Homework 8: Topic Modeling using Latent Dirichlet Allocation

Student: Patrick Walsh
Professor: Dr. Ami Gates
School: Syracuse University
Date: 3/19/2024

INTRODUCTION

In today's data-driven world, extracting meaningful insights from vast amounts of text data is crucial for informed decision-making. The "110" dataset, encompassing the floor debate transcripts of the 110th Congress (House only), presents a valuable opportunity to leverage advanced analytics techniques like topic modeling.

Topic modeling offers a powerful way to uncover hidden patterns and themes within large text datasets. For businesses, this translates into actionable intelligence that can drive various strategic initiatives such as:

- Policy Analysis: By identifying prevalent topics in congressional debates, policymakers can gain a comprehensive understanding of legislative priorities, enabling them to make informed decisions on policy formulation and amendment.

- Public Opinion Monitoring: Analyzing debates can provide insights into public sentiment and concerns, aiding businesses in understanding societal trends and aligning their strategies accordingly.

- Competitive Intelligence: Topic modeling can uncover key issues and discussions within specific political or demographic groups, offering businesses valuable insights into their competitors' priorities and strategies.

- Content Curation: Understanding the dominant topics in congressional debates helps in curating relevant content for stakeholders, customers, and the public, enhancing engagement and brand perception.

- Risk Management: Detecting emerging topics or controversial discussions allows businesses to proactively address potential risks and challenges, mitigating negative impacts on their operations.

ANALYSIS

Dataset

The dataset used in this analysis, referred to as "110," comprises text documents containing the floor debate of the 110th Congress, focusing on the House of Representatives. The dataset is organized into four subfolders, where "m" denotes male

speakers, "f" denotes female speakers, "d" denotes Democrats, and "r" denotes Republicans. The objective of this analysis is to gain insights into the topics discussed during the floor debates using natural language processing techniques.

```
Female Democrats:
['110_baldwin_x_wi.txt', '110_bean_x_il.txt', '110_berkley_x_nv.txt', '110_boyda_x_ks.txt', '110_brown-waite_x_fl.txt', '110_br
own_x_fl.txt', '110_capps_x_ca.txt', '110_castor_x_fl.txt', '110_clarke_x_ny.txt', '110_davis_x_ca.txt']
Female Republicans:
['110_bachmann_x_mn.txt', '110_biggert_x_il.txt', '110_blackburn_x_tn.txt', '110_bono_x_ca.txt', '110_capito_x_wv.txt', '110_cu
bin_x_wy.txt', '110_davis_jo-ann_va.txt', '110_drake_x_va.txt', '110_emerson_x_mo.txt', '110_fallin_x_ok.txt']
Male Democrats:
['110_abercrombie_x_hi.txt', '110_ackerman_x_ny.txt', '110_allen_x_me.txt', '110_altmire_x_pa.txt', '110_andrews_x_nj.txt', '11
0_arcuri_x_ny.txt', '110_baca_x_ca.txt', '110_baird_x_wa.txt', '110_barrow_x_ga.txt', '110_becerra_x_ca.txt']
Male Republicans:
['110_aderholt_x_al.txt', '110_akin_x_mo.txt', '110_alexander_x_la.txt', '110_bachus_x_al.txt', '110_baker_x_la.txt', '110_barr
ett_x_sc.txt', '110_bartlett_x_md.txt', '110_barton_x_tx.txt', '110_bilbray_x_ca.txt', '110_bishop_x_ut.txt']
```

The dataset can be found here:
https://drive.google.com/drive/folders/1hf3ALdntNIm6r-te2z8yAM26e33F0WFt

Data Preparation

The data preparation phase involved combining and cleaning the text data to create a document-term matrix (DTM) suitable for topic modeling. This process included removing irrelevant columns, special characters, and numbers, ensuring that the data is ready for analysis. The text data was vectorized using CountVectorizer() and converted to a dataframe.

