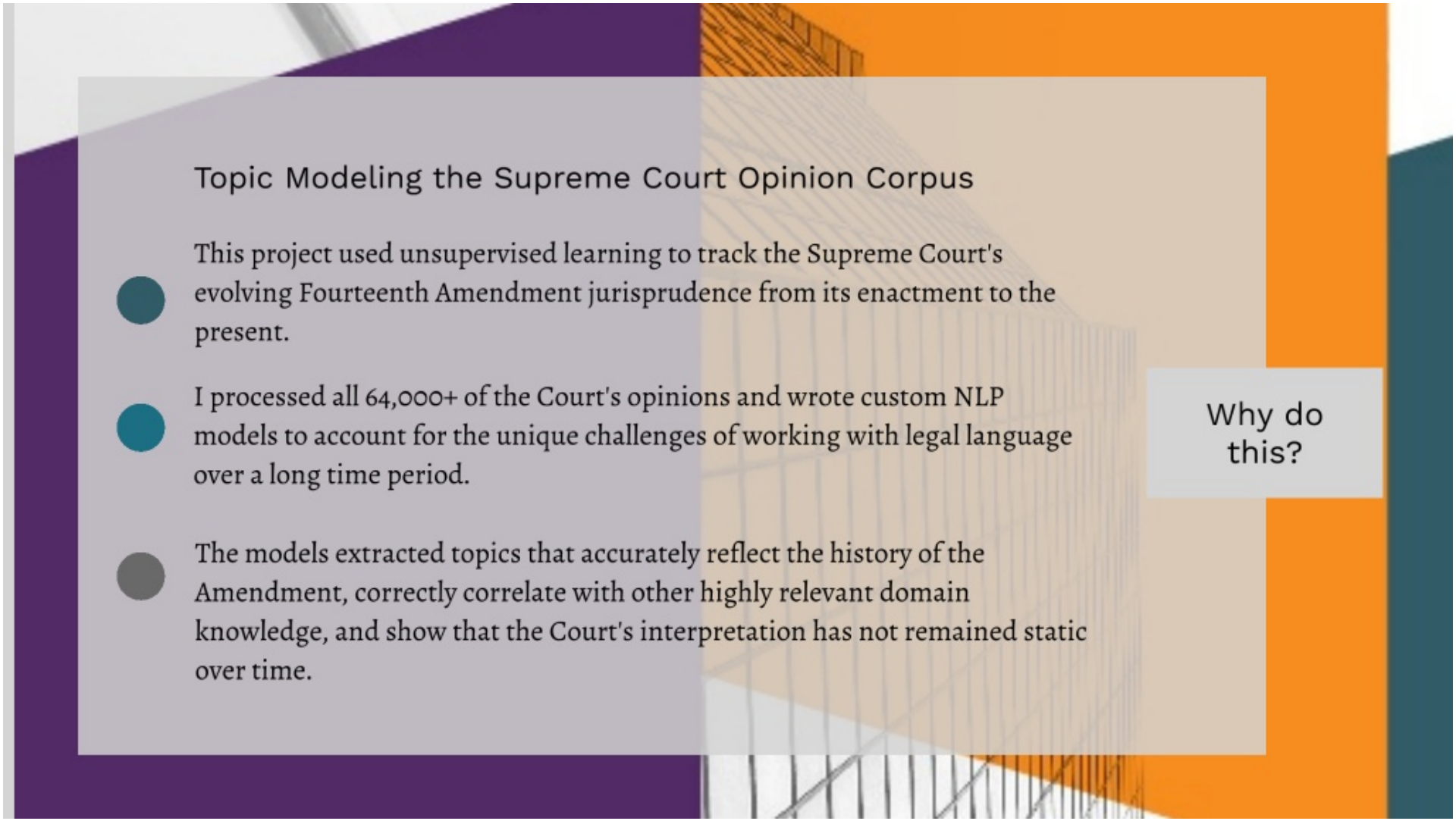
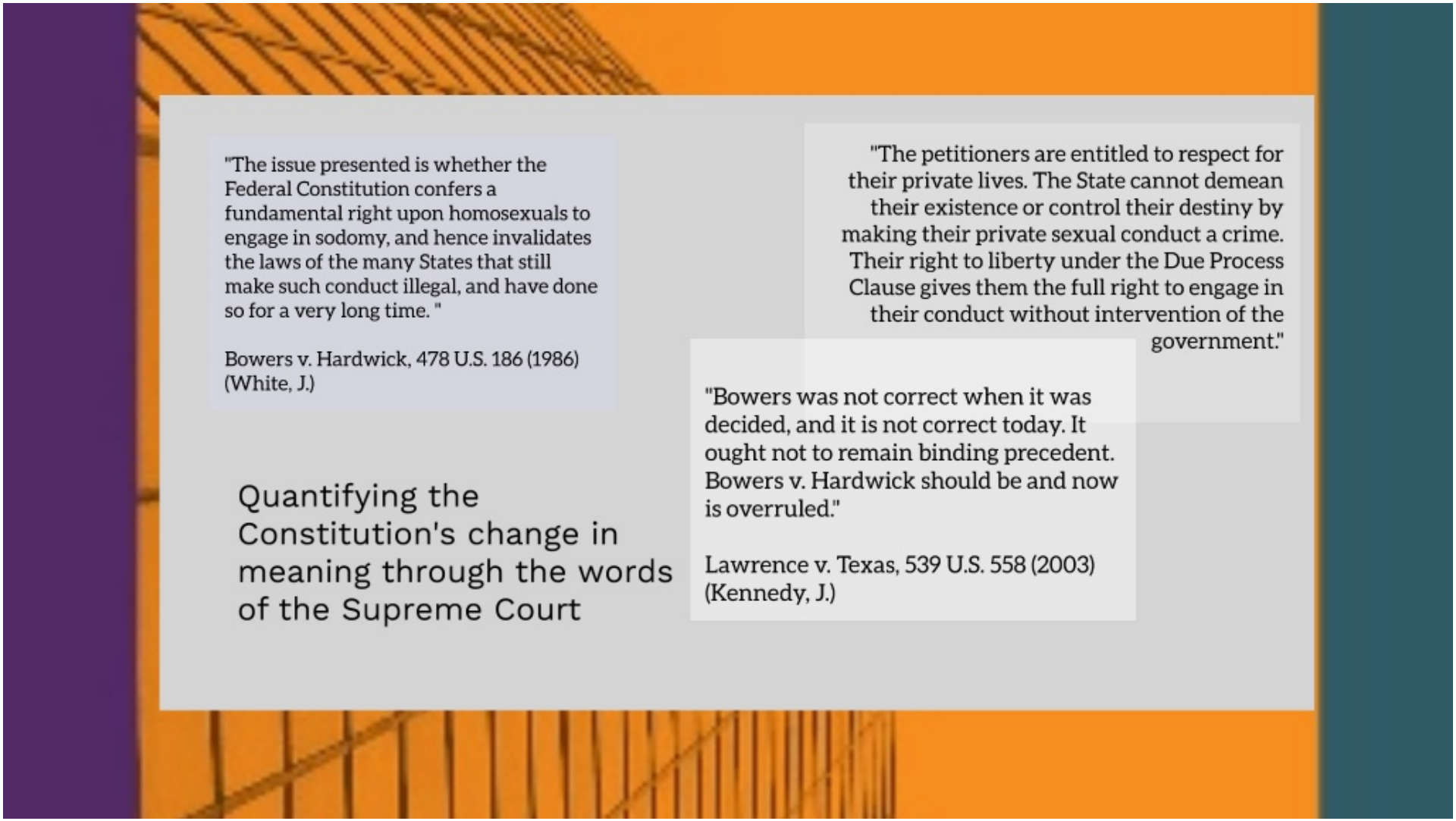
Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

