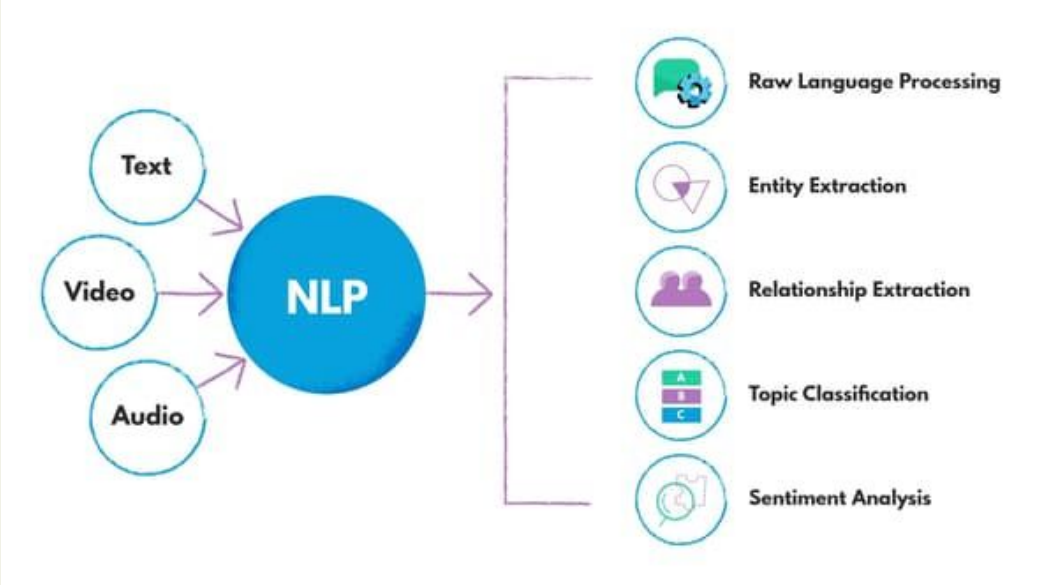
Natural Language Processing

Anne Cuzeau, Sam Mathieu, Eve Dean

Overview

1. NLP Processes
   a. Ways to do NLP
   b. What we used
2. Our Project
   a. Our data & goals
   b. Preprocessing
   c. Sentiment Analysis
   d. Topic Analysis



Ways to do NLP

Approaches:

- Rule-based methods: uses predefined rules and patterns to extract specific information (eg part-of-speech tagging, sentiment analysis)
- Machine learning: uses algorithms to learn patterns and rules from labeled data to classify and extract information from new data (eg named entity recognition, text classification)
- Deep learning: uses artificial neural networks to learn patterns and representations of text data (eg sentiment analysis, language translation)
- Statistical methods: analyze and extract information from text data (eg frequency analysis, clustering)
- Hybrid methods: combines multiple methods, depends on goals

We ar

- Bu
  - on
- In es &
  to
- Se

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

Our Project

Our Data

Kaggle dataset of New York Times Excerpts from 1920-2020.

Each row represents an individual article

Year	1971
Title	Dr J P Eaton of Natl Center for Earthquake Research, other quake specialists, after preliminary study, fear Los Angeles area faces greatest quake ever (W Sullivan rept)
Excerpt	they warn accumulation of strain along fault which skirts San Francisco and passes near Los Angeles threatens quake like one that hit San Francisco in '06; say Los Angeles quake, instead of relieving strain along fault, may have increased it; findings by Dr Franken, Dr Steinbrugge and Dr Franklin outlined Scientists Fear Worse Quake on Coast

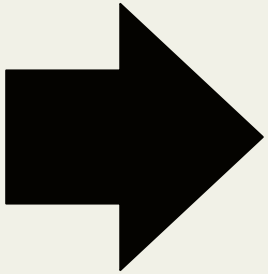
Questions

- Feature extraction
  - What are the common topics, people, and places discussed for each decade?
  - How often are women mentioned in the news over the decades?
- Sentiment Analysis
  - What are the common sentiments associated with women in the news?
  - Has the sentiment in articles involving women shifted?
- Text Classification/Topic Analysis
  - What topics are commonly associated with men vs women?

Dealing with Text Datasets

Challenges with our dataset

- Very large and inconsistent
- File types were not convenient for anything
- Lots of missing excerpts
- Most titles lacked context



How we dealt with it

- Only used a small portion
- Used preprocessing to make text simple & consistent
- Pulled subsets of data to save as CSVs
- Used batch processing
- Combined title and excerpts & treated combined text as a single “text blob”

Preprocessing

PreProcessing

Name Entity Recognition: Looking at text entities

- Identifies noun chunks as entities
- Labels entities: ORG, companies, agencies, institutions, etc.

A new aspirant for Tom Mix's **PERSON** audiences, named Ken Maynard **PERSON**, is introduced in a picture entitled " Señor Daredevil **WORK_OF_ART**," which is sojourning at the Colony **LOC**. As the wages of a cowboy are nominal it is no wonder that occasionally a rough-rider should take a chance before the camera, especially when he hears that Mr. Mix **PERSON**'s income is something like $ 2,000 **MONEY** a day.

PreProcessing

Name Entity Recognition: Part of speech tagging

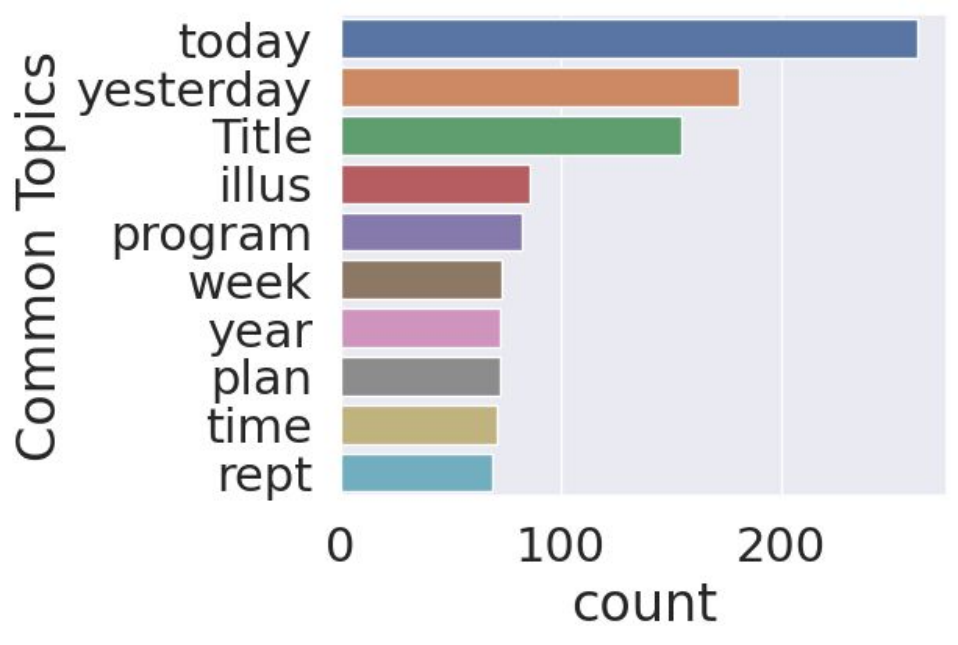
- Identifies what words in excerpts are (noun, preposition, verb, etc.)

```
sentence 1 has noun
chunk 'A new aspirant'
sentence 1 has noun
chunk 'Tom Mix's
audiences'
sentence 1 has noun
chunk 'Ken Maynard'
sentence 1 has noun
chunk 'a picture'
sentence 1 has noun
chunk '"Señor Daredevil'
sentence 1 has noun
chunk 'which'
sentence 1 has noun
chunk 'the Colony'
```

PreProcessing

Looking at what the excerpt talks about: Removing common words

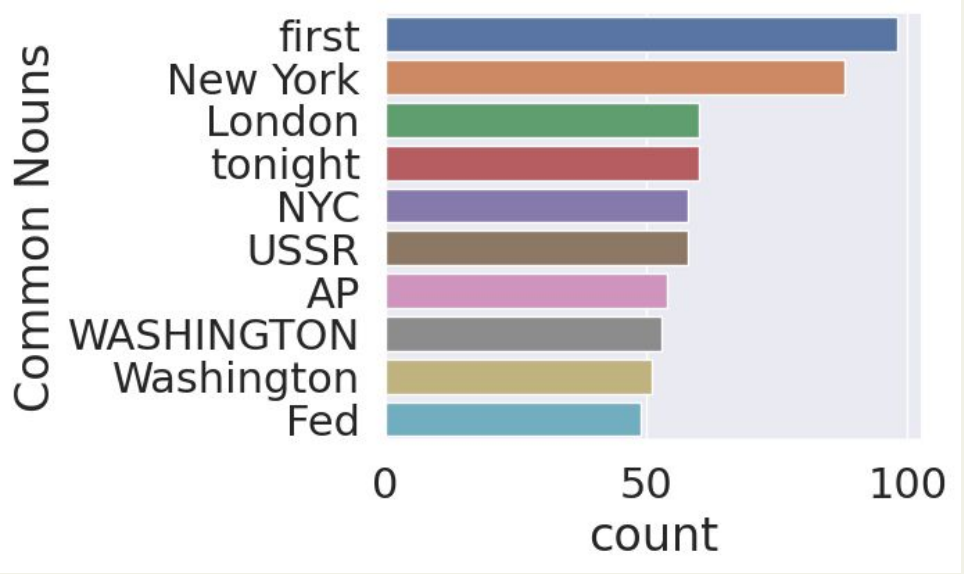
- Common words like a, is, in, the are not useful
- Removing common words provides a better idea of what is important in the excerpt



PreProcessing

Looking at what the excerpt talks about: Exploring common entities

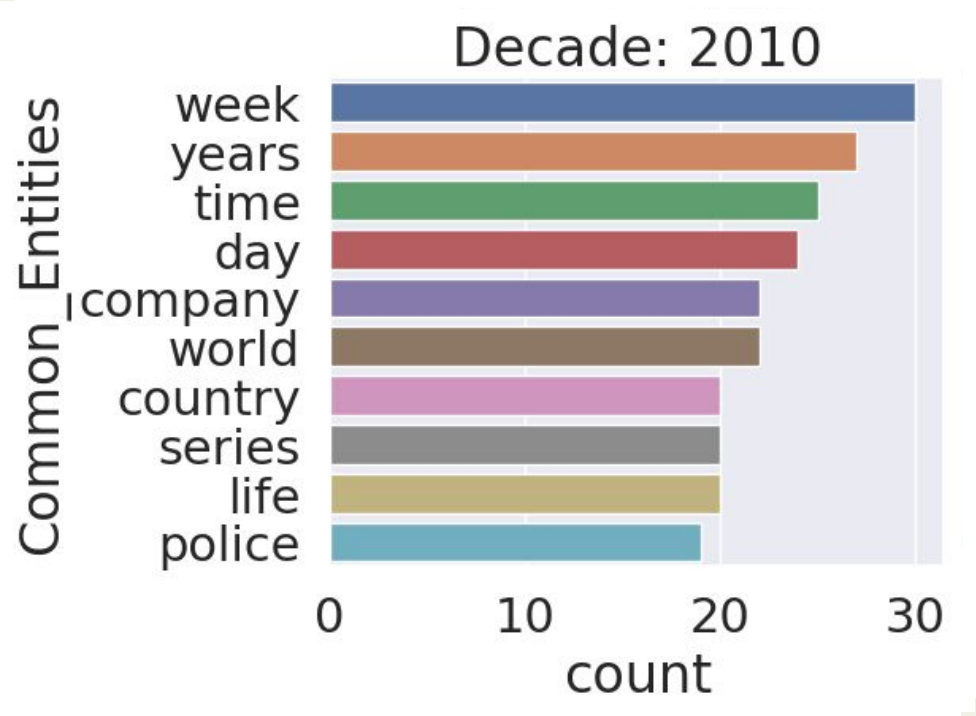
- Removes common words
- Focuses on nouns that are more unique to the excerpt
- Indicates significance of article



