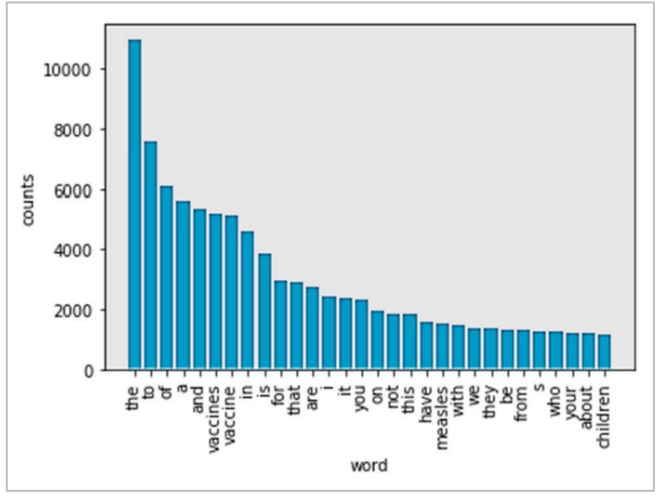


Bag of Words

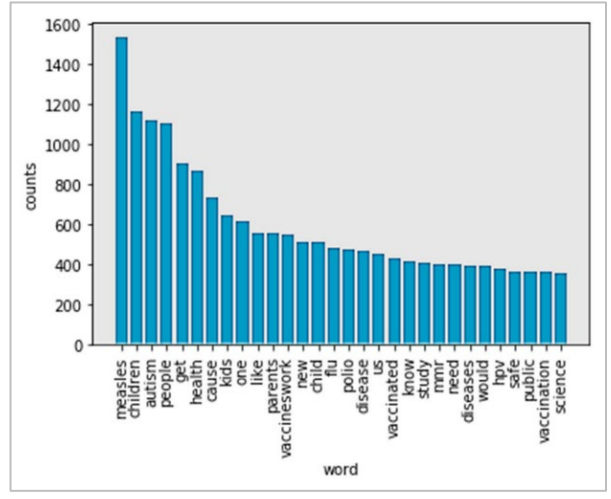
Stop Words

- List of words **removed** from the corpus
- Include **extremely common** words that **aren't** usually very **informative**
- Articles, pronouns, "to be" verbs, etc.

with stop words



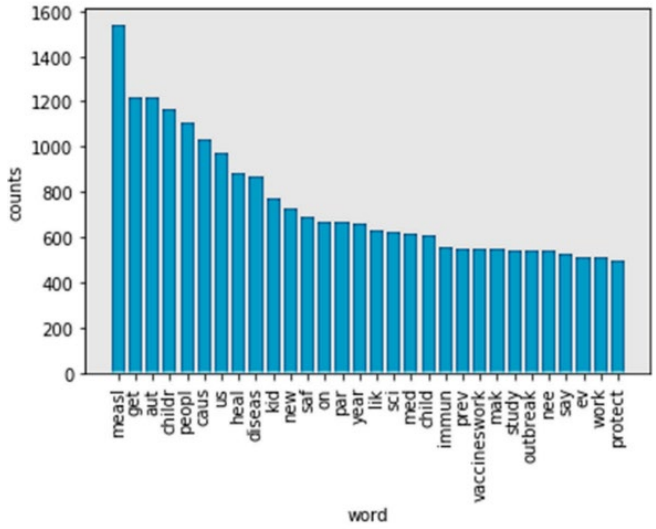
without stop words



Bag of Words

Stemming

- Algorithm **removes** certain **word endings** so similar words can be **counted together** (cars -> car)
- Can result in **inaccurate** word frequencies (caring -> car)



Lemmatization

- Takes **linguistic morphology** into account, providing **better accuracy**
- Algorithm relies on **detailed dictionaries**

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Bag of Words

Term Frequency Method

- Often needs to be **normalized** by the total number of words in a document
- **Number of times a term appears** divided by **the total number of terms**