Tokenize & Vectorize

```
:
```

```python
from sklearn.feature_extraction.text import CountVectorizer

# create CV objects
CV = CountVectorizer(input='filename', stop_words='english', encoding='latin-1')

# create document term matrix
dtm = CV.fit_transform(master_list)

print(dtm[0])
```

```
(0, 18407)    94
(0, 18419)    94
(0, 37175)    95
(0, 7443)     94
(0, 20182)    4
(0, 18111)    4
(0, 22696)    10
(0, 22013)    5
```

```python
import re

def clean_cols(df):
    # drop columns with numbers
    columns_to_drop = [col for col in df.columns if any(char.isdigit() for char in col)]
    df = df.drop(columns = columns_to_drop)

    # drop columns with special characters
    pattern = '^[a-zA-Z]+$'  # regex pattern to match only letters from a-z or A-Z
    columns_to_drop = [col for col in df.columns if not re.match(pattern, col)]
    df = df.drop(columns = columns_to_drop)

    # drop columns that have a length of less than 3 or more than 15
    columns_to_drop = [col for col in df.columns if len(col) < 3 or len(col) > 15]
    df = df.drop(columns = columns_to_drop)

    # drop columns that have three or more consecutive occurrences of the same letter
    pattern_consecutive = r'([a-zA-Z])\1\1'
    columns_to_drop = [col for col in df.columns if re.search(pattern_consecutive, col)]
    df = df.drop(columns=columns_to_drop)

    return df

cleaned_df = clean_cols(df)
display(cleaned_df)
```

	aachen	aap	aapg	aapi	aapis	aaron	aaronsen	aaronson	aarp	aarti	...	zurich	zury	zvi	zweig	zwilling	zwyer	zwyers	zydeco	zygote	zyuganov
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
424	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
425	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
426	0	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0	0
427	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
428	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

429 rows × 57065 columns

Models

The analysis utilized Latent Dirichlet Allocation (LDA) as the main modeling technique. LDA is a probabilistic topic modeling approach that identifies underlying topics within a collection of documents.

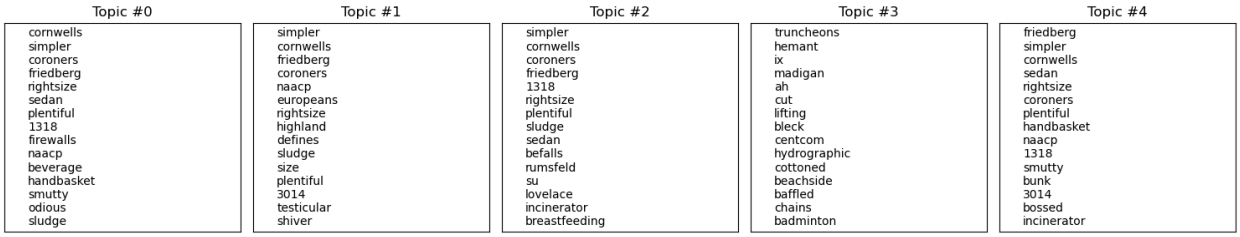
The LDA model was set to 100 max iterations and a learning method of ‘online’ rather than ‘batch’ so that words and documents were processed one at a time rather than in batches.

```
def run_LDA(df, num_topics, colnames):
    lda_model = LatentDirichletAllocation(n_components=num_topics, max_iter=100, learning_method='online')

    # Start the timer
    start_time = time.time()

    # Fit the model with progress bar
    with tqdm(total=len(df)) as pbar:
        lda_fitted = lda_model.fit_transform(df)
        pbar.update(len(df))
```