Common Entities by Decade

Looking at what the excerpt talks about by decade

- Removes common words
- Focuses on nouns that are more unique to the excerpt
- Indicates significance of article



Topic Analysis

Comparing articles: men vs women

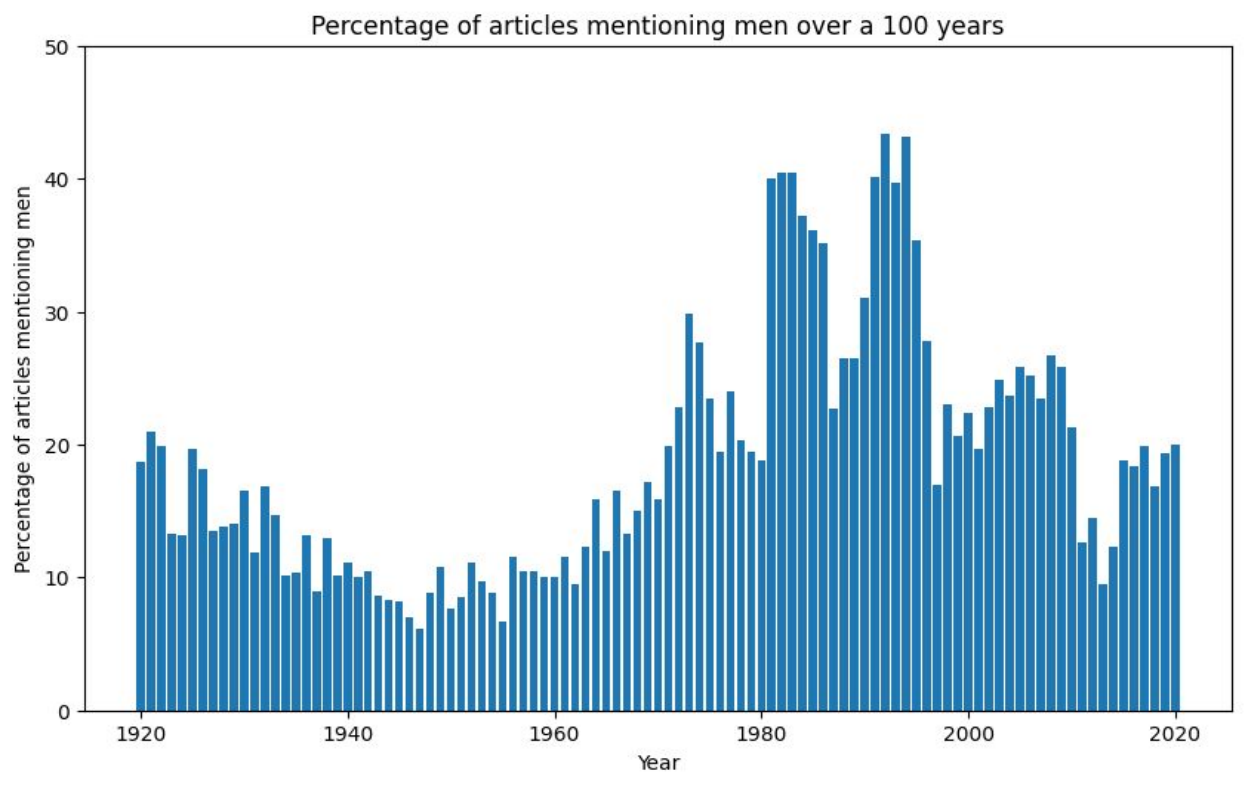
Methodology for 600 articles/year (ran on Google Cloud):

- Isolated articles with pronouns she/her/hers or nouns: woman/women/lady etc.
- Looked at both excerpts and titles: added four boolean columns to our dataset
- Extracted topics from these articles
- Repeated analysis for masculine pronouns and nouns

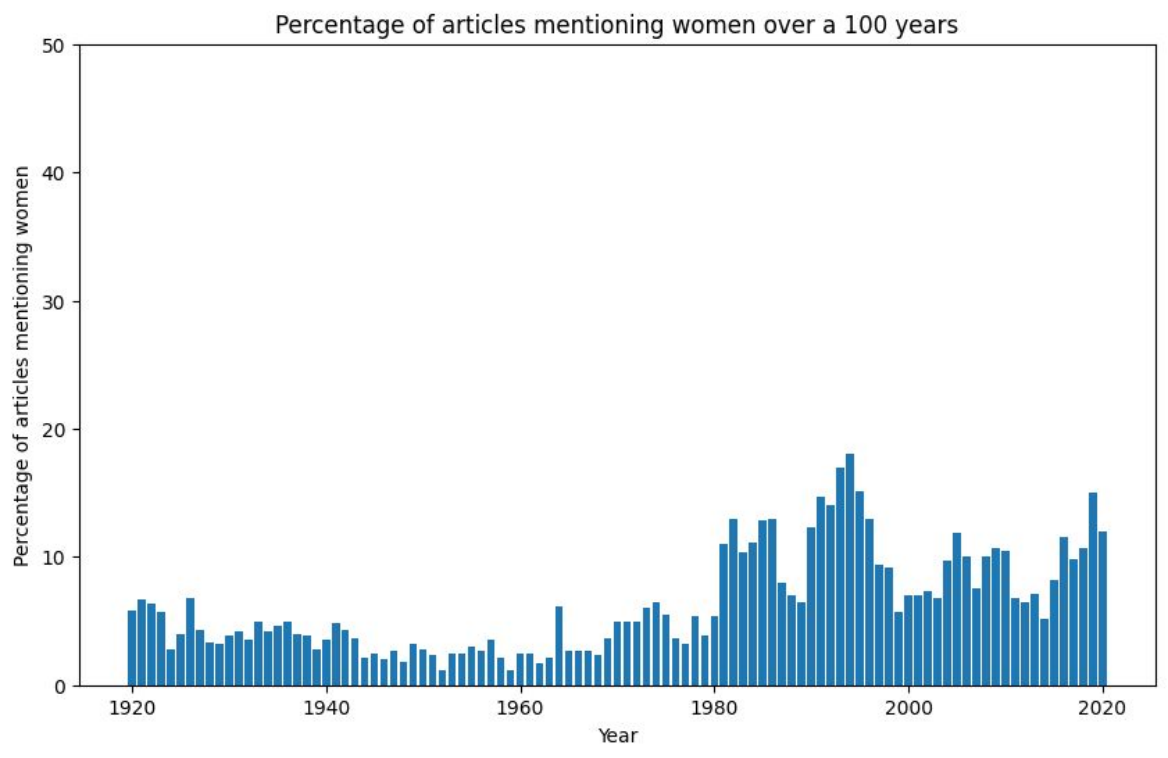
Not using gender-guesser library:

- Storing names in arrays, having to parse through each array, isolate first names → Too much computing for little results