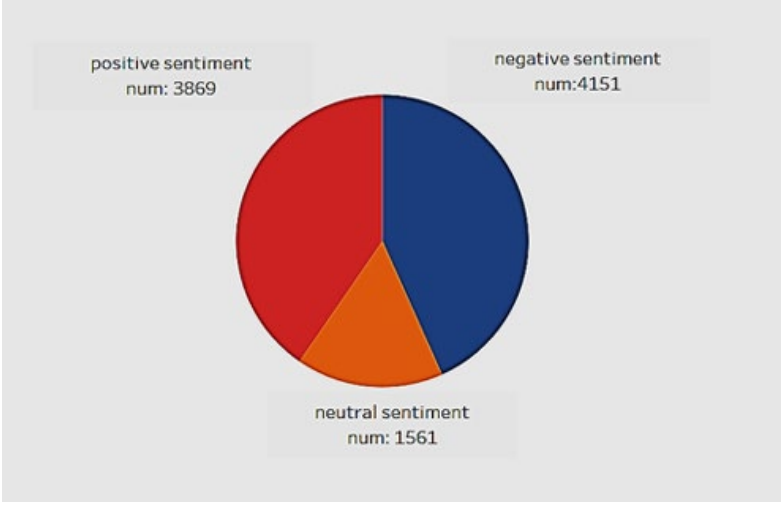
$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

TF-IDF Method

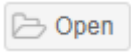
- Uncommon terms are **more informative** than common ones
- TF-IDF takes into account both the **frequency of a term in a document** and its **overall scarcity in the corpus**

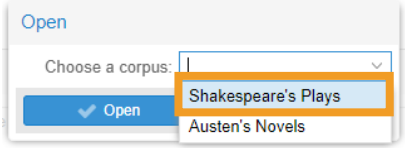


Bag of Words

EXERCISE

Shakespeare's works in Voyant

1. Go to **voyant-tools.org**
2. Click on 

3. Choose 

Tips

- Open tool in a **new tab** (click **Export**)
- **Change tool**
- Change **options** (more at bottom of tool)
- Tool **help**

- What is the **corpus**? What are the **documents**? **How many** documents are in the corpus? *(Hint: Summary tool)*
- What is the **most common 4-gram** in the corpus? In what **document** does it have the **highest** relative **frequency**? *(Hint: Phrases tool and Trends tool)*
- Edit the **stop words** so that the **top 20 terms** are more **informative**. *(Hint: Terms tool - Options)*
- Find a term that Shakespeare uses **more often** in his **early** works than his **later** works. *(Hint: Terms tool and Trends tool)*
- Look at the **highest Significance scores** for **Document Terms**. Why are they all **character names**? *(Hint: Document Terms tool - Help)*

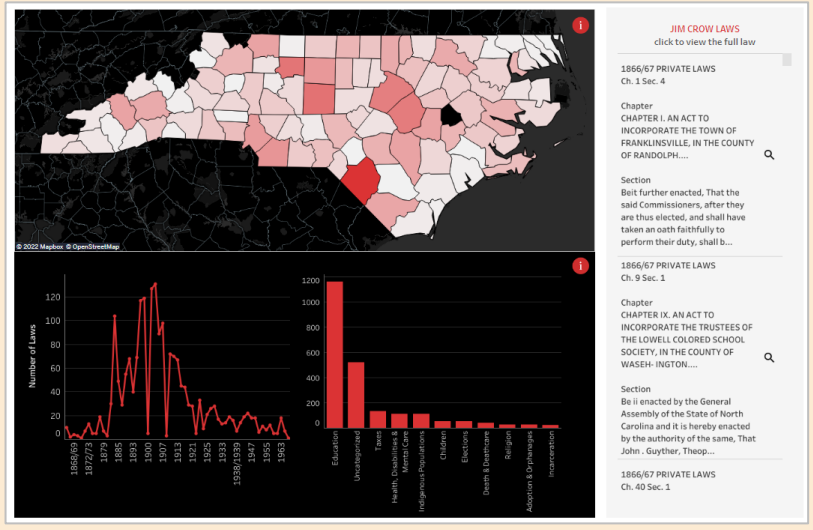
Classification

Classification

EXAMPLE

On the Books: Jim Crow and Algorithms of Resistance

- Discovered **Jim Crow and racially-based legislation** signed into law in North Carolina between 1866 and 1967
- Used **supervised and unsupervised classification techniques** to identify laws with race-based language and the type of legislation covered by those laws



SEC. 9. That it shall be the duty of said School Commissioners to establish graded schools in said town, one for white children and one for colored children, and to appropriate the funds derived from said special taxes and from all other sources, for the maintenance of said schools so as to equalize the school facilities between the races.

