

---

# Budget Text Analysis

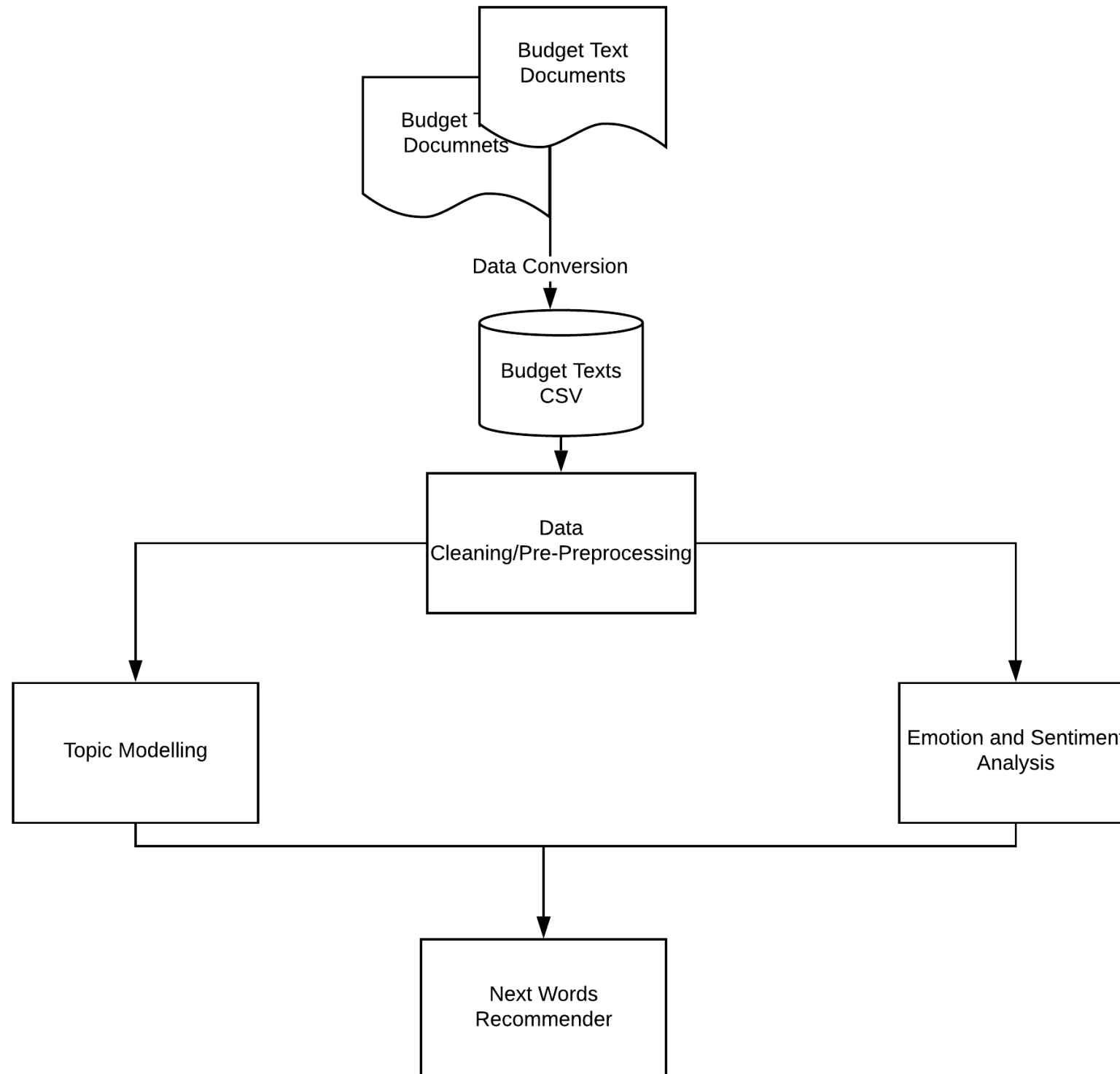
- Datatopian Visionaries

Akash Meghani,  
Miguel Gaspar Utrera,  
Naseeb Thapaliya,  
Sultan Al Bogami,  
Unnati Khivasara