LORA JOHNS

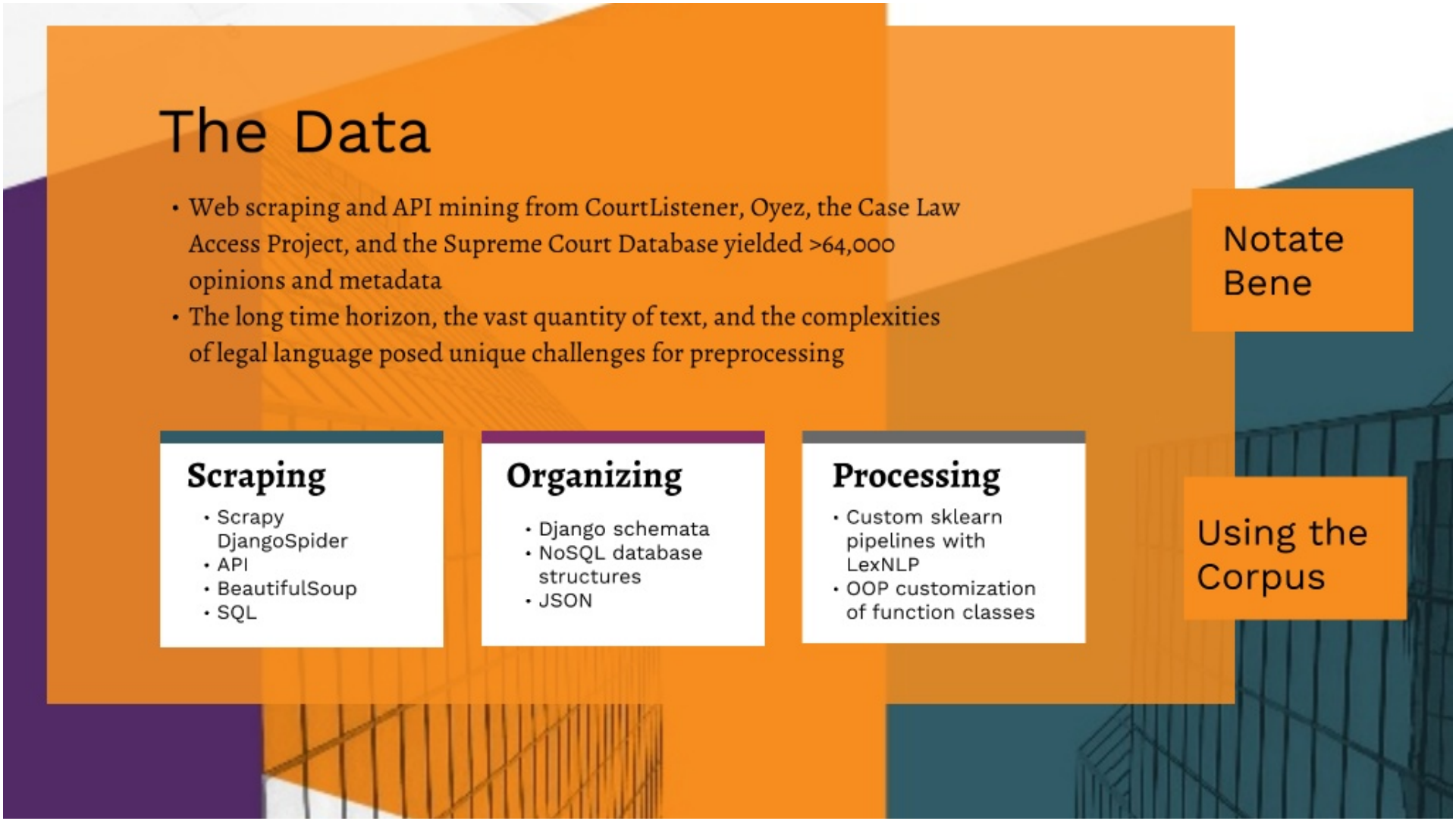

Topic Modeling the Supreme Court Opinion Corpus
This project used unsupervised learning to track the Supreme Court's evolving Fourteenth Amendment jurisprudence from its enactment to the present.
I processed all 64,000+ of the Court's opinions and wrote custom NLP models to account for the unique challenges of working with legal language over a long time period.
The models extracted topics that accurately reflect the history of the Amendment, correctly correlate with other highly relevant domain knowledge, and show that the Court's interpretation has not remained static over time.
Why do this?