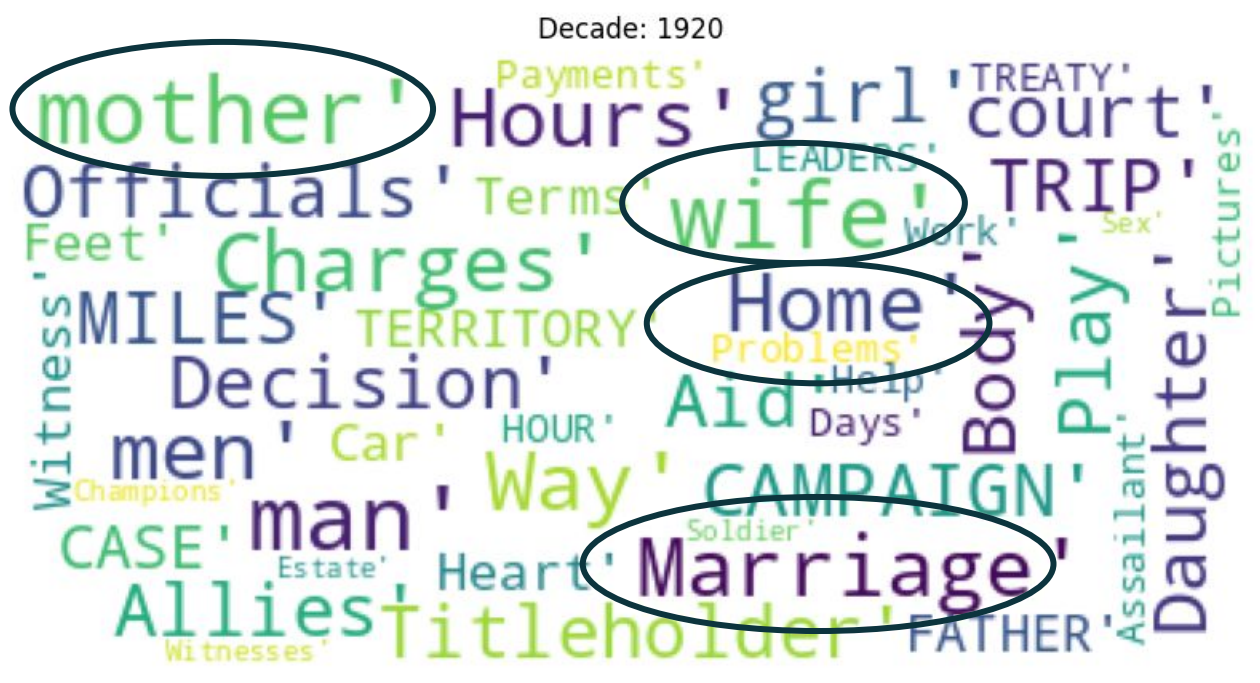
Articles mentioning men vs women



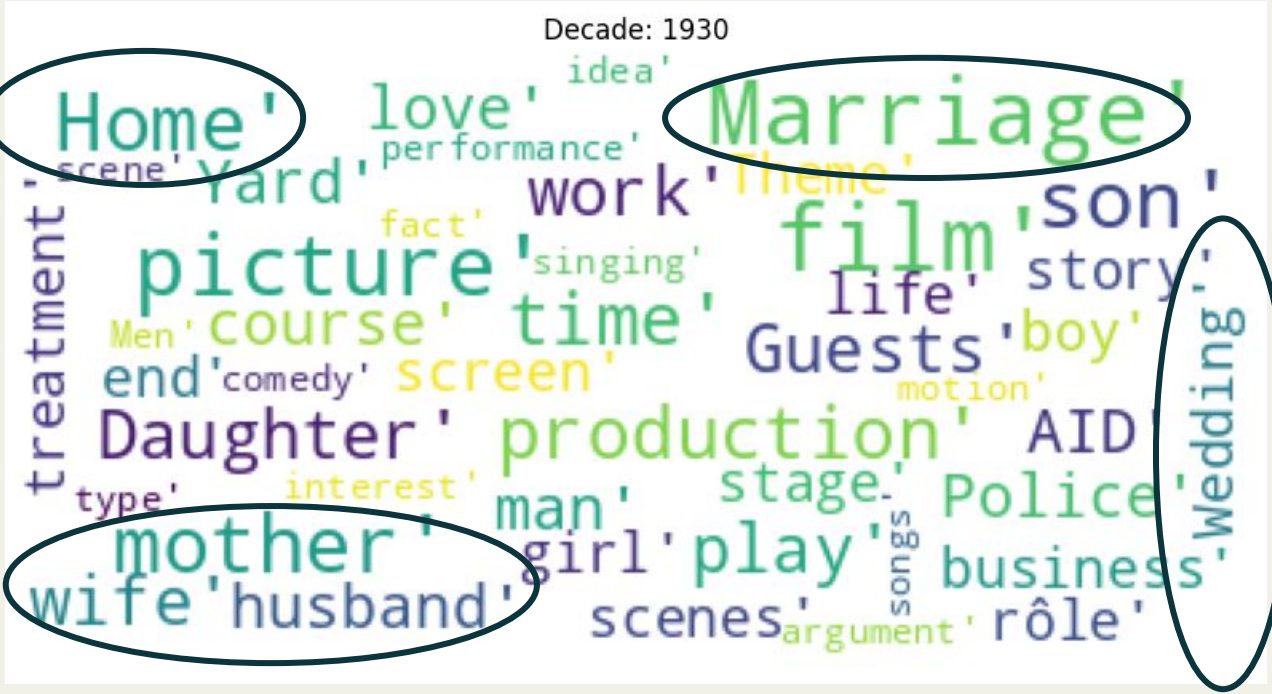
Articles mentioning men vs women



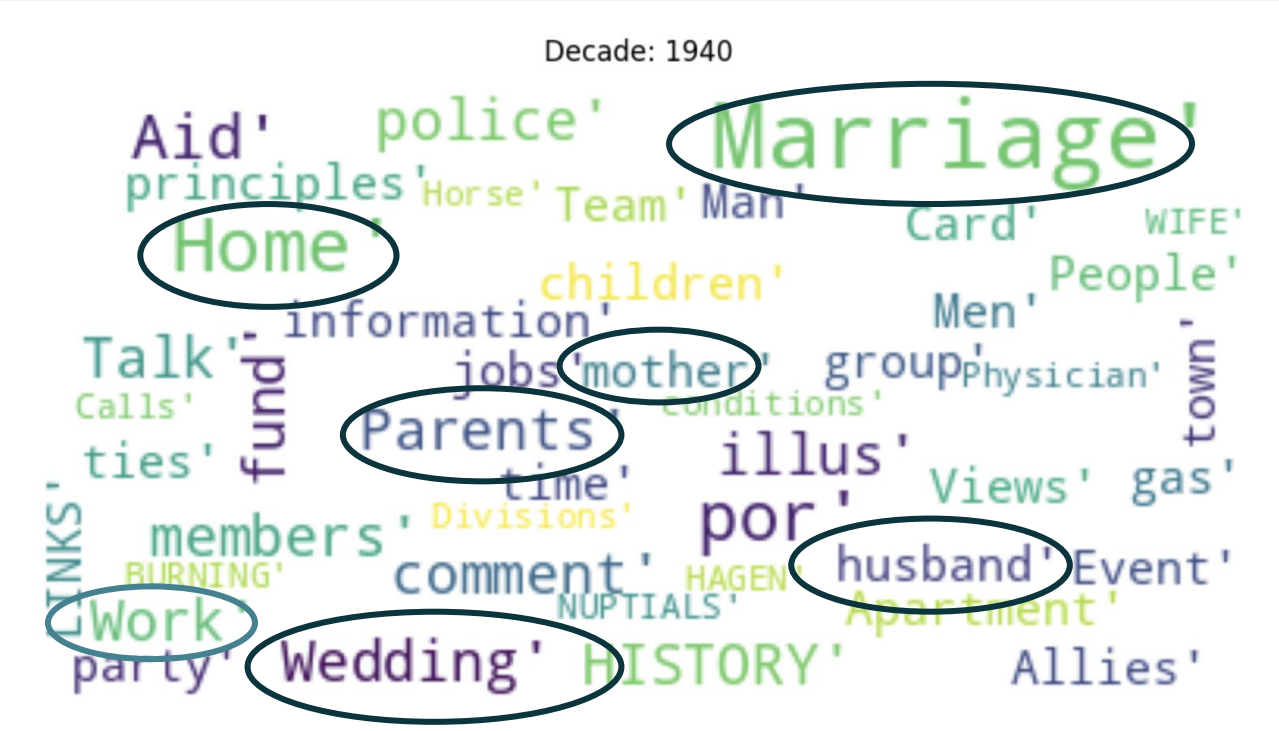
Word Clouds: 1920s



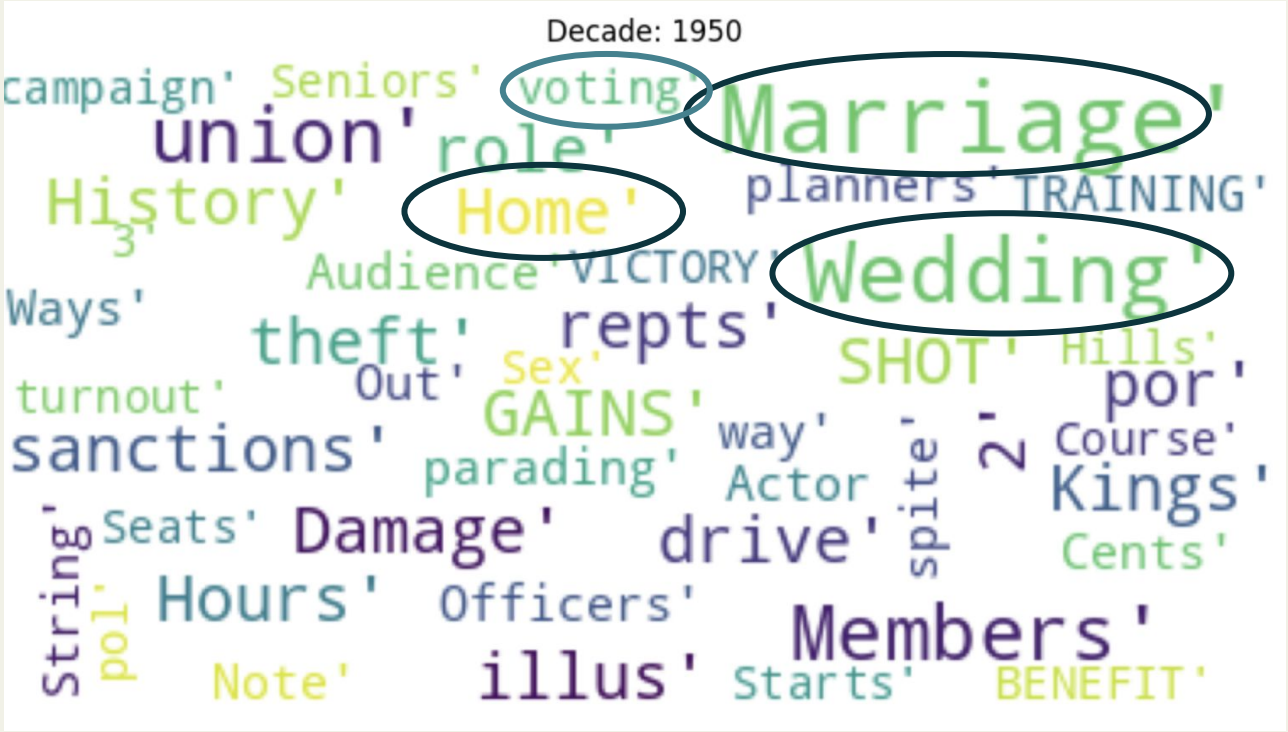
Word Clouds: 1930s



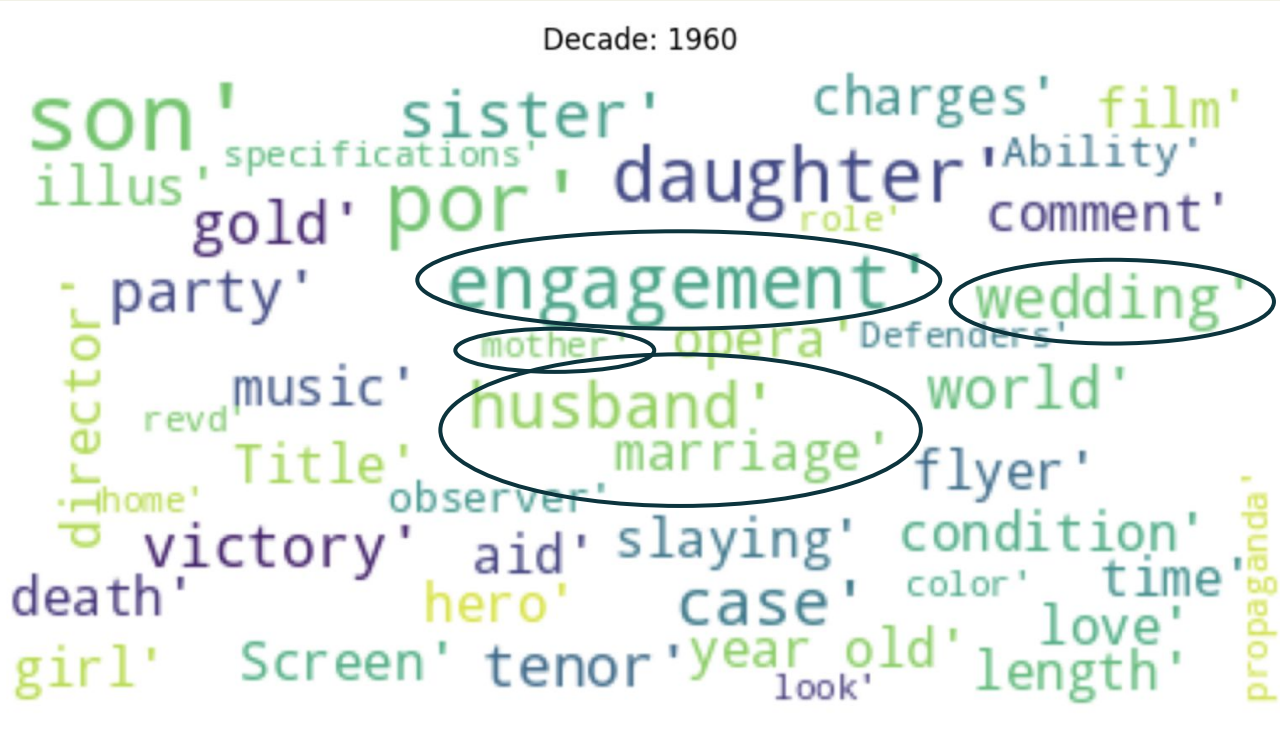
Word Clouds: 1940s



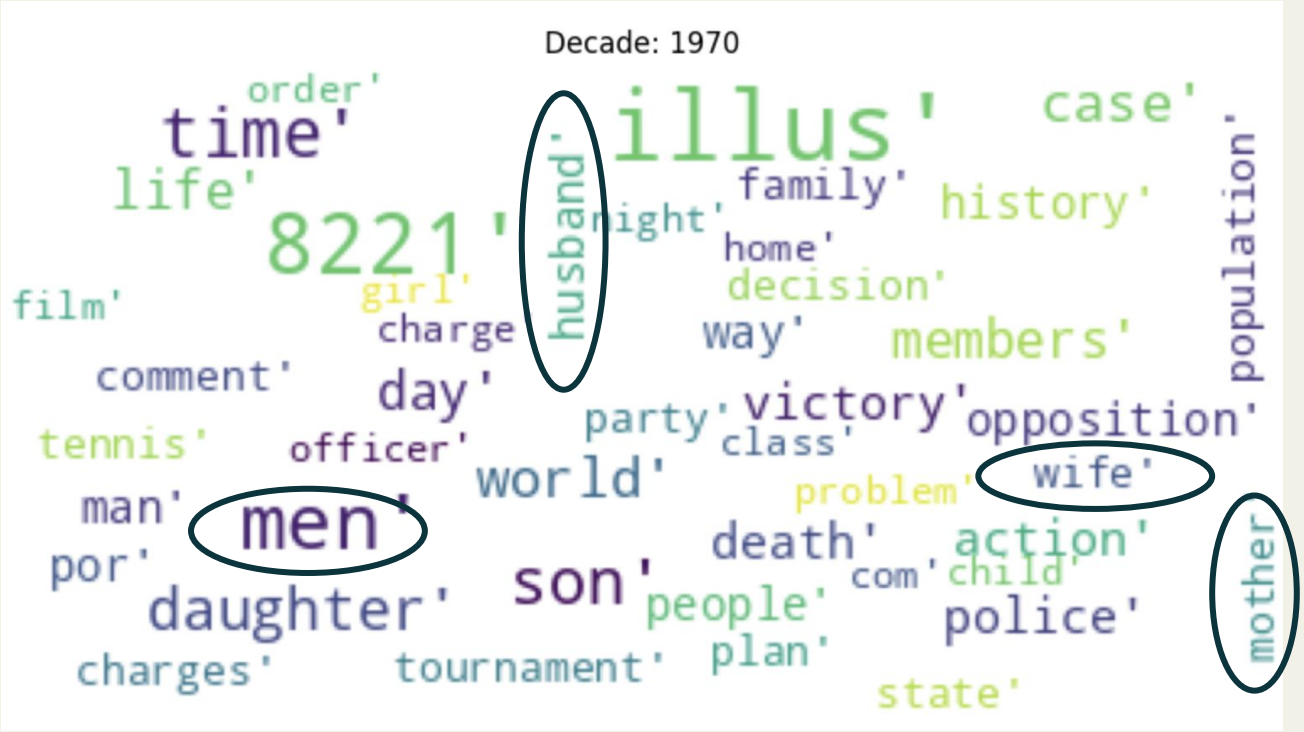