Mentors: Dr. Soumya Mohanty  
Jason Jones (Guilford County)

---

# Overview of the Project



---

# Goals

---

- ❖ Understand the Different sections of budget text data from different counties and create a relation between them.
- ❖ Compare the general funds section of Guilford county, Durham County and Charlotte City (2008 and 2019) and understand the difference between them.
- ❖ Visualization of emotions between 2008 and 2019.
- ❖ Understand the different relevant topics from all the counties 2019 and with computed their coherence score with proper visualization.
- ❖ Compared the topic modeling results over the years (2008,2012,2016,2020)

---

# Team Structure

---

- ❖ All the individuals will work on preparing data i.e. Perform Data cleaning and Data preprocessing.
- ❖ Team will be divided into 2 groups to perform different tasks:
  - Team 1: Topic Modelling  
Members:
    1. Naseeb Thapaliya
    2. Miguel Gasper Utrera
  - Team 2: Emotion and Sentiment Analysis  
Members:
    1. Akash Meghani
    2. Unnati Khivasara

---

# Individual Tasks Done

---

❖ **Sultan Al Bogami**

1. Collected Budget Documents from all the different Counties websites and other sources(2008 to 2020) and organization of github.
2. Converted the pdf documents to csv formats. Extract words from the documents using online tool, and classify them for further processing.

❖ **Naseeb Thapaliya**

1. Compared the topic modeling results over the years (2008,2012,2016,2020)

❖ **Miguel Gasper Utrera**

1. Applied Topic modeling on different relevant topics from all the counties and computed their coherence score with proper visualization.
2. Applied Davis model and showed top 30 words in each topic and their relevance.

❖ **Unnati Khivasera**

1. Analyzing sentiment intensity using Vader.
2. Performed visualization of emotions from different sections of documents.

❖ **Akash Meghani**

1. Applied Emotional and Sentiment analysis with NLTK and got meaningful results.
2. Performed visualization of emotions from different sections of documents.