Henley, A., Jansen, M., Bruckner, L., Byers, N., & Dalwadi, R. (2020). *On the Books: Jim Crow and Algorithms of Resistance White Paper*. https://doi.org/10.17615/hvz4-sr14

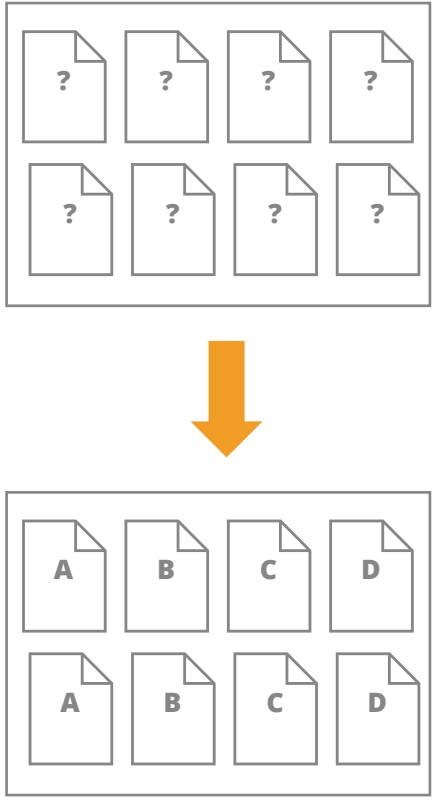
Classification

Uses for Classification

- Separate documents into **different groups** depending on the language used in the text
- Often called **topic modeling** when the groups represent topics

Types of Classification

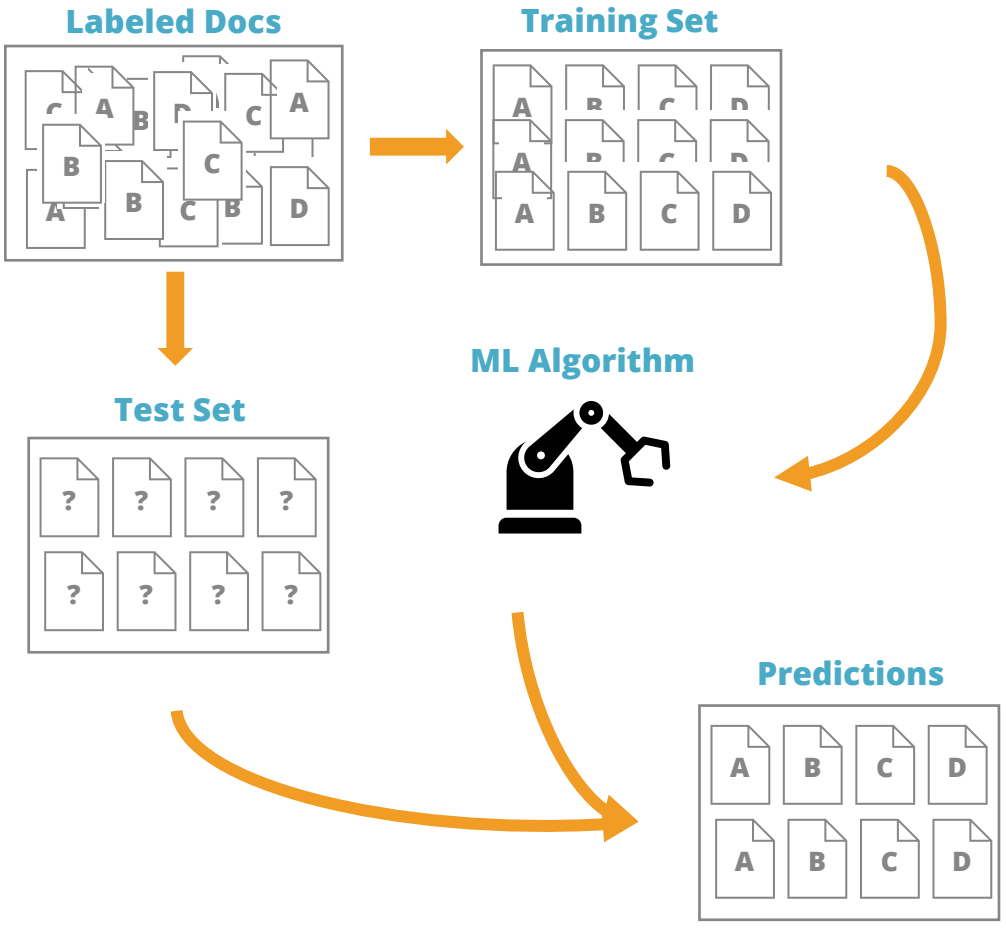
- Supervised
- Unsupervised
- Rule based