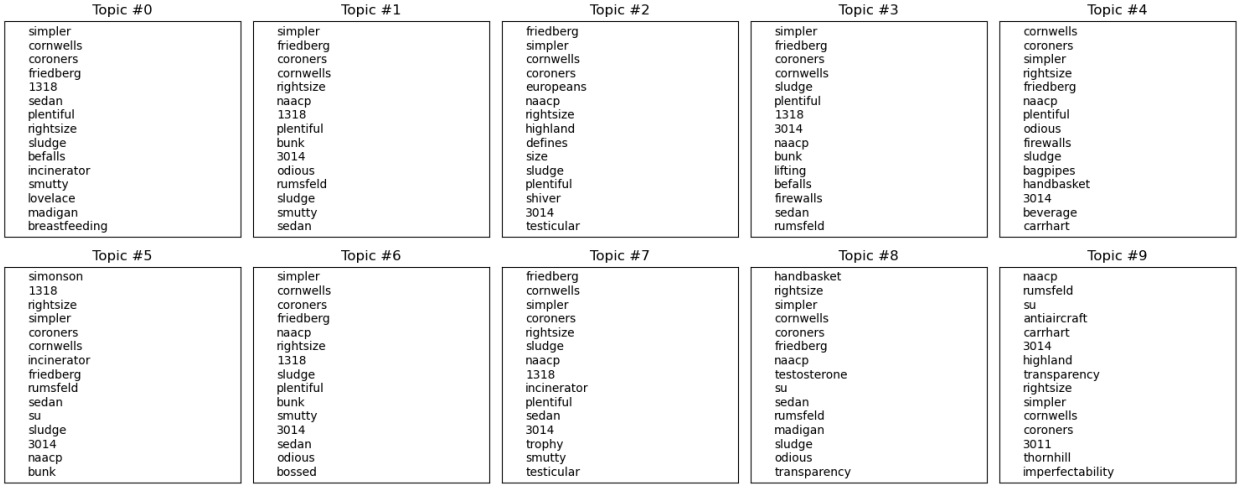
RESULTS

The LDA model was trained with varying numbers of topics (5, 10, 20, and 30) to explore different levels of granularity in topic identification. The top words representing each topic were extracted, providing insights into the main themes discussed during the floor debates.