Word Clouds: 1950s



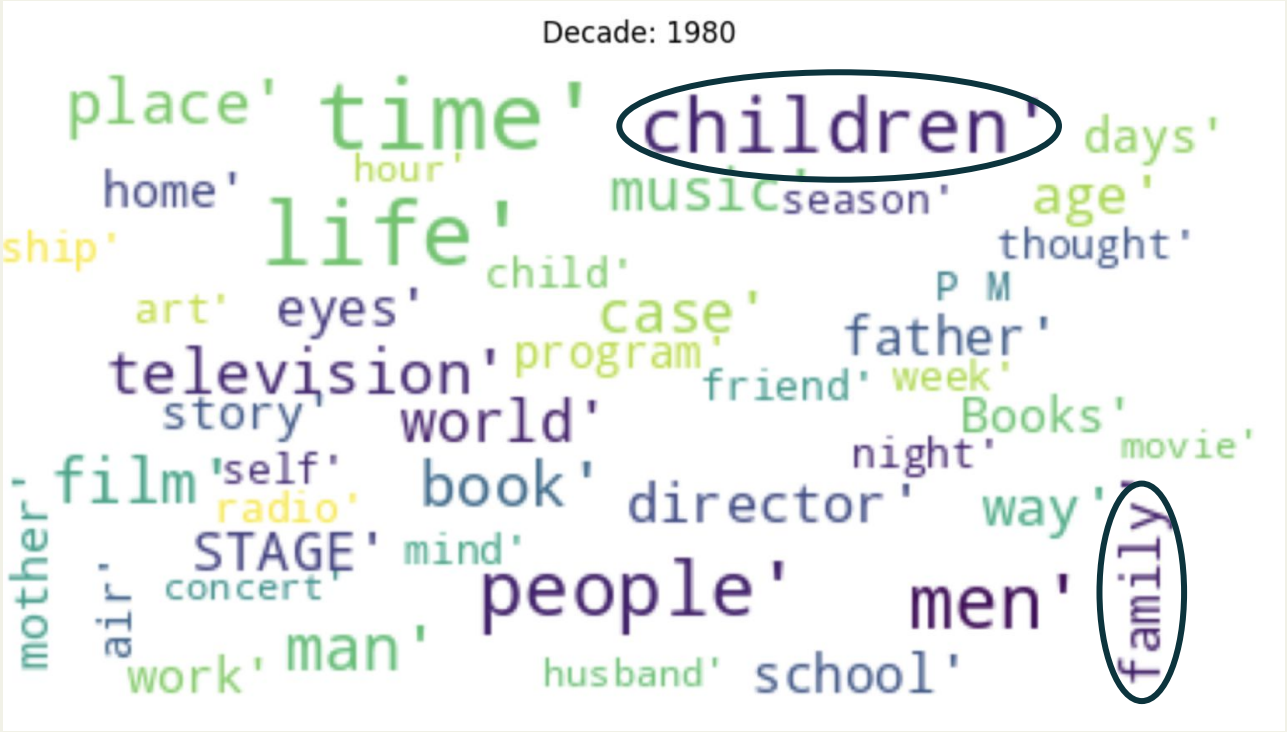
Word Clouds: 1960s



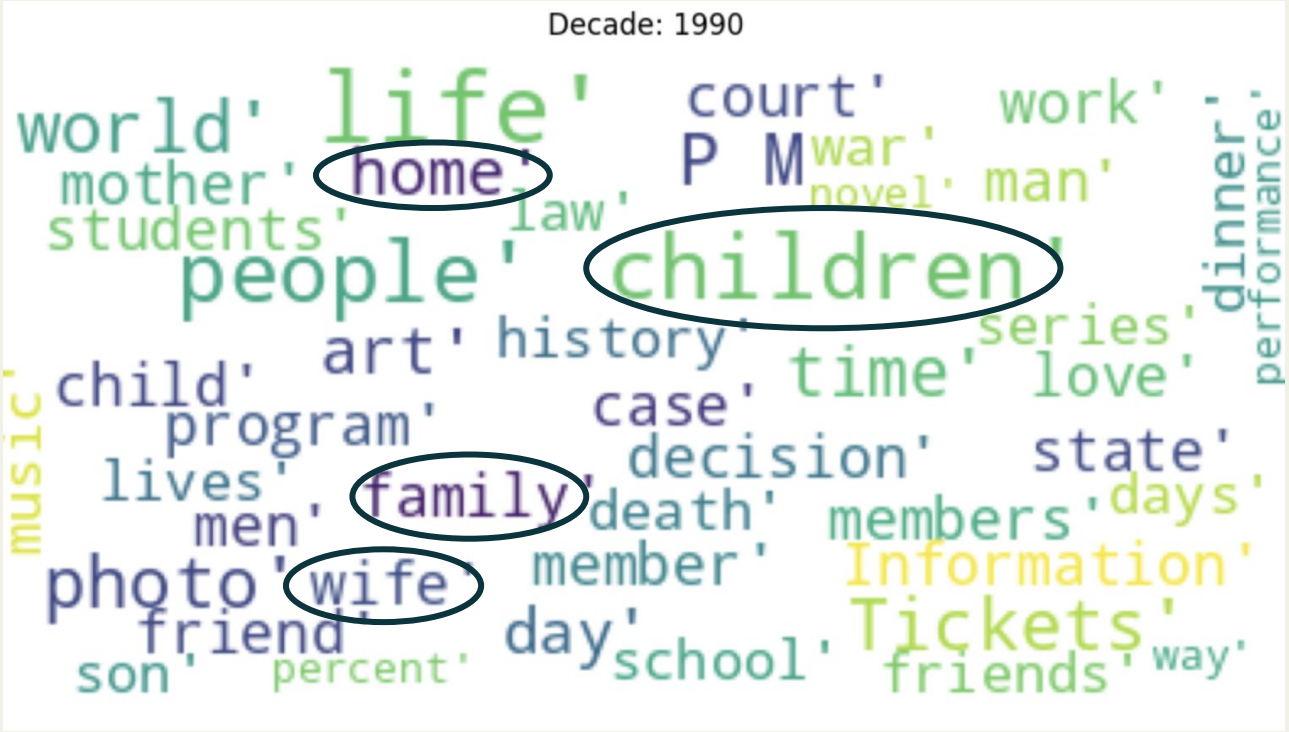
Word Clouds: 1970s



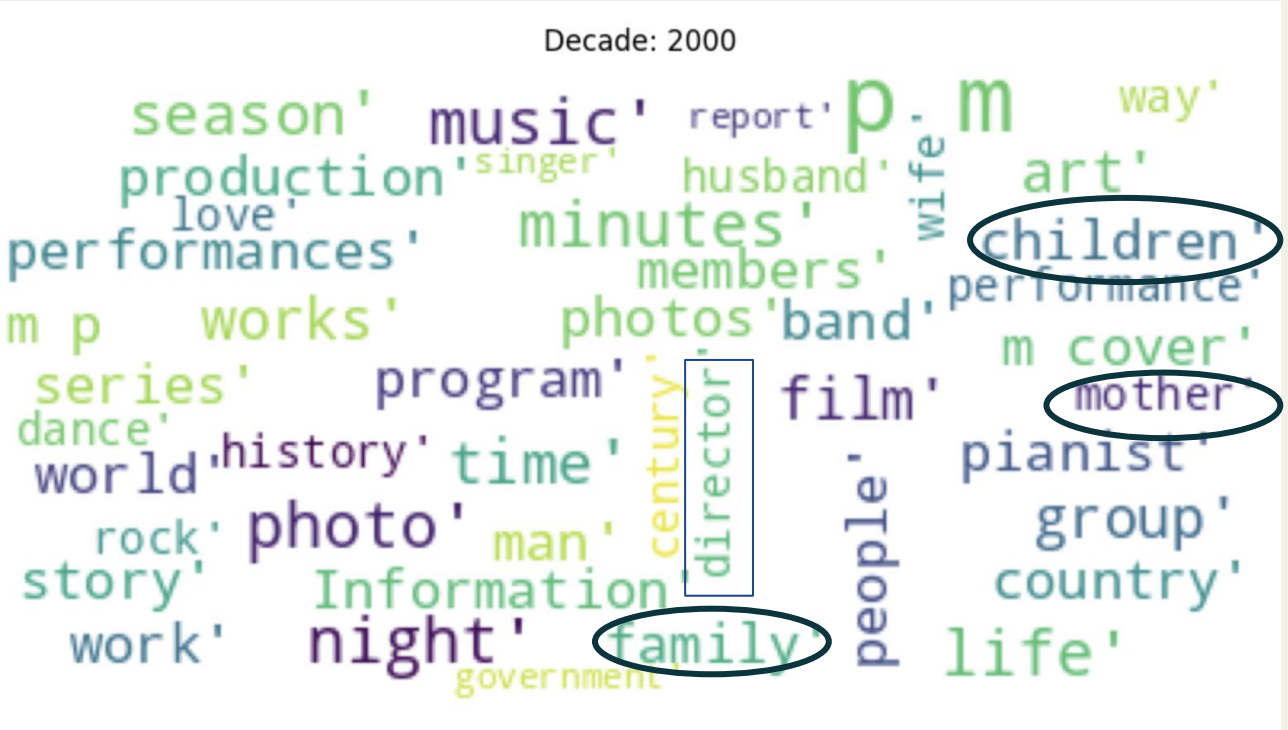
Word Clouds: 1980s



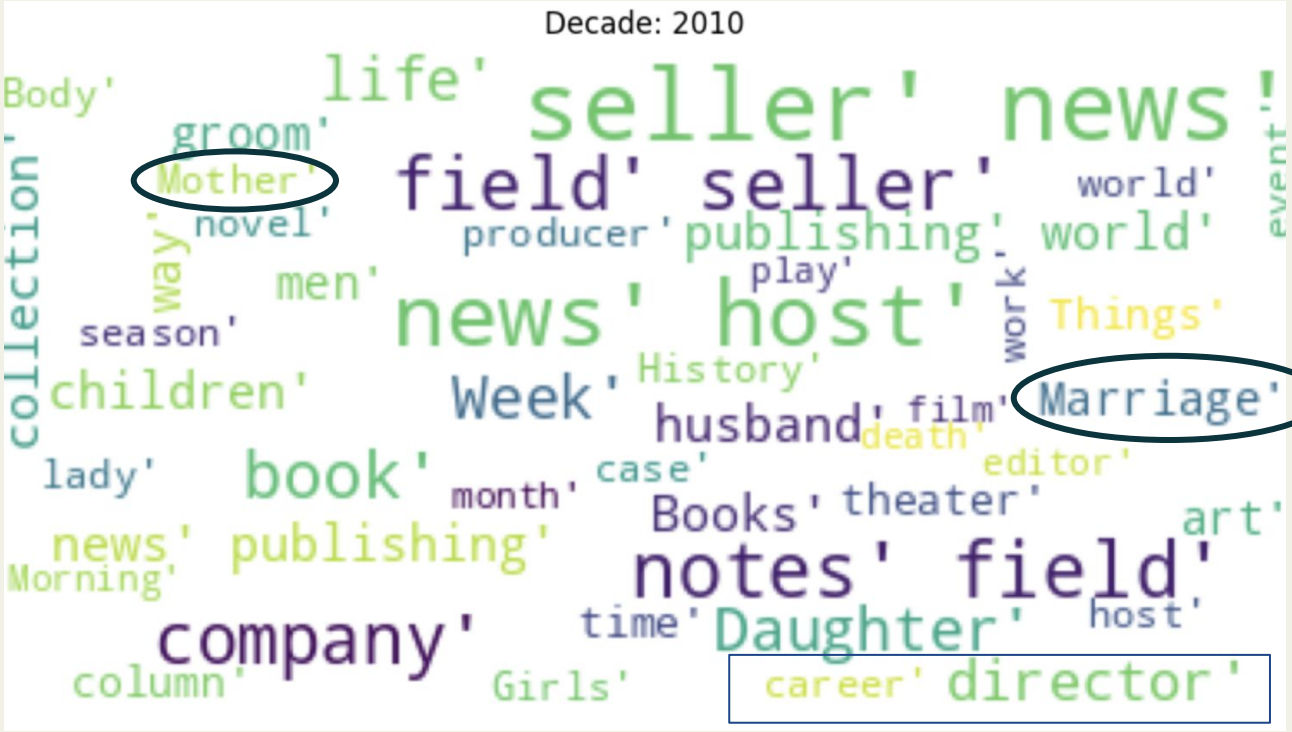
Word Clouds: 1990s



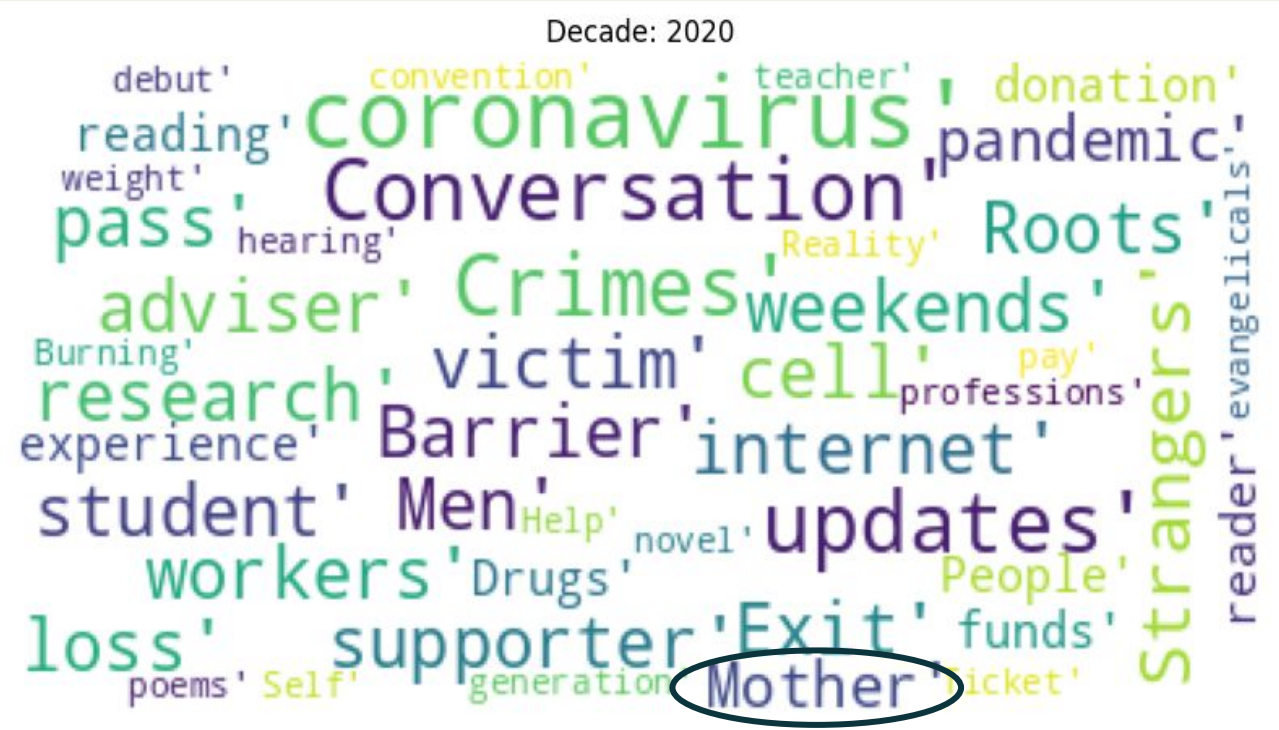
Word Clouds: 2000s



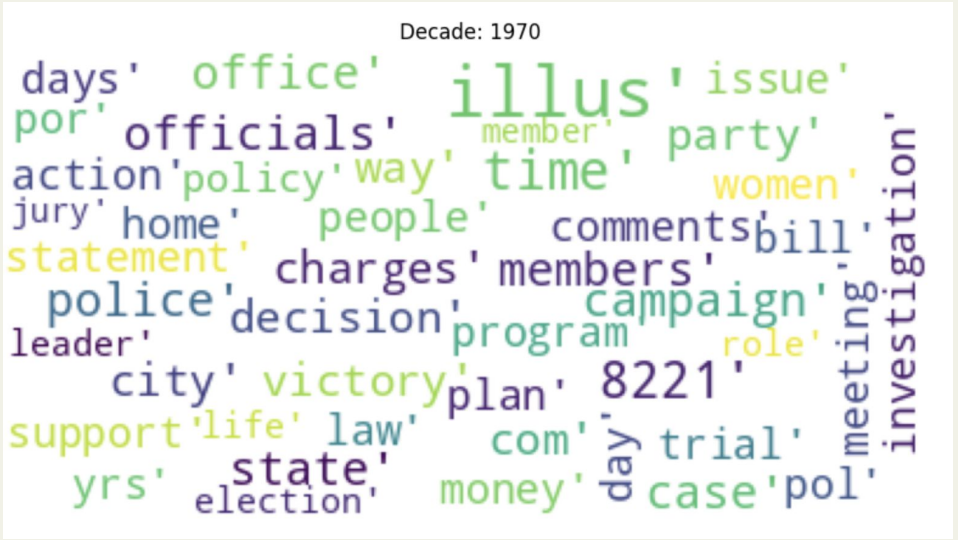
Word Clouds: 2010s



Word Clouds: 2020



Word Clouds: Men



Key takeaways

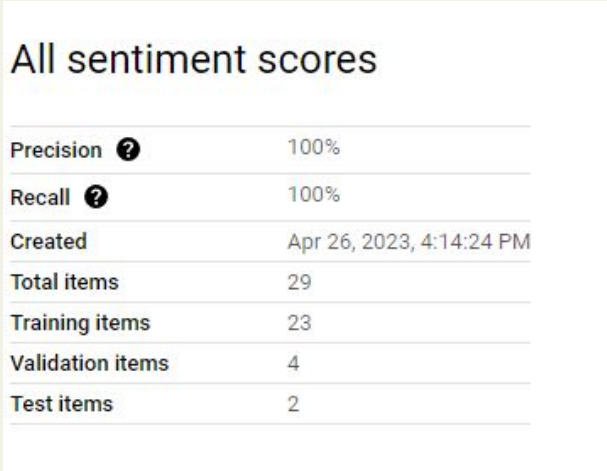
- For women: most of the twentieth century is family oriented
- After 1980: women are mentioned more often
  → but still very family-oriented
- After 2000: women are mentioned more often with more diverse topics
- Men: no clear themes in topics regardless of the decade