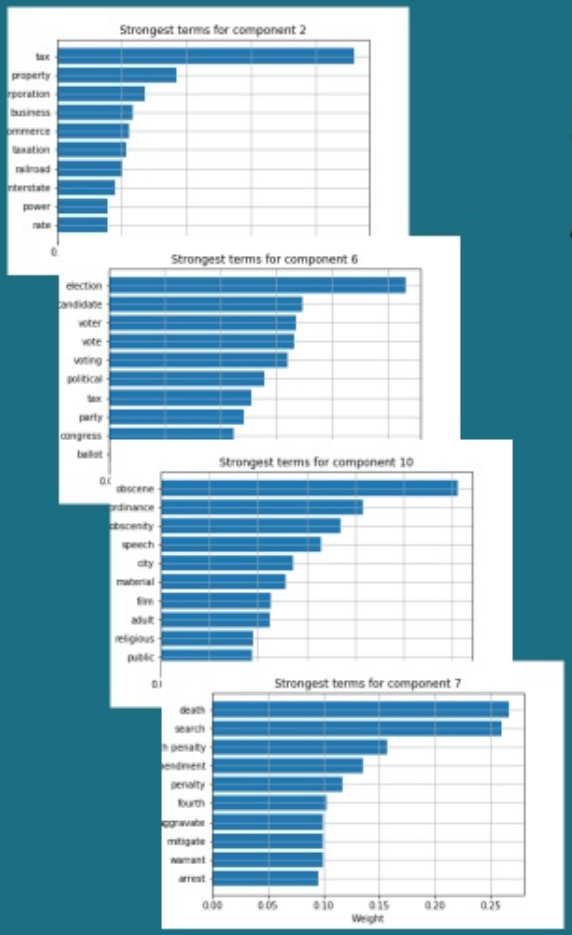

"The issue presented is whether the Federal Constitution confers a fundamental right upon homosexuals to engage in sodomy, and hence invalidates the laws of the many States that still make such conduct illegal, and have done so for a very long time. "
Bowers v. Hardwick, 478 U.S. 186 (1986) (White, J.)
"The petitioners are entitled to respect for their private lives. The State cannot demean their existence or control their destiny by making their private sexual conduct a crime. Their right to liberty under the Due Process Clause gives them the full right to engage in their conduct without intervention of the government."
"Bowers was not correct when it was decided, and it is not correct today. It ought not to remain binding precedent. Bowers v. Hardwick should be and now is overruled."
Lawrence v. Texas, 539 U.S. 558 (2003) (Kennedy, J.)
Quantifying the Constitution's change in meaning through the words of the Supreme Court

Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

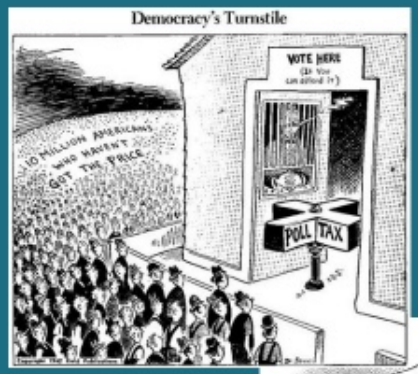
LORA JOHNS


The Data
• Web scraping and API mining from CourtListener, Oyez, the Case Law Access Project, and the Supreme Court Database yielded >64,000 opinions and metadata
• The long time horizon, the vast quantity of text, and the complexities of legal language posed unique challenges for preprocessing
Scraping
• Scrapy DjangoSpider
• API
• BeautifulSoup
• SQL
Organizing
• Django schemata
• NoSQL database structures
• JSON
Processing
• Custom sklearn pipelines with LexNLP
• OOP customization of function classes
Notate Bene
Using the Corpus



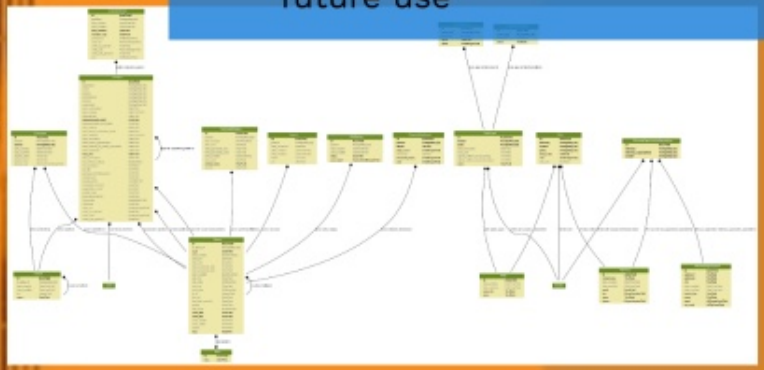
Latent Semantic Analysis

Unsupervised learning using LSA on the corpus as a whole produced coherent topics.

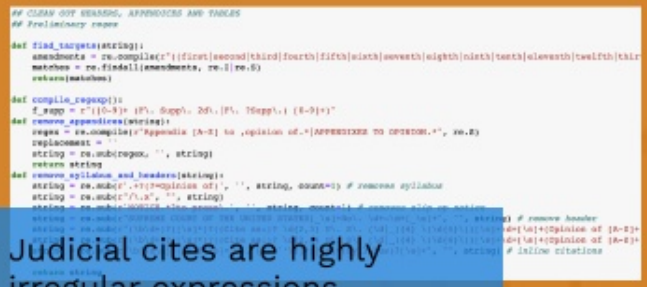


Data Structures

- Original underlying relational structure was poorly designed, resulting in null keys and corrupted API output
- Corrected, connected, and joined 128,000+ such files for future use



Legal Citations



- Judicial cites are highly irregular expressions
- > 200 years of changing citation formats means no easy rule
- Wrote a tailored script to remove "stop words" without over-excluding important information