Classification

Supervised

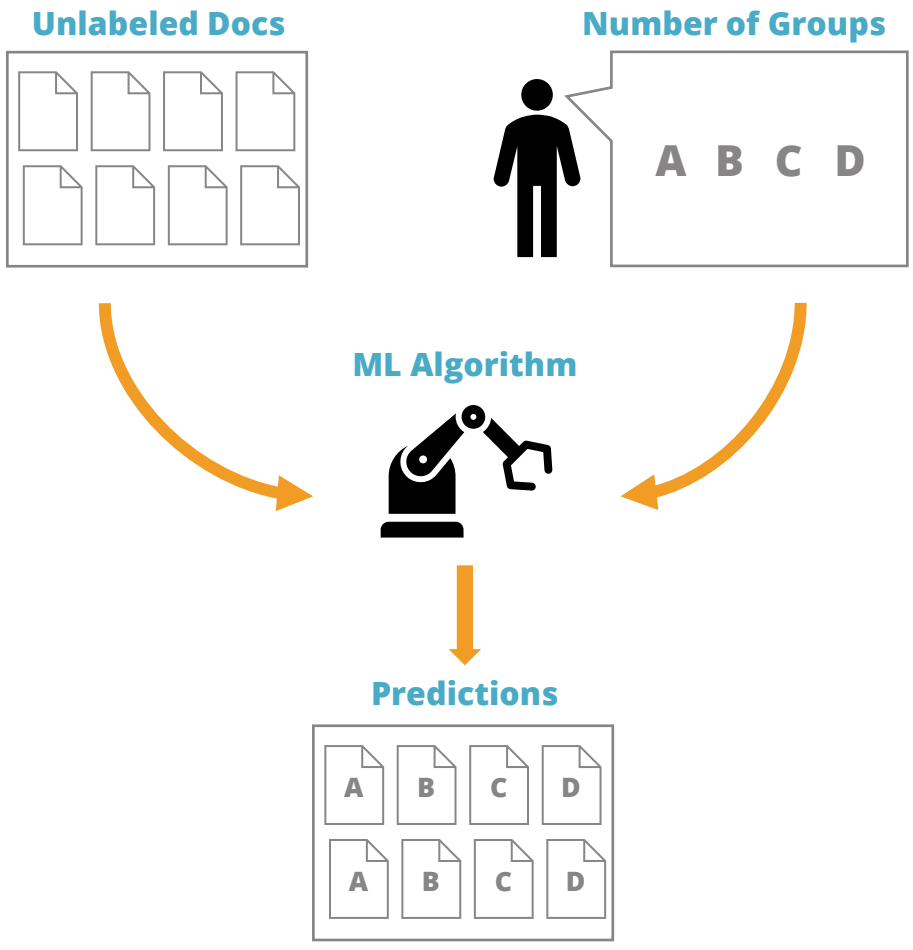
- Portion of the corpus **must be manually labeled** to provide a training set (very **time consuming**)
- **Training set** serves as an example of **correctly classified text** (it's the supervisor)
- **Machine learning algorithm** uses the labels supplied by the training set to **predict classifications** for documents in a **test set,** and eventually unlabeled documents
- Often, **many different algorithms are trained, tested and compared** to find the best results, e.g., Naïve Bayes, Random Forest, Support Vector Machine, XGBoost