Sentiment Analysis

NLP with Vertex AI

Ran a sentiment analysis model but we did not get very far…

- Only accepted text file inputs
  - Max 500 text files
- Model ran for 6 hours & cost ~$25
  - Not time or cost effective with the size of our dataset
- Overall not user friendly or useful for our objectives



Sentiment Analysis with Spacy

Used built in functions to assess polarity & subjectivity

- Made ‘sentiment’ column
  - Very positive
  - Positive
  - Neutral
  - Negative
  - Very Negative
- Made a new column ‘intensity’ which is the absolute value of polarity

Polarity: score -1 - 1 based on how positive or negative a statement is

Subjectivity: score 0-1 based on how “subjective” or “objective” a statement is

Sentiment Analysis with Spacy

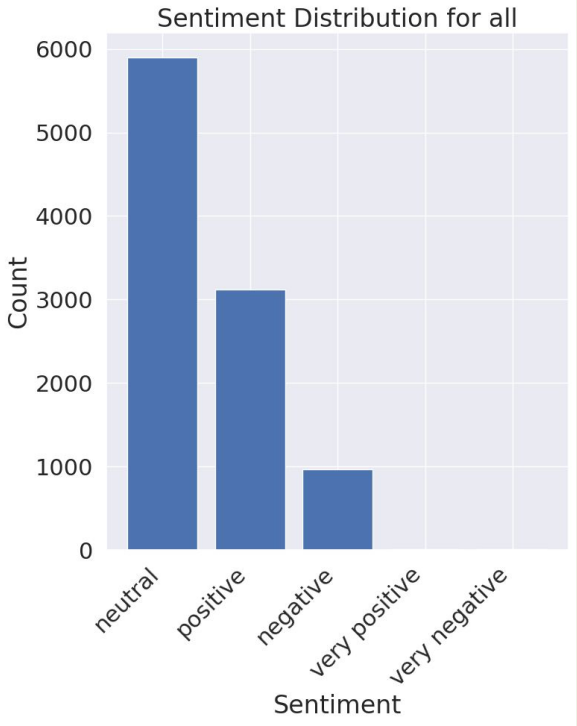
How is it determining polarity & subjectivity:

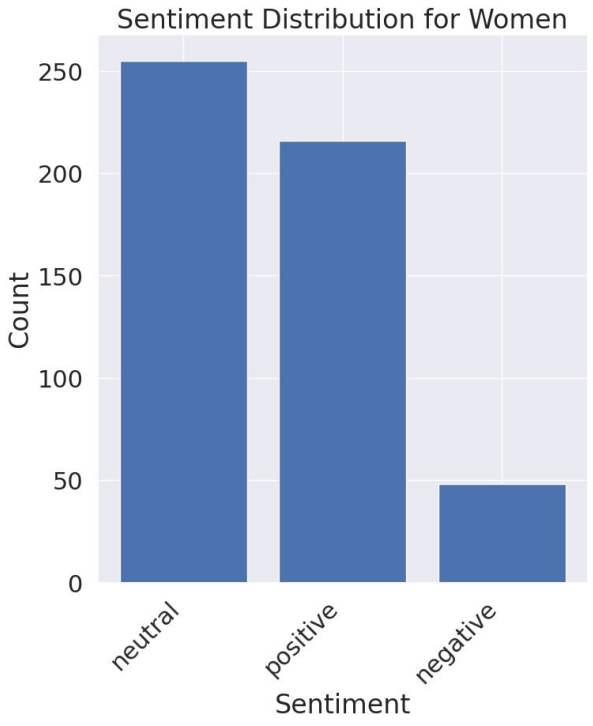
- Pulls words from the dataframe
- Scores each word internally
- Uses combined scores to compute overall sentence score

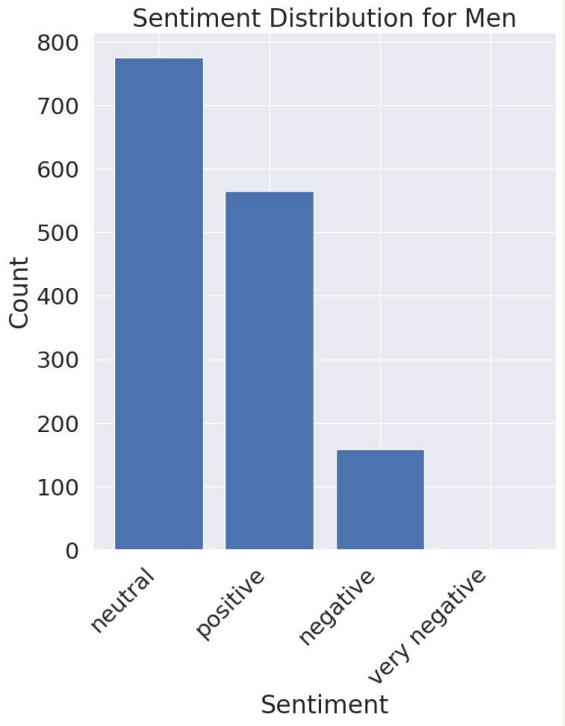
```
Row 5:
        Words: ['greater']
        Polarity: 0.5
        Subjectivity: 0.5

Row 5:
        Words: ['open']
        Polarity: 0.0
        Subjectivity: 0.5

Row 5:
        Words: ['classic']
        Polarity: 0.16666666666666666
        Subjectivity: 0.16666666666666666

Row 5:
        Words: ['third']
        Polarity: 0.0
        Subjectivity: 0.0
```

```
Row 5:
        Words: ['round']
        Polarity: -0.2
        Subjectivity: 0.4

Row 5:
        Words: ['back']
        Polarity: 0.0
        Subjectivity: 0.0

Row 5:
        Words: ['new']
        Polarity: 0.13636363636363635
        Subjectivity: 0.45454545454545453
```