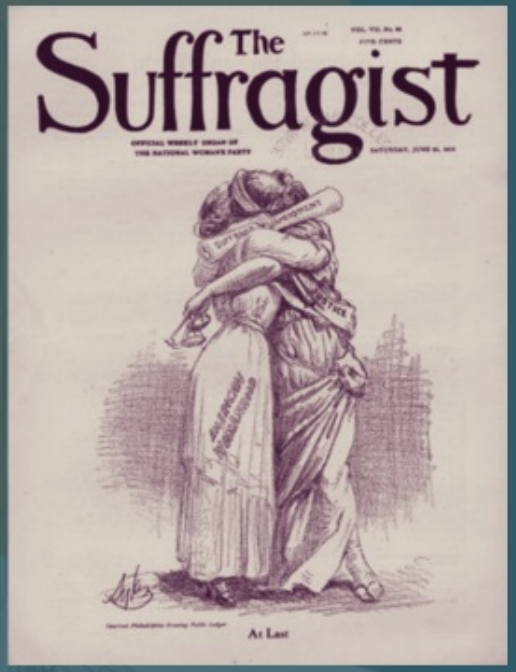
Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

LORA JOHNS

Methods of analysis

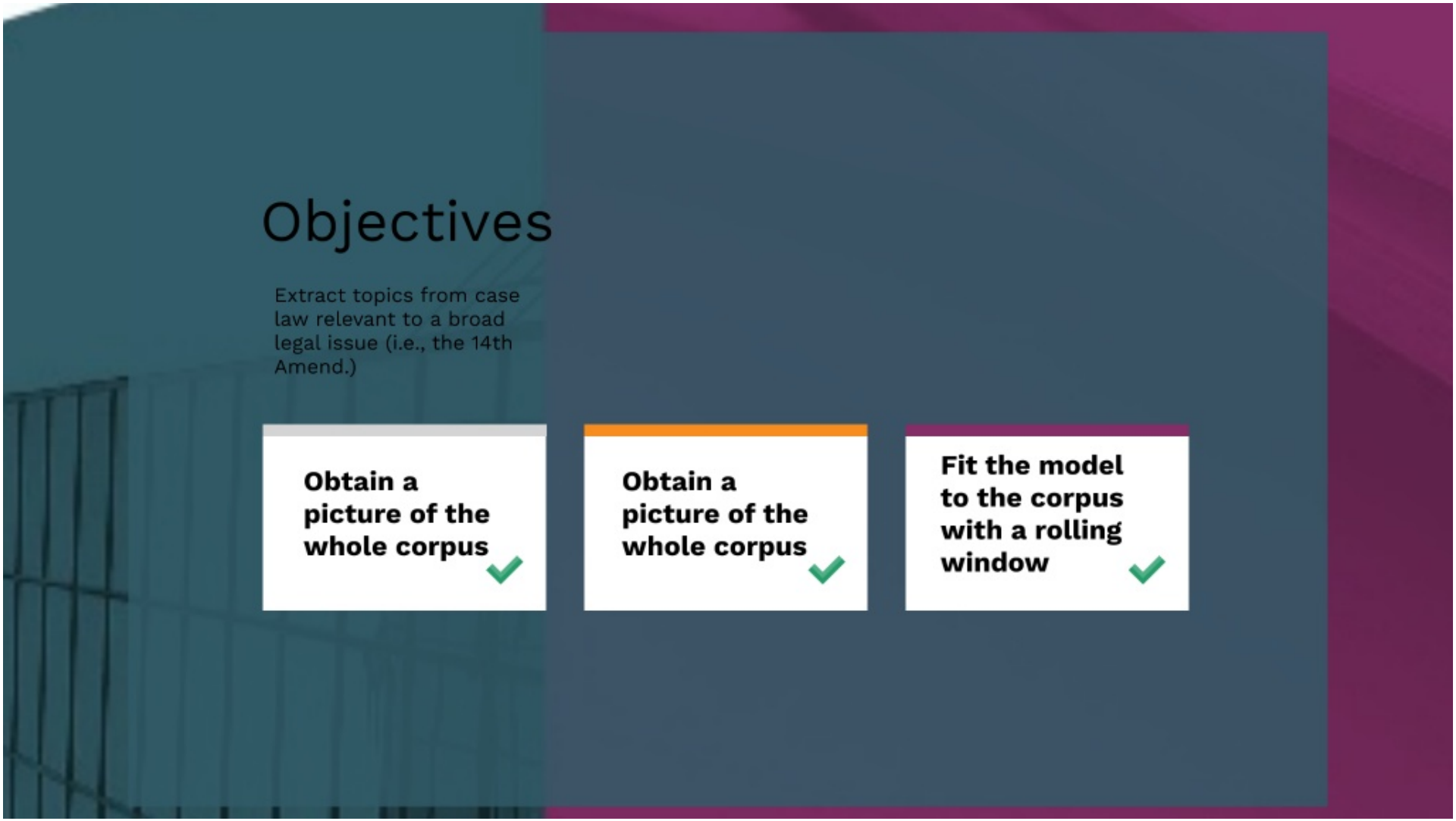
Objectives

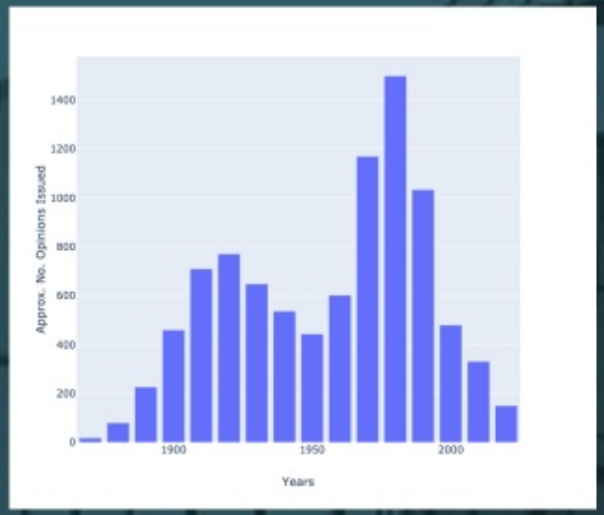


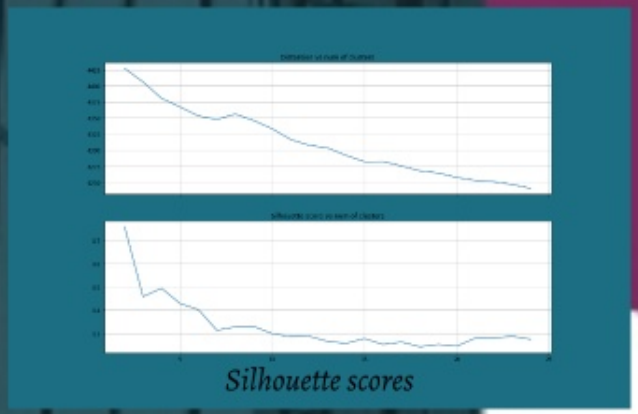
Tools and libraries:

- Scikit-learn, numpy, pandas
- Matplotlib, plotly, Plotly Express, OpenCV
- SpaCy, NLTK
- Django, Scrapy, REST API
- SQL, MongoDB, neo4j

Topic Modeling


Objectives
Extract topics from case law relevant to a broad legal issue (i.e., the 14th Amend.)
Obtain a picture of the whole corpus
Obtain a picture of the whole corpus
Fit the model to the corpus with a rolling window





Topic extraction workflow

- Make TF-IDF matrix, bag of words, and extract features
- Perform LSA and singular value decomposition
- K-means test with silhouette method
- Create rolling window function class
- Evaluate and visualize results
- Revise and apply to additional document classes
- Expand and refine methods

Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

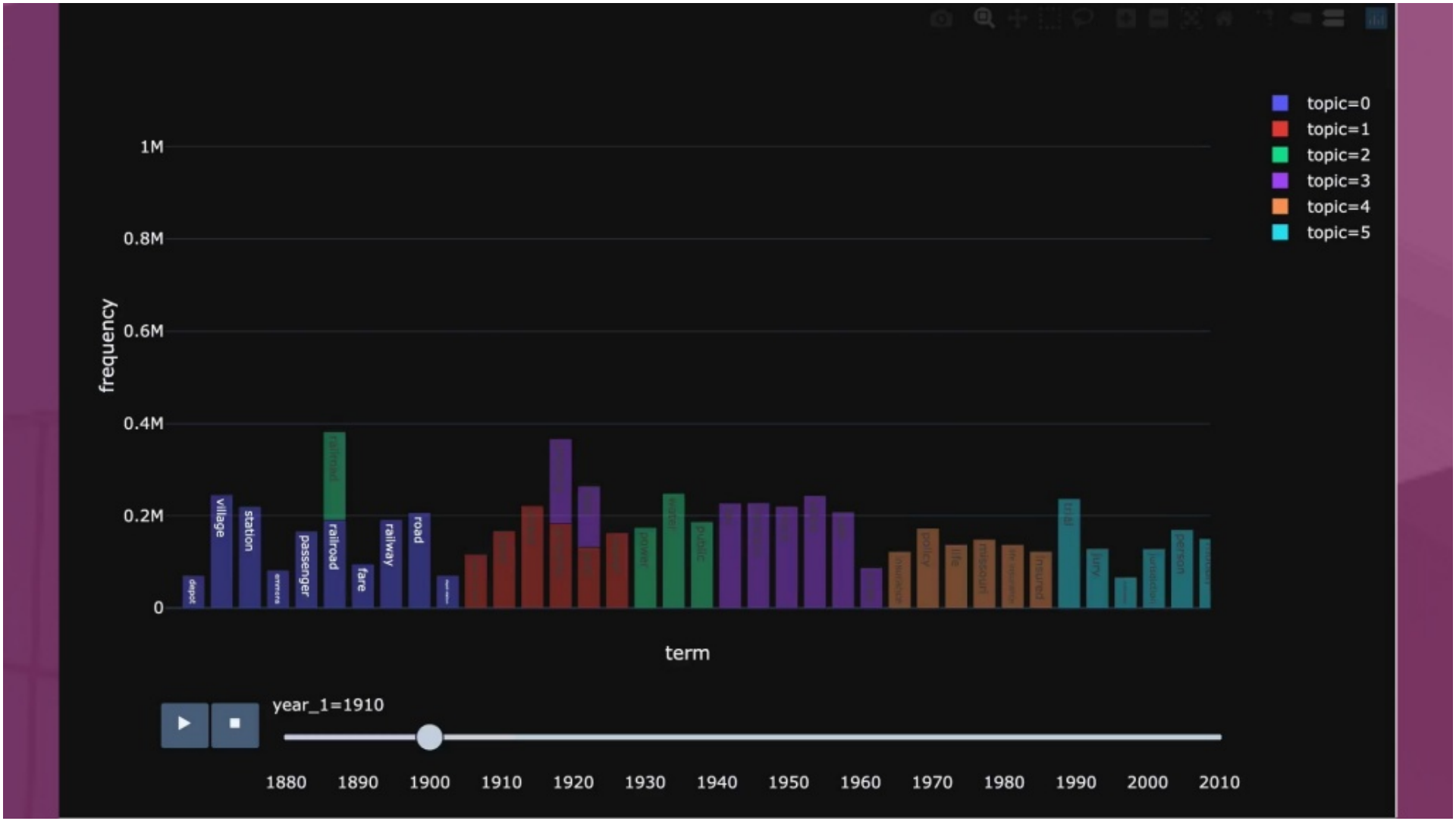
LORA JOHNS

Evolving 14th Amendment Jurisprudence

Visualized through k-means clusters

The 14th Amendment

Distinct Eras

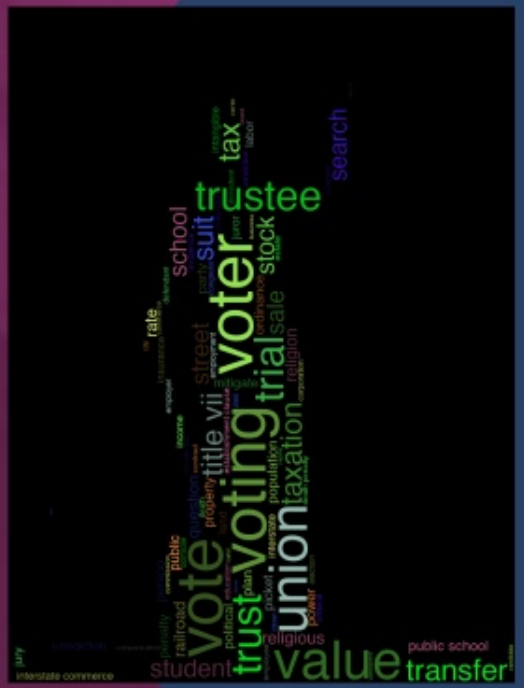

topic=0
topic=1
topic=2
topic=3
topic=4
topic=5
1M
0.8M
0.6M
0.4M
0.2M
0
frequency
village
station
passenger
railroad
fare
railway
road
power
water
public
policy
life
mission
injured
trial
jury
person
term
year_1=1910
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010

