Jeffrey Levesque

jlevesqu@syr.edu | Topic Modeling

IST-736: HW8

PRofessor gates

A drawing of a cartoon character

Description automatically generated

Table of Contents

[Introduction 2](#_Toc10550916)

[Data Preparation 2](#_Toc10550917)

[Stop Words 3](#_Toc10550918)

[Ngram 4](#_Toc10550919)

[Results 4](#_Toc10550920)

[Latent Dirichlet Allocation 4](#_Toc10550921)

[Non-negative Matrix factorization 4](#_Toc10550922)

[Overall Sentiment 4](#_Toc10550923)

[Overall Topic Model 4](#_Toc10550924)

[Conclusions 4](#_Toc10550925)

# Introduction

The United States Congress, consisting of the Senate and the House of Representatives was enacted during the first two years of George Washington’s presidency[[1]](#footnote-1). While roles and responsibilities of the House of Representative have not changed, topics pervading society have grown and fluctuated through the years. Today, legislative terms vary from any number of 552 different subjects[[2]](#footnote-2), with topics generally varying non-inclusively from Defense, Education, Healthcare, and Economy.

Congressional Records are often detail oriented, providing transcripts of discussions and policy proposals made. However, these records are often not organized into a hierarchy of topics. The ability to explore topic trends, would enhance the understanding of Bill proposition and resolution.

# Data Preparation

Data was provided in a directory consisting of four subfolders. Specifically, each folder was prefixed by “110” (name of the dataset) and suffixed by two letters separated by a hyphen. The first letter represents the gender of the speaker, denoted by “m” for male, or “f” for female. Finally, the second letter represents the party affiliation, “d” for democrat, and “r” for republican. Thus, each text file represents a speaker of the House of Representative, associated with latter directory connotation. Furthermore, a total of 100M of text was distributed among the 400 text files.

Each file consists of numerous instances of XML style tags:

<DOC>

<DOCNO>Mrs. BOYDA of Kansas. (DEPARTMENT OF THE INTERIOR, ENVIRONMENT, AND RELATED AGENCIES APPROPRIATIONS ACT, 2008 -- (House of Representatives - June 26, 2007))</DOCNO>

<TEXT>

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

</TEXT>

</DOC>

**Figure 1:** entry from 110\_boyda\_x\_ks.txt, with contents of <TEXT> replaced with placeholder[[3]](#footnote-3).

Using the lxml package[[4]](#footnote-4), each file was wrapped into a common <data> xml tag, providing the ability to iterates nodes and children of the given structure. This provided the ability to restructure content into a pandas dataframe. Furthermore, a reproposed cleanse function was utilized to remove URLS, newlines, punctuation, and non-ascii characters[[5]](#footnote-5).

## Stop Words

A rich set of stop words was created[[6]](#footnote-6), utilized prior to vectorization:

* ser
* year
* today
* time
* thank
* make
* people
* going
* say
* like
* new
* need
* know
* chairman
* list
* act
* years
* house
* congress
* speaker
* madam
* colleagues
* gentleman
* let
* said
* way
* want
* support
* state
* yield
* time
* good
* right
* urge
* important
* come
* rise
* dont
* members
* help
* country
* american
* legislation
* president
* united
* states
* mr
* hr
* thats
* things
* day
* use
* ensure
* look
* got
* fact
* really
* thing
* percent
* trying
* tell
* im
* "im'",
* sure
* saying
* understand
* floor
* actually
* came
* says
* talking
* believe
* didnt
* congressman
* great
* lets
* heard
* rule
* amendment
* committee
* went
* ago
* conference
* section
* provision
* able
* yes
* republican
* republicans
* democrat
* democrats
* vote
* majority
* democratic
* north
* south
* bipartisan
* washington
* carolina
* debate
* question
* talk
* friend
* district
* issue
* certainly
* appreciate
* making
* middle
* including
* res
* program
* doesnt
* ought
* simply
* talked
* weve
* wanted
* order
* party
* place
* body
* coming
* kind
* point
* little
* lot
* aisle
* maybe
* ms
* seen
* hear
* getting
* theyre
* provide
* area

## 

## Ngram

A custom create\_ngram[[7]](#footnote-7) function was created, such that after preprocessing finished (cleansing, lowercase, stopwords), remaining words were combined into bigrams. Specifically, bigrams were concatenated with an underscore delimiting the two words. Thus, the TfidfVectorizer[[8]](#footnote-8) and CountVectorizer[[9]](#footnote-9) treats the associated input text as unigram.

# Results

## Latent Dirichlet Allocation

## Non-negative Matrix factorization

## Overall Sentiment

## Overall Topic Model

# Conclusions

1. <https://en.wikipedia.org/wiki/1st_United_States_Congress> [↑](#footnote-ref-1)
2. <https://www.congress.gov/help/field-values/legislative-subject-terms> [↑](#footnote-ref-2)
3. <https://github.com/jeff1evesque/ist-736-hw/blob/master/data/110/110-f-d/110_boyda_x_ks.txt> [↑](#footnote-ref-3)
4. <https://lxml.de/> [↑](#footnote-ref-4)
5. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw8/utility/dataframe.py> [↑](#footnote-ref-5)
6. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw8/utility/stopwords.py> [↑](#footnote-ref-6)
7. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw8/algorithm/topic_model.py> [↑](#footnote-ref-7)
8. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html> [↑](#footnote-ref-8)
9. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html> [↑](#footnote-ref-9)