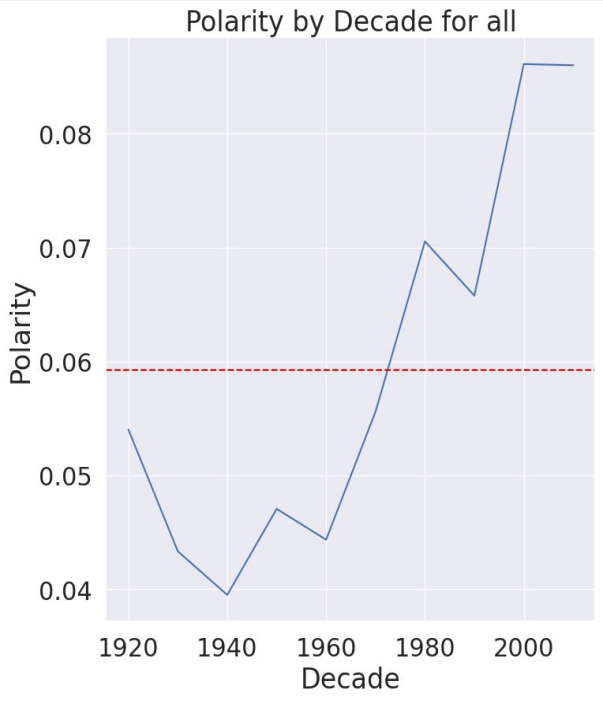
Sentiment Analysis: Sentiment Frequency

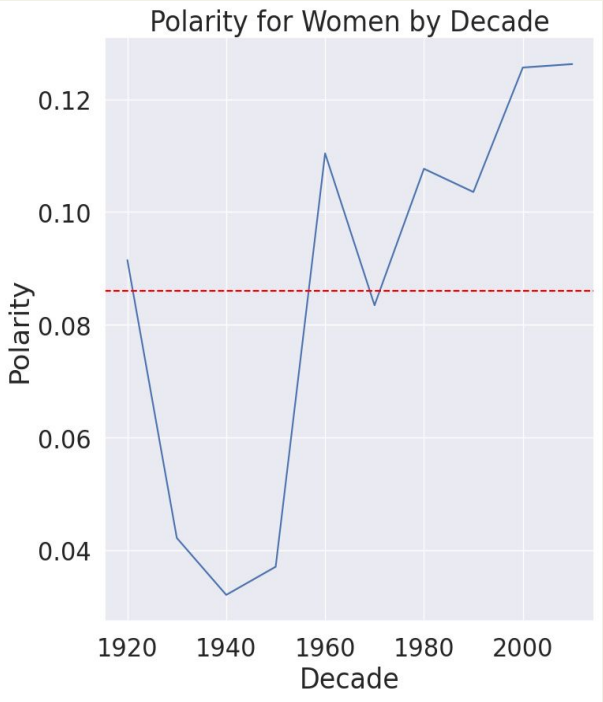


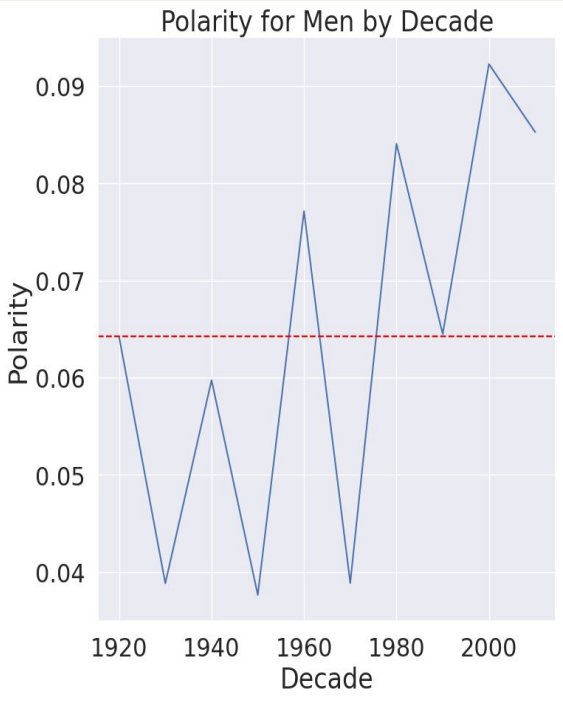




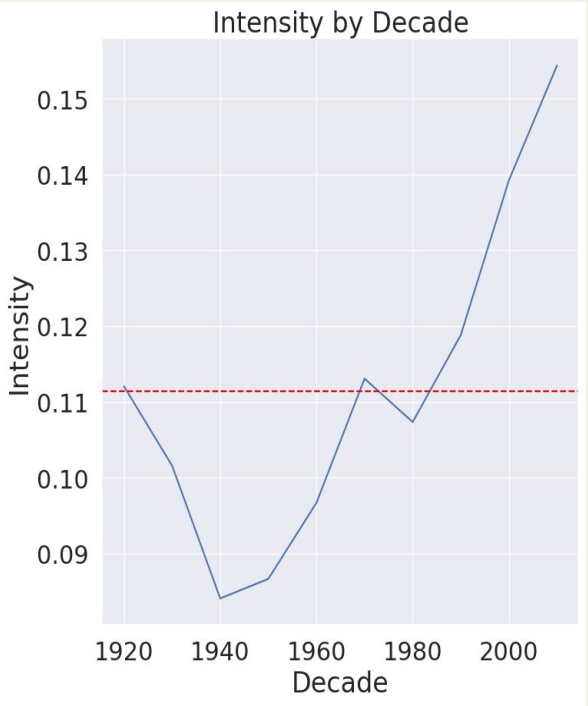
Sentiment Analysis: Polarity

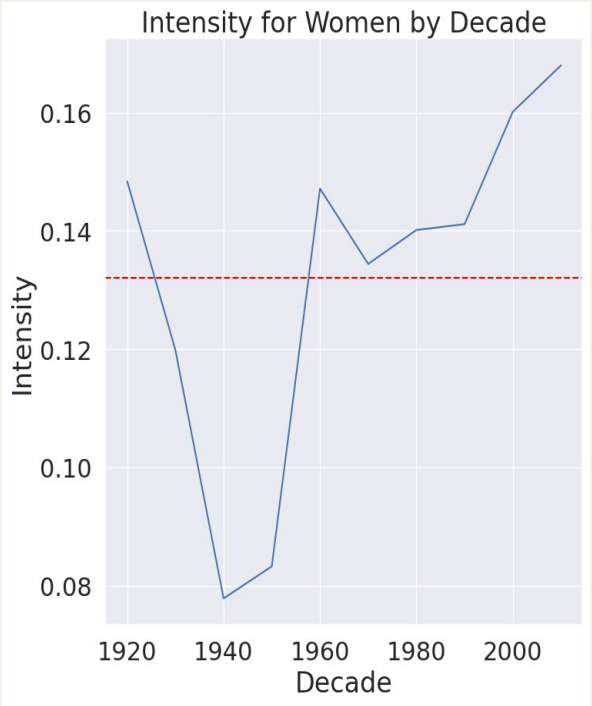


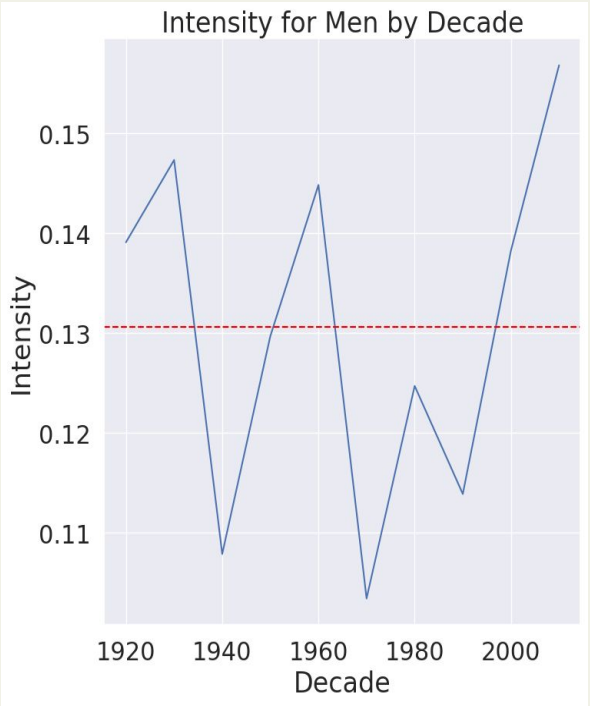




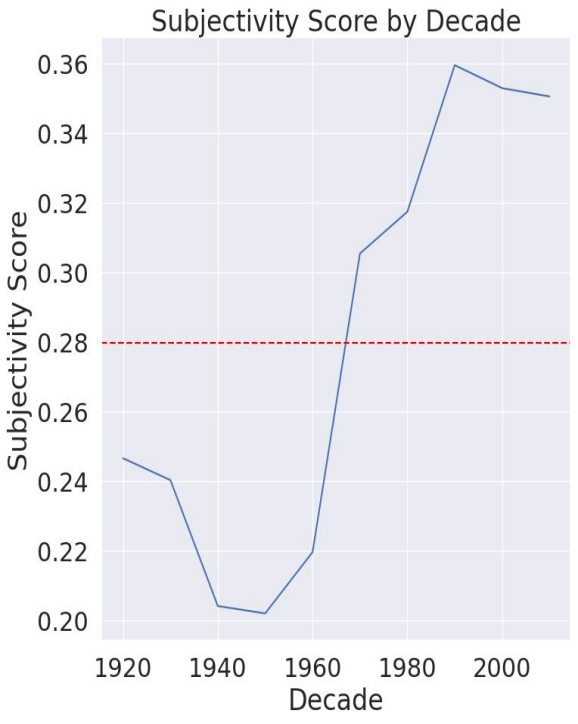
Sentiment Analysis: Intensity

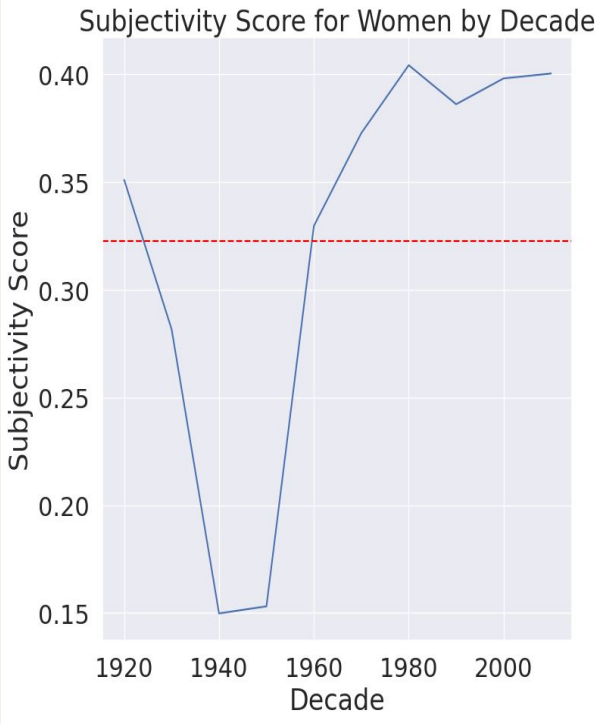


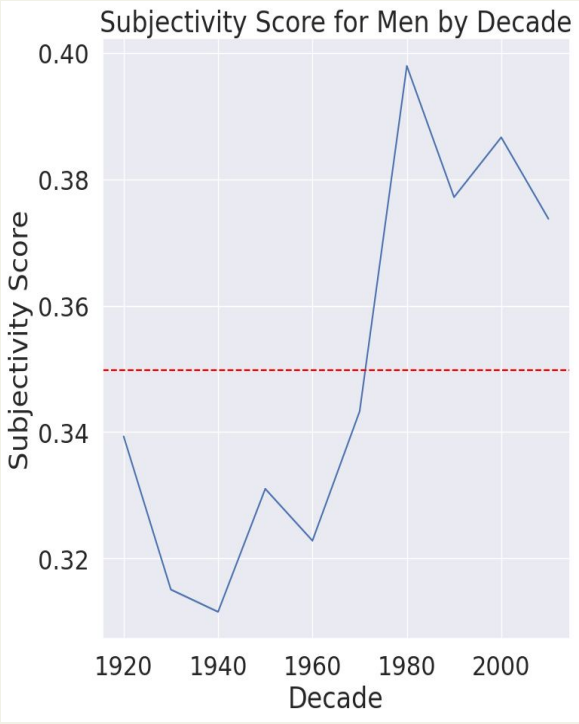




Sentiment Analysis: Subjectivity







Overall Takeaways

- Vertex AI seems awesome but can be challenging to use
- There are a variety of approaches and applications available for NLP
- Women are mentioned far less often than men
- Entities in the news vary a lot per decade
- Women are more often associated with familial topics
    - Marriage, children, family, etc.
    - Clear shift around the 2000s
- When women are in the news feelings are on average slightly more positive and less subjective

Thank You

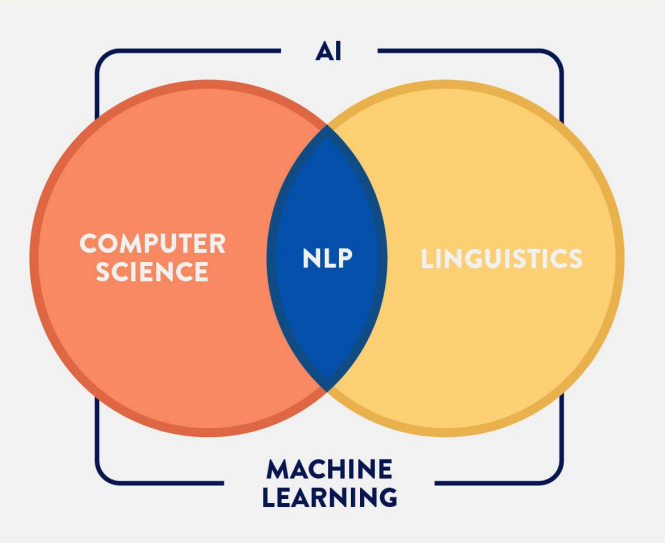
Questions?

Extra slides

What is NLP?

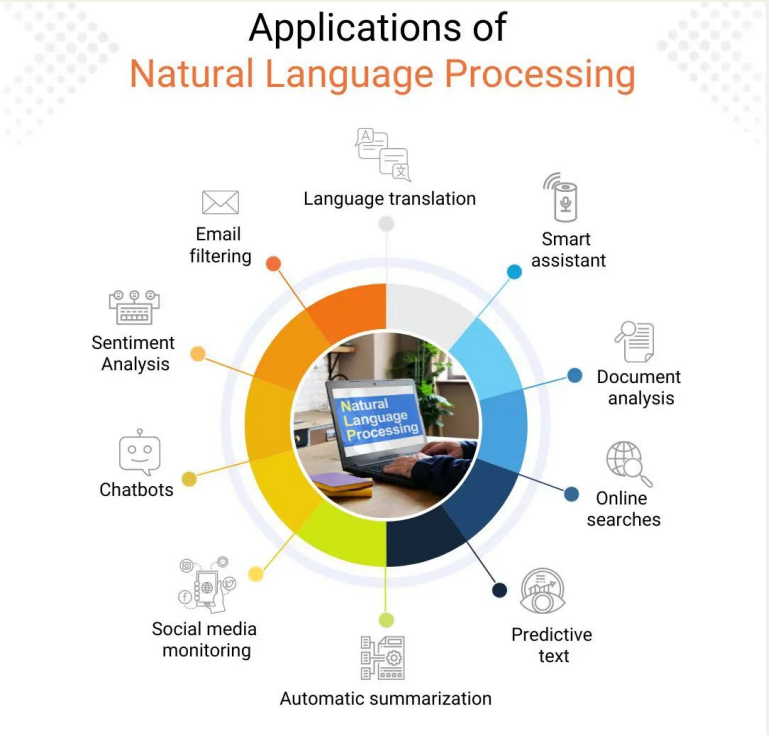
Natural Language Processing combines computer science and linguistics

- Allows computers to understand human language
- Subfield within Artificial Intelligence which uses machine learning



How is NLP used?