---

# Emotion And Sentiment Analysis

---

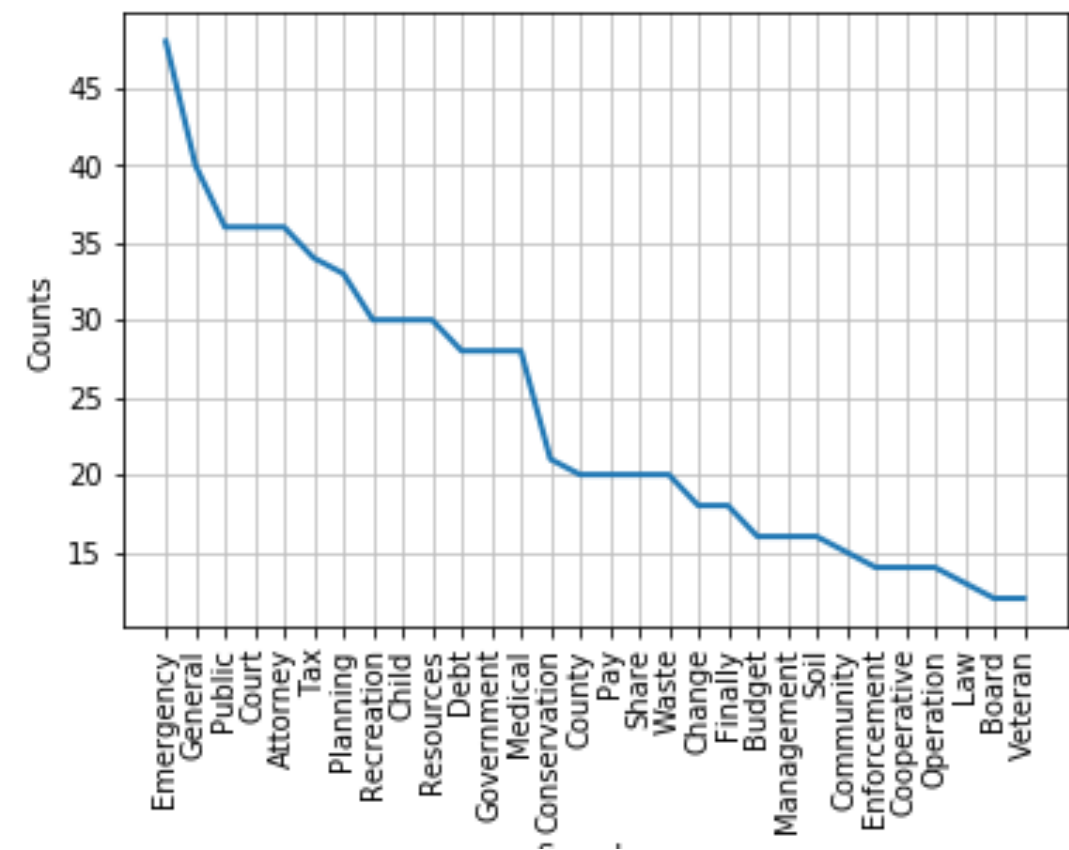
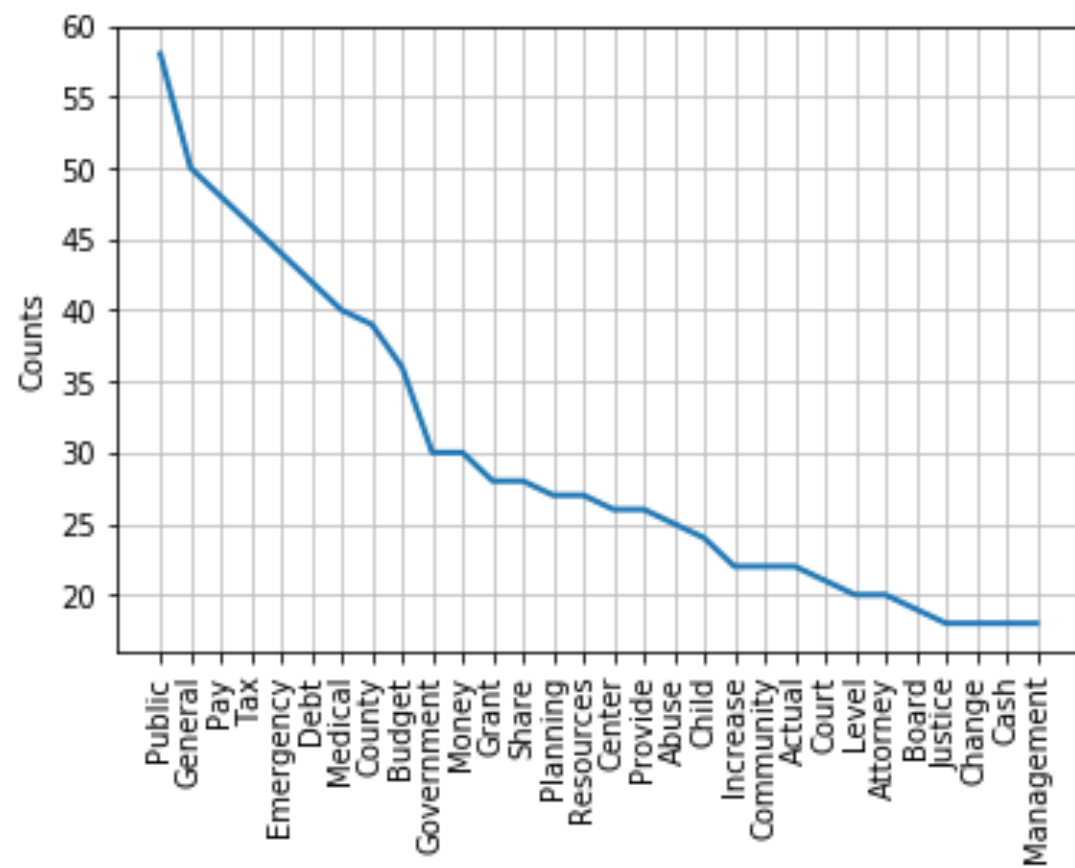
## Most Influential Words in Guilford County (2020 and 2008)

```
[('Public', 58),  
 ('General', 50),  
 ('Pay', 48),  
 ('Tax', 46),  
 ('Emergency', 44),  
 ('Debt', 42),  
 ('Medical', 40),  
 ('County', 39),  
 ('Budget', 36),  
 ('Government', 30)]
```

```
[('Emergency', 48),  
 ('General', 40),  
 ('Public', 36),  
 ('Court', 36),  
 ('Attorney', 36),  
 ('Tax', 34),  
 ('Planning', 33),  
 ('Recreation', 30),  
 ('Child', 30),  
 ('Resources', 30)]
```