Classification

Unsupervised

- **Does not require manually labeled text** and is **less time consuming**
- Due to lack of labeling, results are often **less accurate and harder to evaluate**
- **Researcher determines number of groups** the algorithm will organize the data into
- **Machine learning algorithm looks for similarities in the data** and creates groups based on those similarities
- **Many different types of algorithms** including Latent Dirichlet allocation, K-Means clustering, word embeddings, etc.



Classification

EXERCISE

Topic Modeling Tool

1. Follow the steps at **go.unc.edu/topmod.**
2. Go to your **Desktop** and open the **output_html** folder.
3. Double-click on **all_topics.html**

Tips

To **learn more about the topic modeling algorithm** we are using, visit:

go.unc.edu/lda

(written by Google employee, Ria Kulshrestha)

- If you had to **come up with a name for each topic**, what would it be?
- **Click on each topic and look at the top five laws within it.** Based on the text of those laws, **how representative** are the topic names you picked?
- **How useful** do you think the results of the topic modeling algorithm are?

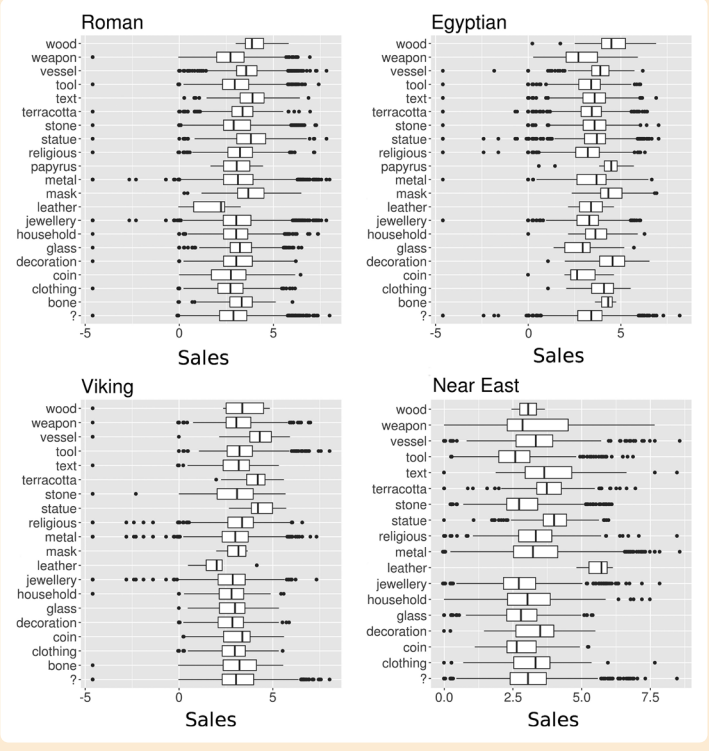
Named Entity Recognition

Named Entity Recognition

EXAMPLE

The Market for Heritage: Evidence From eBay Using Natural Language Processing

- **Examined the trade of cultural objects** and antiquities through **eBay listings**
- Used **Named Entity Recognition** techniques to **detect the names of cultures, items, and materials** within the text of the listings