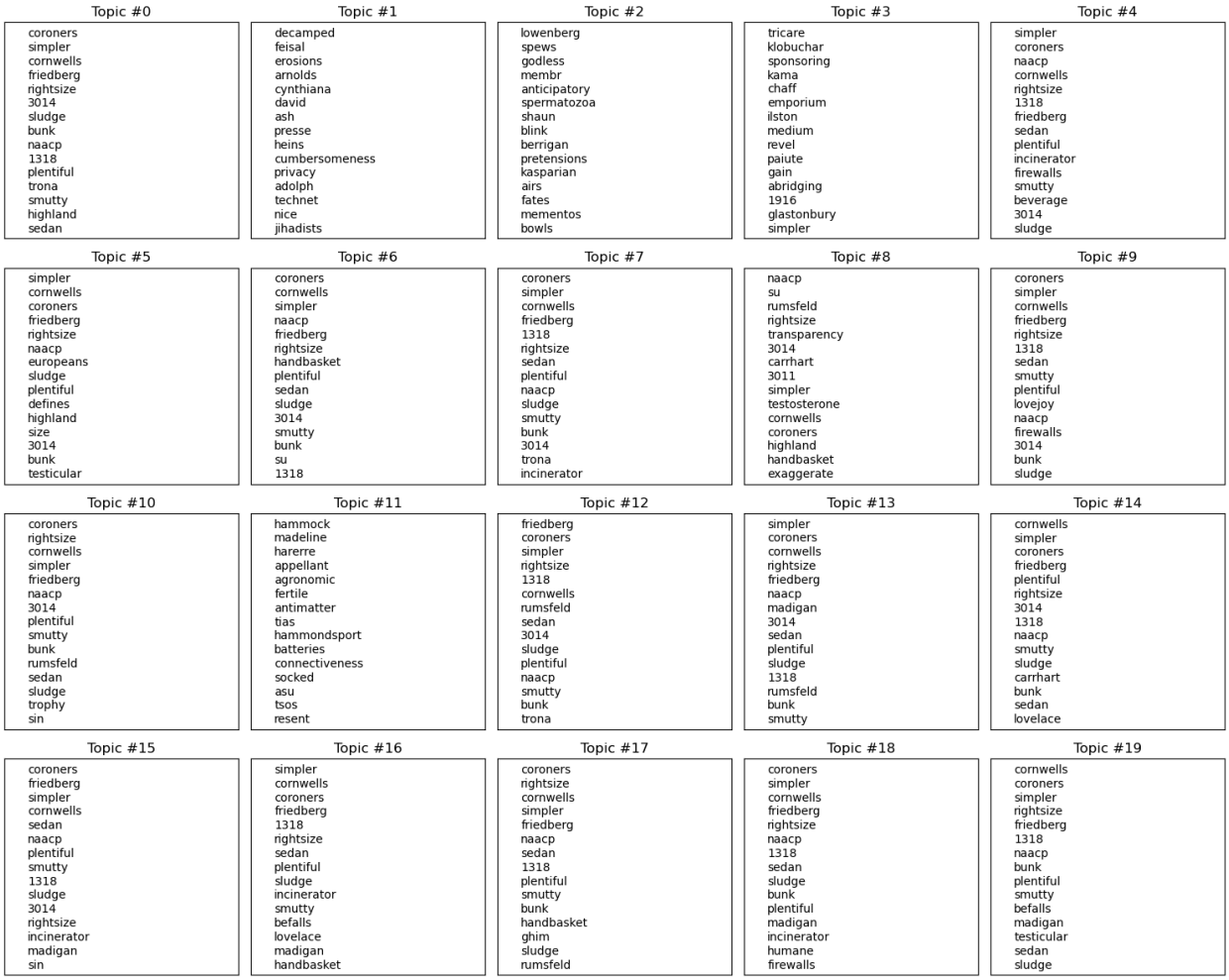
5 Topics



10 Topics



20 Topics


Topic #0
coroners
simpler
cornwells
friedberg
rightsize
3014
sludge
bunk
naacp
1318
plentiful
trona
smutty
highland
sedan
Topic #1
decamped
feisal
erosions
arnolds
cynthiana
david
ash
presse
heins
cumbersomeness
privacy
adolph
technet
nice
jihadists
Topic #2
lowenberg
spews
godless
membr
anticipatory
spermatozoa
shaun
blink
berrigan
pretensions
kasparian
airs
fates
mementos
bowls
Topic #3
tricare
klobuchar
sponsoring
kama
chaff
emporium
ilston
medium
revel
paiute
gain
abridging
1916
glastonbury
simpler
Topic #4
simpler
coroners
naacp
cornwells
rightsize
1318
friedberg
sedan
plentiful
incinerator
firewalls
smutty
beverage
3014
sludge
Topic #5
simpler
cornwells
coroners
friedberg
rightsize
naacp
europeans
sludge
plentiful
defines
highland
size
3014
bunk
testicular
Topic #6
coroners
cornwells
simpler
naacp
friedberg
rightsize
handbasket
plentiful
sedan
sludge
3014
smutty
bunk
su
1318
Topic #7
coroners
simpler
cornwells
friedberg
1318
rightsize
sedan
plentiful
naacp
sludge
smutty
bunk
3014
trona
incinerator
Topic #8
naacp
su
rumsfeld
rightsize
transparency
3014
carrhart
3011
simpler
testosterone
cornwells
coroners
highland
handbasket
exaggerate
Topic #9
coroners
simpler
cornwells
friedberg
rightsize
1318
sedan
smutty
plentiful
lovejoy
naacp
firewalls
3014
bunk
sludge
Topic #10
coroners
rightsize
cornwells
simpler
friedberg
naacp
3014
plentiful
smutty
bunk
rumsfeld
sedan
sludge
trophy
sin
Topic #11
hammock
madeline
harerre
appellant
agronomic
fertile
antimatter
tias
hammondsport
batteries
connectiveness
socked
asu
tsos
resent
Topic #12
friedberg
coroners
simpler
rightsize
1318
cornwells
rumsfeld
sedan
3014
sludge
plentiful
naacp
smutty
bunk
trona
Topic #13
simpler
coroners
cornwells
rightsize
friedberg
naacp
madigan
3014
sedan
plentiful
sludge
1318
rumsfeld
bunk
smutty
Topic #14
cornwells
simpler
coroners
friedberg
plentiful
rightsize
3014
1318
naacp
smutty
sludge
carrhart
bunk
sedan
lovelace
Topic #15
coroners
friedberg
simpler
cornwells
sedan
naacp
plentiful
smutty
1318
sludge
3014
rightsize
incinerator
madigan
sin
Topic #16
simpler
cornwells
coroners
friedberg
1318
rightsize
sedan
plentiful
sludge
incinerator
smutty
befalls
lovelace
madigan
handbasket
Topic #17
coroners
rightsize
cornwells
simpler
friedberg
naacp
sedan
1318
plentiful
smutty
bunk
handbasket
ghim
sludge
rumsfeld
Topic #18
coroners
simpler
cornwells
friedberg
rightsize
naacp
1318
sedan
sludge
bunk
plentiful
madigan
incinerator
humane
firewalls
Topic #19
cornwells
coroners
simpler
rightsize
friedberg
1318
naacp
bunk
plentiful
smutty
befalls
madigan
testicular
sedan
sludge