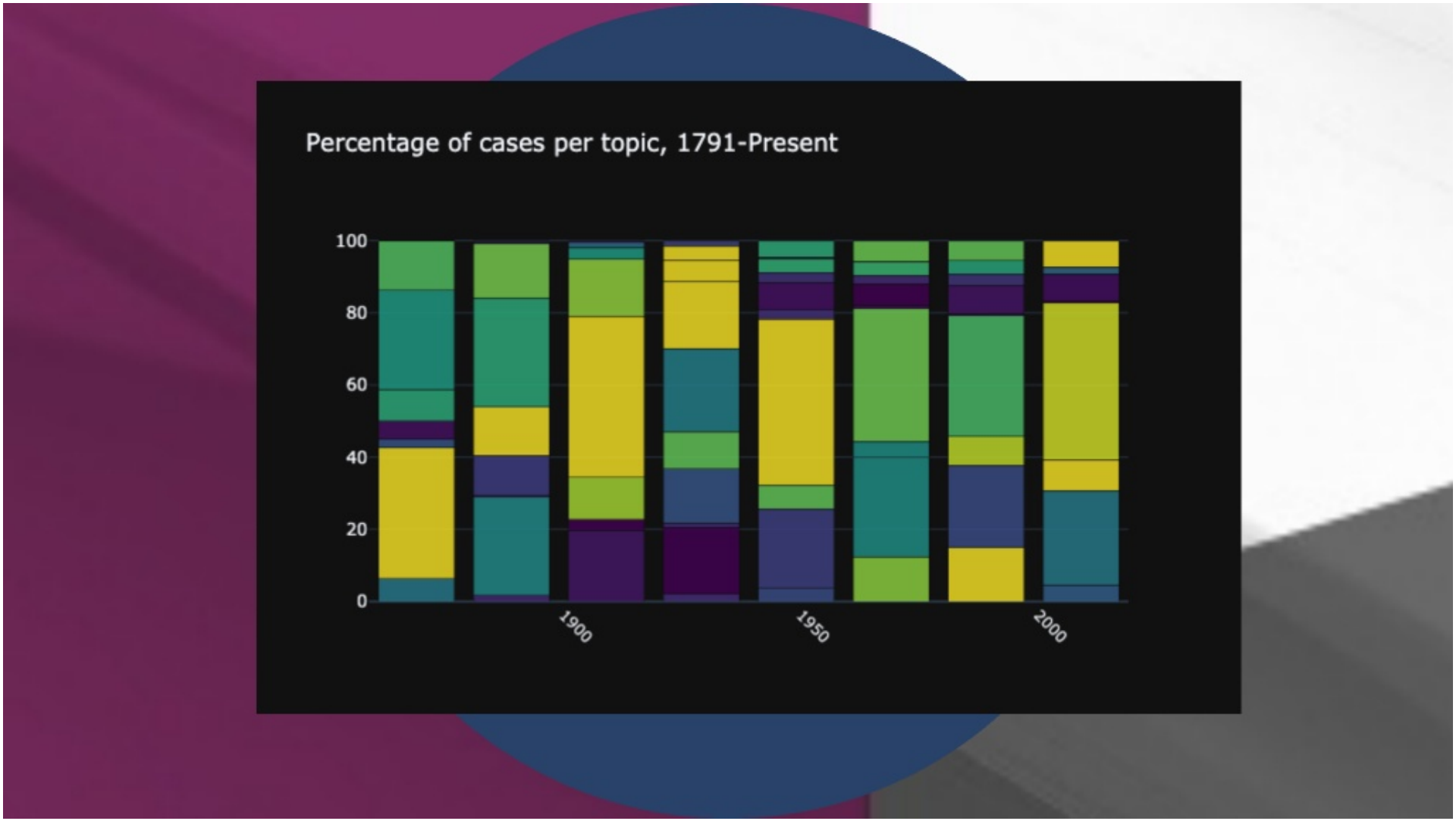
Topics: The Corpus as a Whole

- Mini-batch k-means pipeline, using TF-IDF, with n-grams up to 2 words and 10 clusters
- Metrics: elbow plot, silhouette score
- Additional hyperparameter tuning with sklearn
- Additionally, I included the size of the time window and span of overlap in the training pipeline

Top Terms over Time

Top Topics over Time


vote
voting
voter
union
trustee
trial
value
transfer
trust
taxation
search
tax
school
suit
stock
sale
street
title vii
student
public school
religious
railroad
rate
public
interstate commerce
jury


Percentage of cases per topic, 1791-Present
100
80
60
40
20
0
1900
1950
2000


Topics through American History

Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

LORA JOHNS

Next Steps

- Expand the analysis to more documents and corpora
- Incorporate additional analyses and refine those already used
- Build a graph database to contain the vast stores of information acquired
- Use the semantic analysis uncovered to build more relevant legal research tools

Counting and Apportioning Legal History with Natural Language Processing

2019/08/01

LORA JOHNS