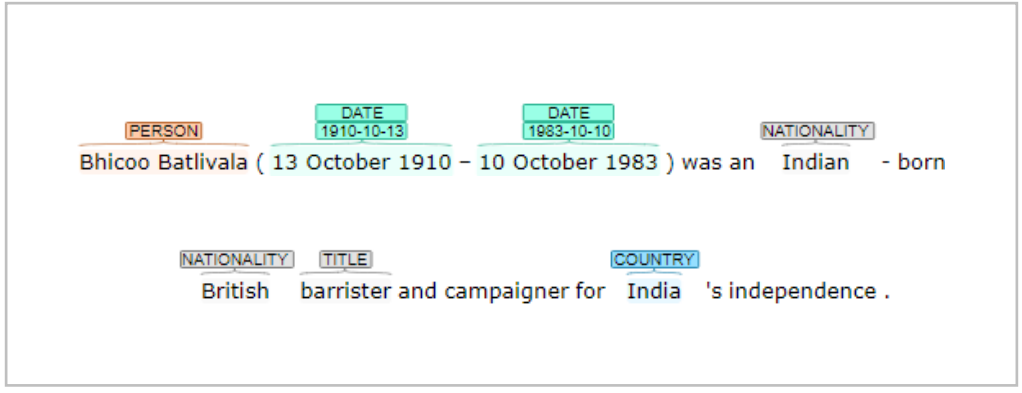


Altaweel, M. (2019). The Market for Heritage: Evidence From eBay Using Natural Language Processing. *Social Science Computer Review*, *39*(3), 391-415. doi:https://doi.org/10.1177/0894439319871015

Named Entity Recognition

Detect Entities in Text

- NER tools **label text** with certain tags
- Tags refer to **various entities** that may represent **important information** such as **people, places, times, etc.**

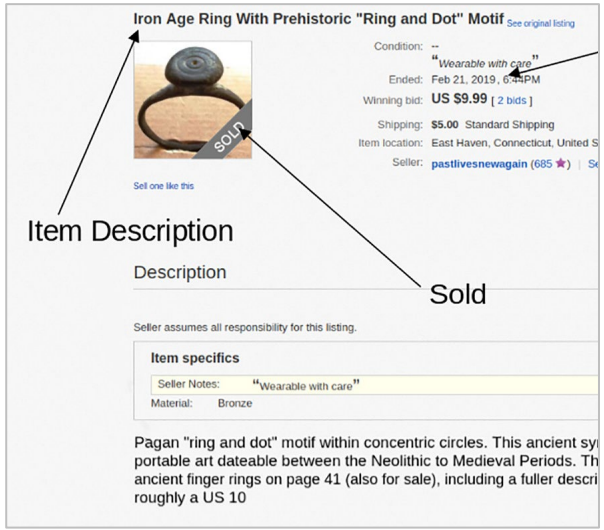


Uses Previously Trained Models

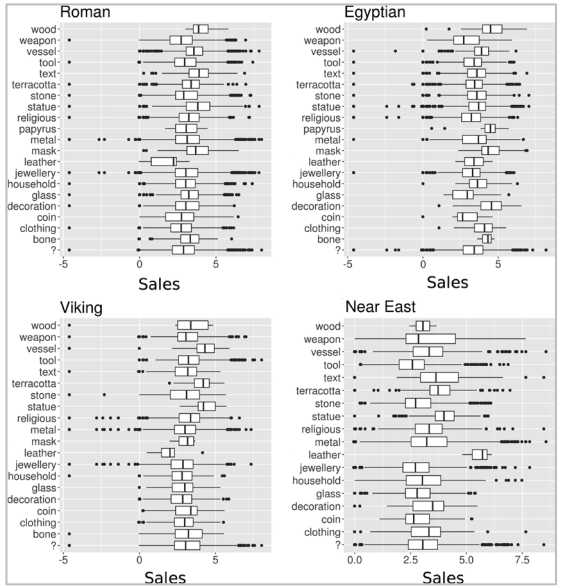
- NER tools typically come **pre-trained** through **supervised machine learning**
- Some tools also **allow you to train the model** with your own labeled data
- Current tools are commonly based on the **Stanford Model**

Named Entity Recognition

NER can turn unstructured text data into a structured dataset, allowing for exploratory analysis.