30 Topics

Topic #0	Topic #1	Topic #2	Topic #3	Topic #4
coroners	simpler	soothes	su	cornwells
friedberg	cornwells	gear	rumsfeld	coroners
simpler	coroners	frieze	naacp	simpler
cornwells	rightsize	odious	rightsize	rightsize
rightsize	friedberg	1937	transparency	friedberg
naacp	sludge	bourgeois	handbasket	1318
plentiful	incinerator	transitioning	preeclampsia	bunk
1318	1318	hellenic	fulfills	sedan
smutty	plentiful	3011	exaggerate	3014
incinerator	madigan	astronautical	testosterone	sludge
madigan	odious	letpp	antiaircraft	naacp
sludge	trophy	nate	simpler	befalls
sedan	handbasket	doster	coroners	madigan
bunk	highland	brewers	madigan	humane
rumsfeld	naacp	saroyan	cornwells	ghim

Topic #5	Topic #6	Topic #7	Topic #8	Topic #9
friedberg	balkanizing	simpler	simpler	friedberg
simpler	masquerading	cornwells	cornwells	simpler
coroners	athleticism	friedberg	friedberg	cornwells
cornwells	hendry	coroners	coroners	coroners
3014	raceptives	sludge	sedan	rightsize
rightsize	biorefining	plentiful	plentiful	naacp
sludge	jsf	3014	naacp	1318
naacp	carriers	sedan	1318	sludge
plentiful	transformative	rightsize	rightsize	trophy
sedan	allstate	naacp	incinerator	plentiful
handbasket	evacuating	1318	madigan	europeans
1318	gunpowder	incinerator	sludge	madigan
carrhart	dave	smutty	befalls	sedan
incinerator	pffs	highland	firewalls	3014
odious	abhors	handbasket	beverage	smutty

Topic #10	Topic #11	Topic #12	Topic #13	Topic #14
lifting	coroners	cornwells	coroners	coroners
melding	friedberg	coroners	cornwells	simpler
hendry	simpler	friedberg	simpler	cornwells
trainmen	cornwells	simpler	friedberg	friedberg
godlessness	naacp	rightsize	rightsize	sedan
firewalls	rightsize	3014	plentiful	rightsize
backhanded	plentiful	sludge	naacp	sludge
banc	1318	naacp	1318	plentiful
bunkhouses	incinerator	sedan	sludge	1318
beverage	sludge	smutty	madigan	rumsfeld
heterosexual	sedan	plentiful	trona	smutty
bars	handbasket	1318	smutty	incinerator
1842	smutty	bunk	3014	beverage
allstate	bunk	rumsfeld	firewalls	bunk
blockade	madigan	incinerator	rumsfeld	su

