The State of the Union Is...

An NLP and Unsupervised Learning Project

Stephen DeFerrari

Motivation

According to the Constitution, the president must periodically *"give to the Congress Information of the State of the Union, and recommend to their Consideration such measures as he shall judge necessary and expedient."*

The president's annual State of the Union (SOTU) is the president's fulfillment of this and typically includes a budget and economic report as well as the president's **national priorities** for that year.

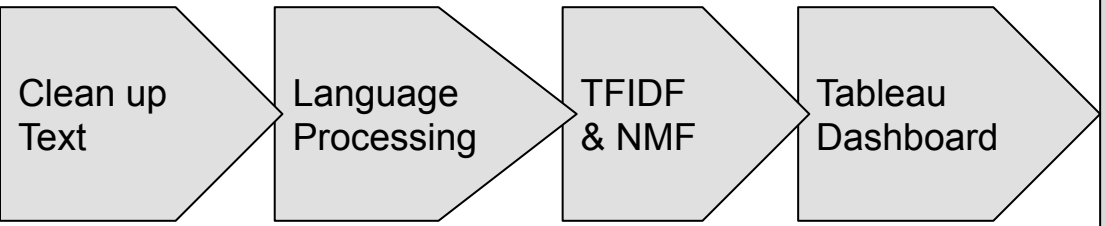
The U.S. is over 200 years old, how have these national priorities shifted over time?

Objective

Using natural language processing (NLP) and unsupervised learning, I want to find the topics that most stand out as "national priorities" through the years in SOTUs.

The ultimate goal will be to turn this into a Tableau dashboard which is interactive and easily understandable so that the average citizen can be better informed about their nation.

Methodology



Data:

- State of the Unions from 1791 until 2018
- 227 speeches, 5k length avg

Processing:

- NLTK's stemming followed by TFIDF

Unsupervised Learning:

- NMF

Data Processing and Text Clean Up

- Differentiating between same last names
- Assigning parties
- Cleaning up speeches
  - Numbers, punctuation
  - all lowercase, etc

Processing - To Stem or not to Stem

Deciding between NLTK's SnowballStemmer and WordNetLemmatizer was a tough decision to make.

Feature Name Lengths:

Stemming: 1723

Lemmatization: 2604

Ultimately, Stemming was picked.

Lemmatizer:

terrorism becomes **terrorism**

terror becomes **terror**

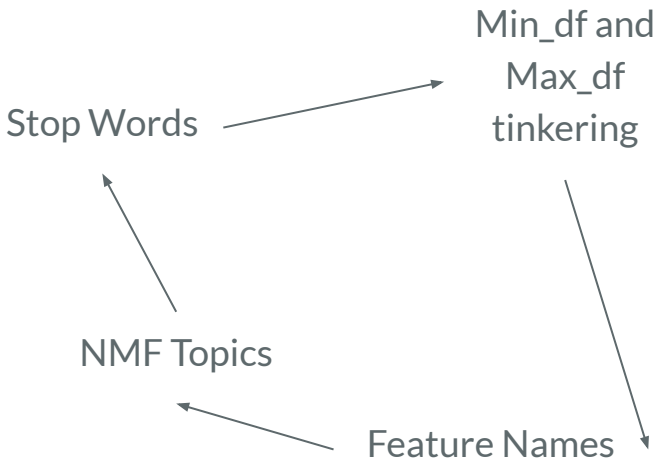
Stemmer:

terrorism becomes **terror**

terror becomes **terror**

Vectorization and Learning

- Repeated the cycle of adding stop words and the minimums and maximums to see what clusters appeared
- Ended up with 25 Topics ranging from Terrorism to Communism



Topic Naming Examples

Taxation

Topic 12
tax, billion, spend, percent, budget, job, inflat, energi, cut, propos

Topic 13
soviet, world, free, defens, freedom, militari, econom, strength, threat, effort

Topic 14
spain, treati, articl, minist, majesti, territori, spanish, likewis, respect, appoint

Civil War

Topic 15
constitut, territori, union, slave, general, republ, duti, thus, elect, question

Topic 16
minist, british, communic, council, instruct, french, commerc, franc, author, relat

Topic 17
necessari, railway, action, seem, duti, ship, develop, immedi, industri, counsel