Type	Total (US$)	Mean
Total	US$2,556,092.00	US$46.7
Roman	US$873,809.24	US$46.94
Egyptian	US$357,256.92	US$47.93
Unknown culture	US$299,434.10	US$48.54
Viking	US$273,632.65	US$41.30
Near East	US$232,599.42	US$45.94
Greek	US$178,332.60	US$49.28
Islamic	US$152,316.15	US$156.06
Medieval	US$139,970.33	US$30.09



Named Entity Recognition

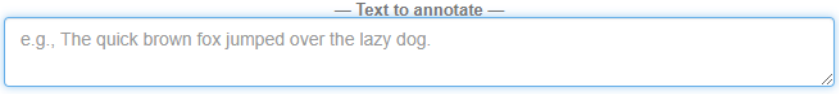
EXERCISE

NER with CoreNLP

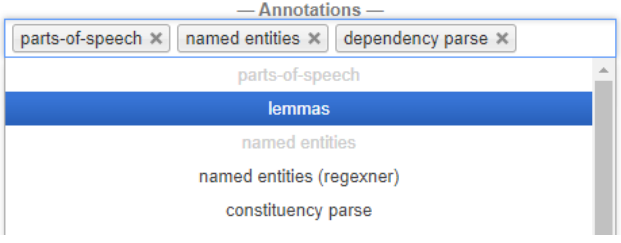
1. Go to **corenlp.run**

Tips

1. Paste text into the box:



2. Add or remove label types:



3. Run the NER tool: Submit

- Go to **wikipedia.com** and search for **"Ynes Mexia"**
- Copy the **first paragraph** of the article and paste it into the text box.
- **Remove** the **parts-of-speech** and **dependency parse** annotations so that **only named entities** is showing.
- Click **Submit**.
- **Carefully read** the text and the labels.
- Are there any entities that were **missed**?
- Are there any entities that were **wrongly labeled**?

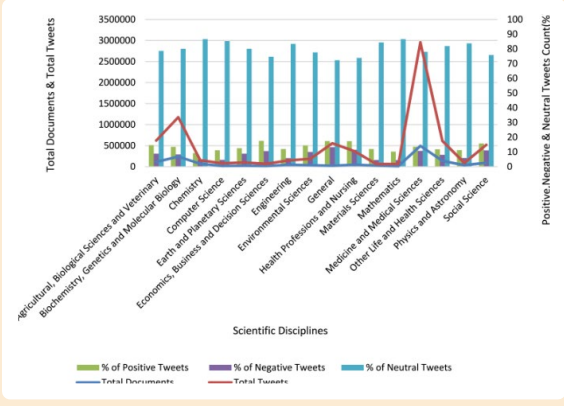
Sentiment Analysis

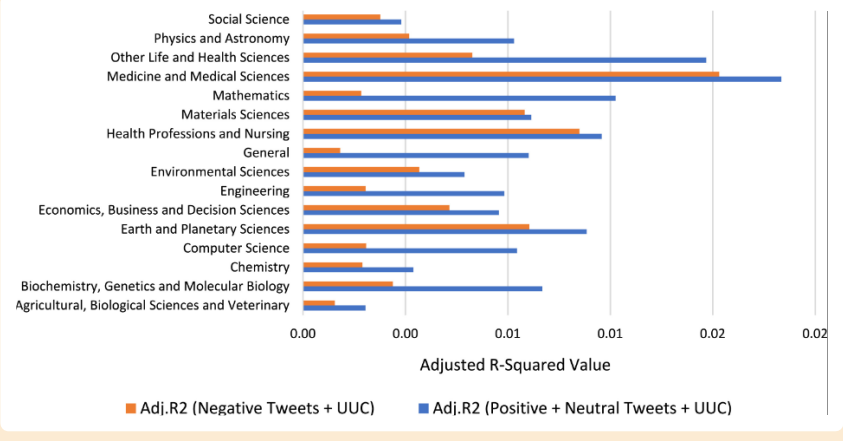
Sentiment Analysis

EXAMPLE

Predicting literature's early impact with sentiment analysis in Twitter

- Examined how **sentiment in posts** on Twitter related to the **early impact of research articles**
- **Used the SentiStrength lexicon to determine sentiment** of tweets, then compared sentiment to citation counts





Hassan, S., Aljohani, N., Idrees, N., Sarwar, R. Nawaz, R., Martínez-Cámara, E., Ventura, S., & Herrera, F., (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems*,192, 105383.doi:https://doi.org/10.1016/j.knosys.2019.105383.