Topic #15	Topic #16	Topic #17	Topic #18	Topic #19
cornwells	cornwells	simpler	simpler	coroners
coroners	coroners	coroners	cornwells	simpler
simpler	friedberg	cornwells	coroners	1318
friedberg	1318	friedberg	rightsize	friedberg
rightsize	rightsize	rightsize	friedberg	cornwells
plentiful	simpler	naacp	incinerator	rightsize
sedan	sludge	bunk	1318	naacp
1318	naacp	3014	sludge	plentiful
naacp	madigan	handbasket	plentiful	sedan
sludge	bunk	1318	bunk	bunk
incinerator	sedan	plentiful	lovelace	incinerator
smutty	smutty	smutty	naacp	3014
3014	incinerator	sludge	beverage	beverage
madigan	plentiful	madigan	carrhart	sludge
su	firewalls	sedan	sedan	befalls

Topic #20	Topic #21	Topic #22	Topic #23	Topic #24
simpler	coroners	cornwells	cornwells	simpler
cornwells	cornwells	friedberg	simpler	friedberg
coroners	friedberg	coroners	coroners	cornwells
friedberg	simpler	simpler	rightsize	coroners
1318	naacp	rightsize	1318	1318
rightsize	1318	1318	plentiful	plentiful
plentiful	3014	handbasket	friedberg	incinerator
sedan	incinerator	3014	naacp	rightsize
sludge	sedan	naacp	incinerator	naacp
incinerator	bunk	plentiful	bunk	sludge
smutty	rightsize	rumsfeld	sludge	beverage
befalls	plentiful	madigan	sedan	carrhart
lovelace	rumsfeld	sedan	trophy	sedan
bunk	simonson	incinerator	madigan	smutty
breastfeeding	odious	firewalls	handbasket	bunk

Topic #25	Topic #26	Topic #27	Topic #28	Topic #29
simpler	juveniles	authenticity	cornwells	eyeshade
cornwells	thinnest	carriers	friedberg	ploesti
coroners	kalaukaua	punching	coroners	dressing
friedberg	ky1	interceded	simpler	simpler
naacp	incredibly	metropolitan	rightsize	friedberg
europeans	3494	thermodynamics	naacp	cornwells
rightsize	interceptors	purportedly	1318	coroners
highland	berrigan	extrapolations	sludge	1318
3014	herodotus	taped	trophy	rightsize
size	2671	herodotus	sedan	plentiful
sludge	rewind	incredibly	plentiful	sedan
defines	hirohito	mujumdar	su	naacp
plentiful	elan	sorenson	carrhart	smutty
testicular	enabling	3229	ghim	handbasket
bunk	apportion	324	handbasket	bunk

CONCLUSIONS

The analysis of the 110th Congress floor debate dataset using topic modeling techniques has yielded significant insights:

- Identified Topics: The analysis successfully identified and categorized topics discussed during the congressional debates, providing a structured view of the key themes and issues addressed.

- Trends and Priorities: The identified topics shed light on the legislative priorities and societal trends during the 110th Congress, offering valuable historical context for policymakers and researchers.

- Public Perception: By understanding the prevalent topics and discussions, businesses and policymakers can gauge public sentiment and concerns, aiding in decision-making and policy formulation.

- Strategic Implications: The insights gained from topic modeling have strategic implications for various sectors, including policy development, public relations, risk management, and content strategy.

In conclusion, leveraging advanced analytics techniques like topic modeling on textual data sets such as the "110" dataset provides a robust framework for understanding complex narratives, uncovering hidden insights, and driving informed decision-making across diverse domains.