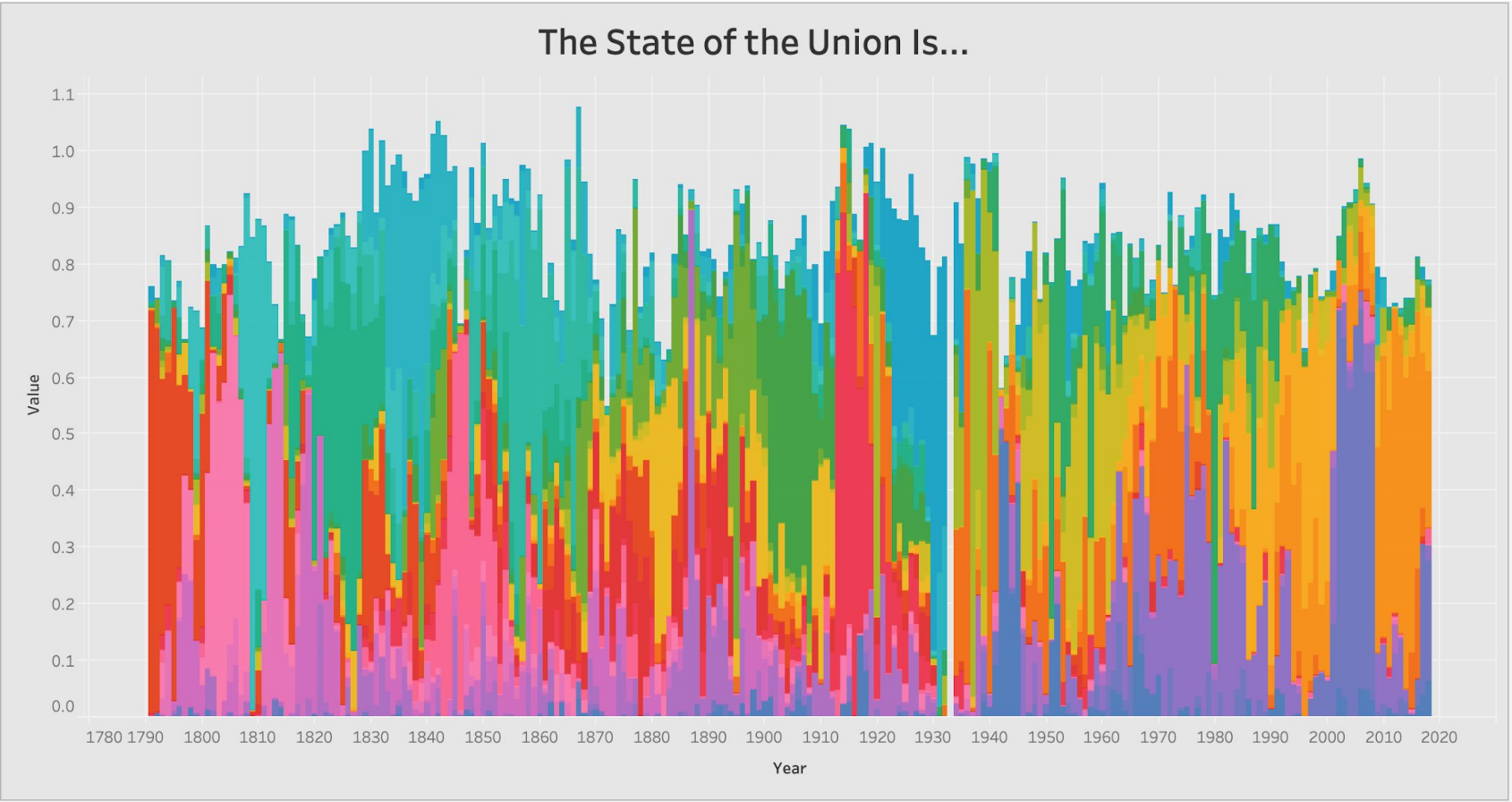
Topic 18
vessel, million, sea, necessari, debt, within, whether, princip, harbor, call

Terrorism

Topic 19
terror, weapon, world, freedom, tax, children, health, women, fight, worker

Topic 20
indian, provis, general, consider, measur, execut, happi, attent, militia, tribe

Topic Visualization!



Next Steps

- Cluster Presidents
- Better tune topics
- Correlate with government statistics (tough before mid 20th century)

Thank you!
Questions?