- Useful for any case where you need to process human language data:
  - Text analysis
  - Text summarization
  - Information search and retrieval
  - Language translation
  - Chatbots

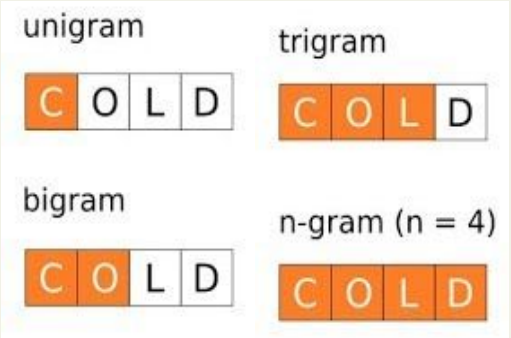


How does NLP work?

- Data Preprocessing: cleaning and prep for data analysis
  - Removing punctuation, capitalization, misspellings, dealing with special characters
  - **Make the text as simple as possible**
- Text Tokenization: breaking text into ‘tokens’
  - Word or sentence level, can also use custom tokens (eg split by commas)
- Feature Extraction: converting text to numerical representation
  - Bag of words, TF-IDF, word embeddings

How does NLP work?

- Analysis: using statistics and visualization for detecting text patterns
  - Word distribution & frequency, n-grams
  - Visualization using word clouds, histograms and heat maps
- Modeling: using machine learning algorithms and other models for tasks like text classification, sentiment analysis, etc.



Relevant Terms

- Name entity recognition
  - information extraction task which groups named entities into predefined categories
- Lemmatization
  - grouping together the inflected forms of a word into a single item
- Stemming
  - cutting the inflected words into the root form
- Part-of-speech tagging
  - marking part of speech based on definition and context
- Parsing
  - Analyzing grammatical structure of text
- Morphological segmentation
  - dividing words into morphemes to understand word structure
- Word segmentation
  - Breaking text into words or tokens
- Sentence breaking
  - Breaking text into sentences

Exploratory Data Analysis (EDA)

- Working with a small subset first: one year, first 2,000 entries
- Questions:
  - For articles mentioning a person, can we classify them as mentioning a male and female?
  - What is the context in which women are mentioned (wordcloud)?

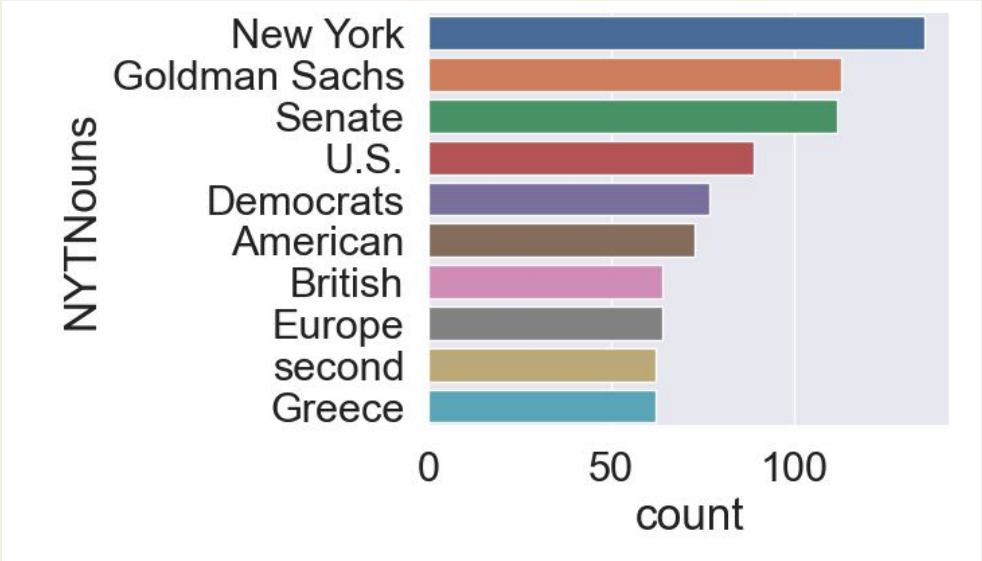
Goal: Get used to Spacy, see if we can do a simple analysis.

Exploratory Data Analysis (EDA)

1st step: Noun chunks

- Chunking is a process of extracting phrases from unstructured text

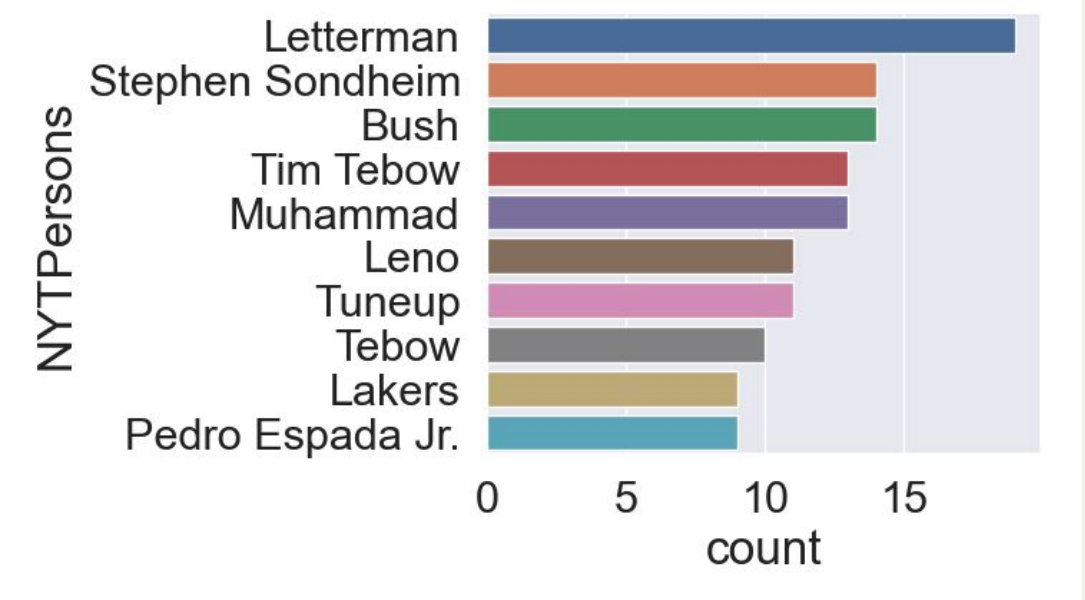
- Extracted nouns from article titles from the first 2,000 articles in 2010



Exploratory Data Analysis (EDA)

2nd step: exploring the Person entity

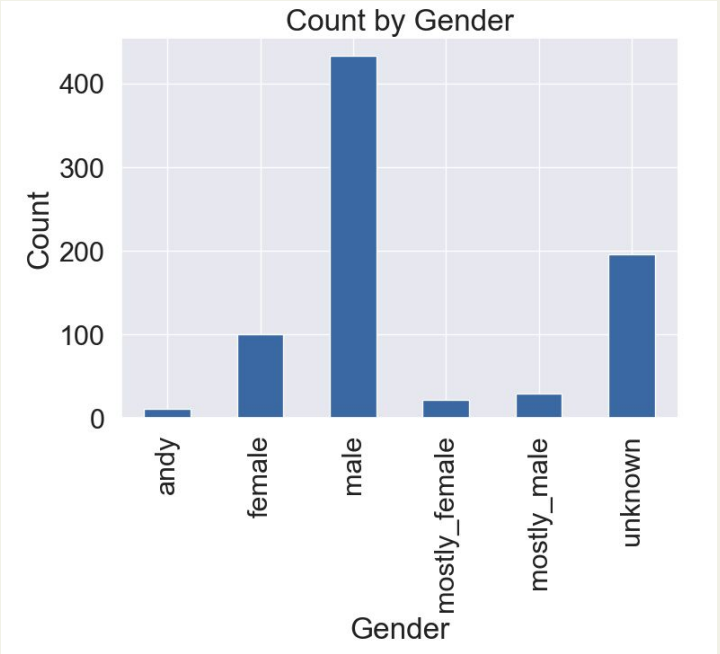
- Spacy can recognize names of people, including fictional
- Not always accurate
- Duplicates



Exploratory Data Analysis (EDA)

3rd step: Using Gender Guesser library

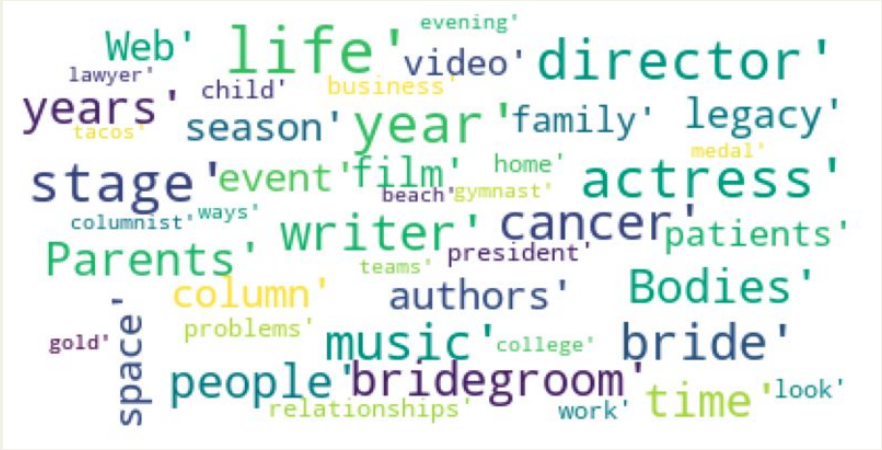
- Extracting first names
- Classifying names: andy (androgynous), unknown (not in the DB), female, male, mostly_male, mostly_female
- Then, isolate articles with female name



Exploratory Data Analysis (EDA)

4th step: Using Word Cloud

- Only ran on a very small df
- Not very insightful but scalable
- Goal: contextualize mentions of Female Names in a visual way



Next Steps

- Continued data analysis
  - Map how topics change over time
  - Visualize the frequency of women in the news over time
- Google Cloud
  - Data too big, gotta get to the cloud
  - Fine tune processing for whole dataset
- Model Selection & Testing
  - Text classification: binary classification for male vs female topics
  - Sentiment analysis: multi-class classification