# Emotion And Sentiment Analysis

## Distribution of most influential Words in Guilford County (2020 and 2008)



---

# Emotion And Sentiment Analysis

---

We have assigned numerical value to every emotion present in the document. Here is the list :

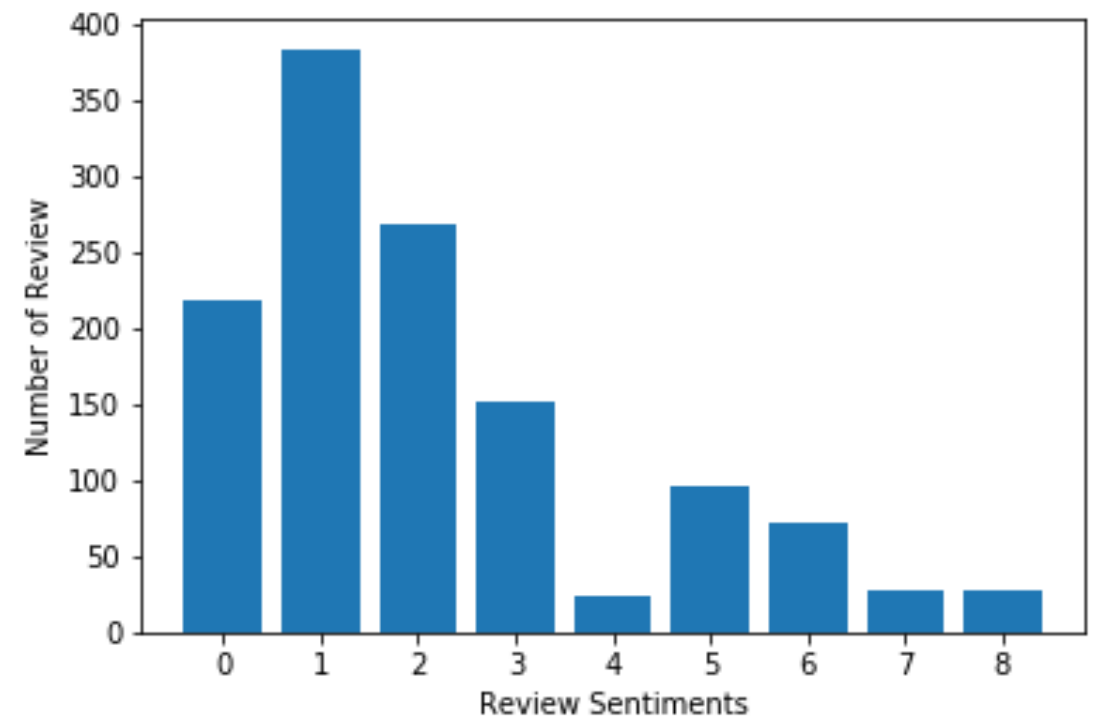
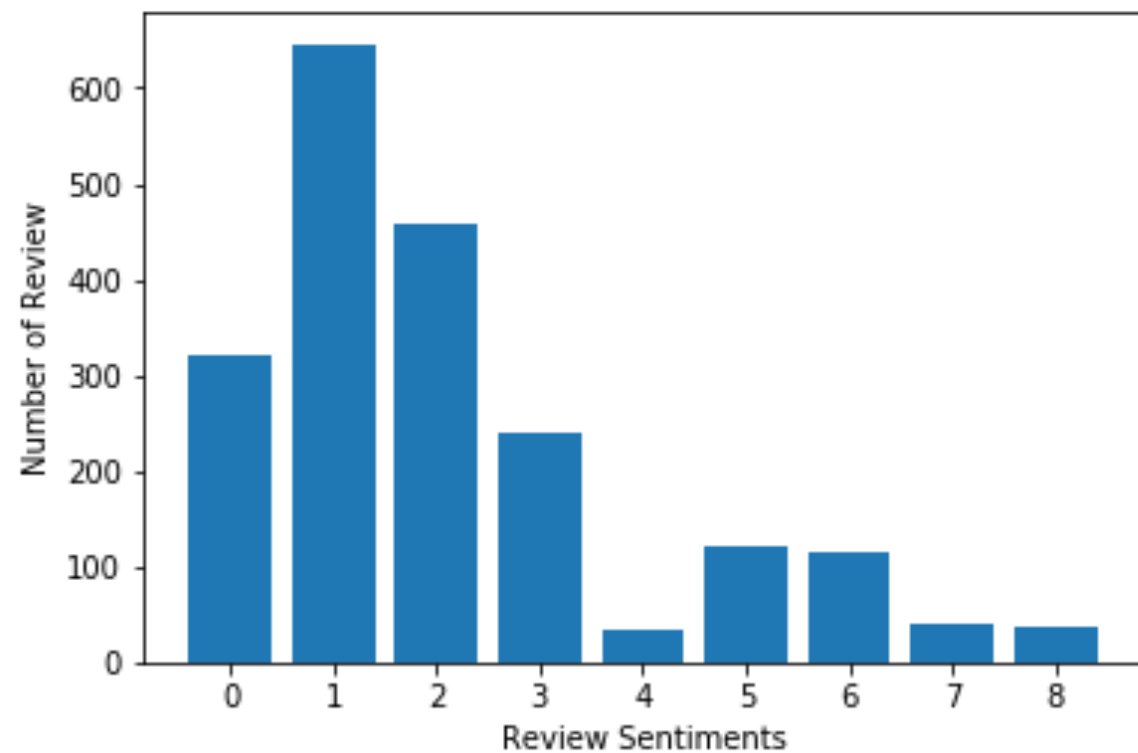
"Negative": "0", "Positive":