Sentiment Analysis

More of a Goal Than a Technique

- Goal is usually to **determine** whether a given document ultimately maintains a **positive, negative or neutral sentiment**
- More **recently** has branched into detecting **specific emotions, intentions, sarcasm**, etc. (can be **very challenging**)

Techniques Used for Sentiment Analysis

- Thematic Coding (Qualitative Text Analysis)
- Lexicons
- Supervised Classification
- Unsupervised Classification

Sentiment Analysis

Lexicons

- **Dictionaries with "scores"** that reflect the positive or negative sentiment of a word
- Documents are given a **total sentiment score** by adding up scores from words in the lexicon
- Often important to **normalize sentiment by number of words** in the document
- Results can be **finely tuned** by adjusting scores for **negation** ("not great"), **amplifiers** ("very good"), and **de-amplifiers** ("somewhat nice")
- Benefits: can be made **domain specific**; **programmatically easy** to use – **rule based**
- Drawbacks: difficult to accurately score **words with multiple meanings** (e.g., "kind"); rule based systems are often **closely tied to the domain**

AFINN

WORD	SCORE
outstanding	5
distressed	-2
fraud	-4
clever	2

SOCIALSENT

WORD	SCORE
#inspirational	0.984
cancelled	-0.406
#norespect	-0.563
cooperation	0.625

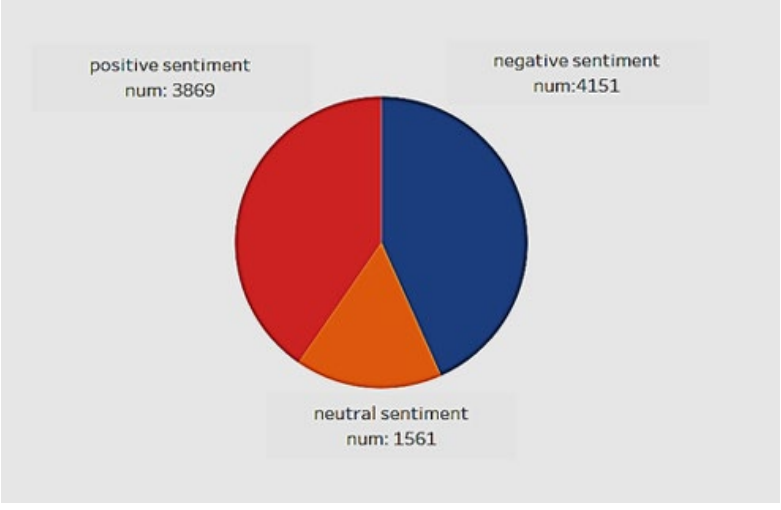
NRC EMOTION

WORD	EMOTION	SCORE
abomination	anger	1
abomination	anticipation	0
abomination	disgust	1
abomination	fear	1
abomination	joy	0
abomination	sadness	0
abomination	surprise	0
abomination	trust	0
abomination	negative	1
abomination	positive	0

Sentiment Analysis

Classification

- Documents are **grouped by sentiment using classification** methods previously discussed
- Can be **supervised or unsupervised**
- **Clustering** is one common method for unsupervised classification as is described in the **Public Perception about Vaccination study**



Sentiment Analysis

EXERCISE

SentiStrength Lexicon

1. Go to **sentistrength.wlv.ac.uk**

Tips

1. Change the text in the box:

Enter text: I love you but hate the current political climate.

2. Change the output type:

Output: ○ Dual, ○ binary, ○ trinary, ◉ scale

3. Click the button: Detect Sentiment

4. Click your browser's back button to retry:

← → ↻ ⚠ Not secure | sentistrength.wlv.ac.uk

- Find the **scaled** sentiment of the following sentence:

  This book is great.

- Without changing or removing words, **add one word to the above sentence** that changes the sentiment to **-1**.

- Without changing or removing words, **add one word to the above sentence** that changes the sentiment to **3**.

- **Change the word "great"** in the above sentence so that the resulting sentiment is **0**.

Discussion:
Problems & Ethics in Text Analysis

Problems in Text Analysis

Discussion

- What issues can lead to inaccurate results when using text analysis techniques?
- How can bias in data affect text analysis?
- How can bias in researchers affect text analysis?
- Could text analysis techniques cause harm to marginalized communities or people?