"1", "Trust" : "2", "Sadness": "0", "Anticipation": "3", "Surprise": "4", "Fear": "5", "Joy": "6", "Anger": "7", "Disgust": "8"



# Emotion And Sentiment Analysis

## Distribution of Emotions in General fund summary section (2020 and 2008)



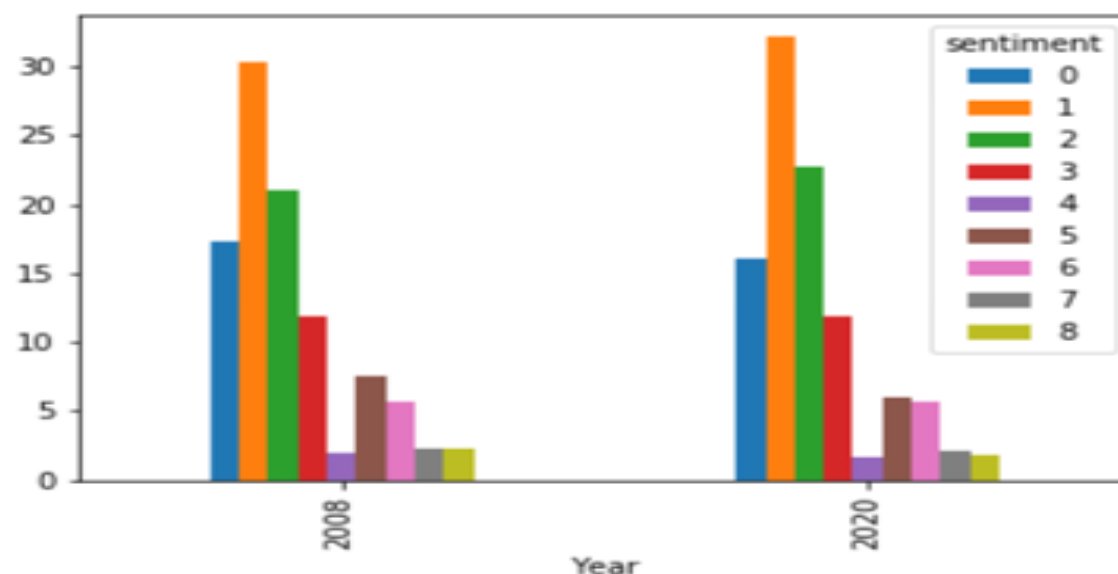
# Emotion And Sentiment Analysis

## Distribution of Emotions in General fund summary section (2020 and 2008) for Guilford County

|           |           |           |           |           |          |          |
|-----------|-----------|-----------|-----------|-----------|----------|----------|
| 32/8      |           |           |           |           |          |          |
| sentiment | 0         | 1         | 2         | 3         | 4        | 5        |
| Year      |           |           |           |           |          |          |
| 2008      | 17.219589 | 30.252765 | 21.090047 | 11.927330 | 1.895735 | 7.582938 |
| 2020      | 16.003976 | 32.107356 | 22.763419 | 11.928429 | 1.689861 | 5.964215 |

|           |          |          |          |
|-----------|----------|----------|----------|
| sentiment | 6        | 7        | 8        |
| Year      |          |          |          |
| 2008      | 5.608215 | 2.211690 | 2.211690 |
| 2020      | 5.715706 | 2.037773 | 1.789264 |

<matplotlib.axes.\_subplots.AxesSubplot at 0x20cf2a6e4a8>



# Emotion And Sentiment Analysis

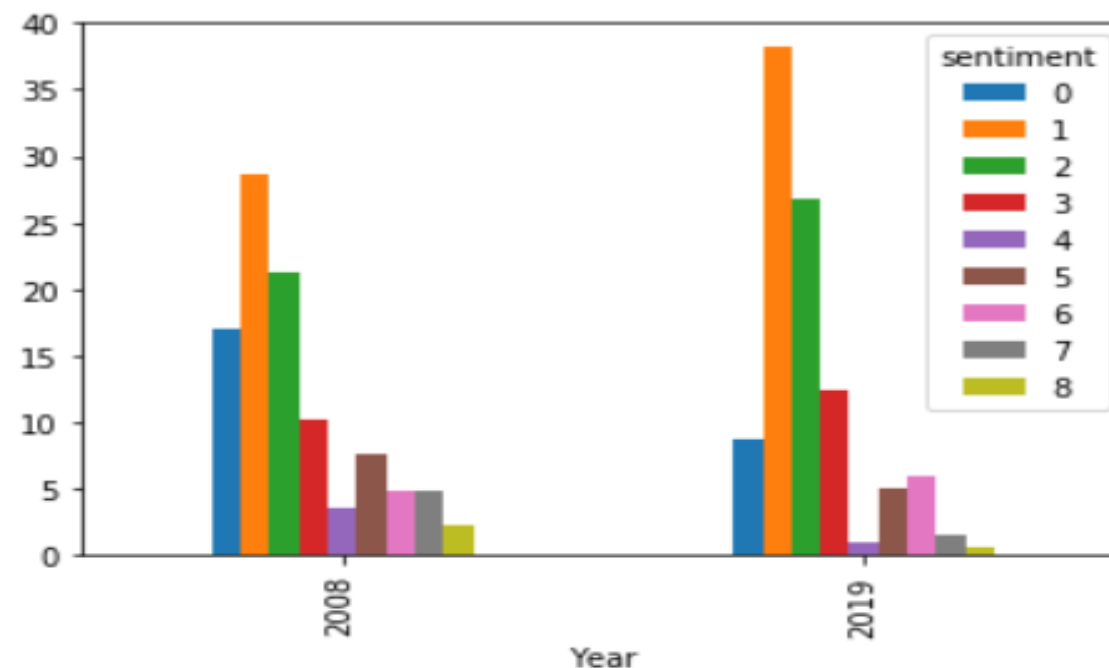
## Distribution of Emotions in General fund summary section (2020 and 2008) for Charolette

|           |           |           |           |           |          |          |
|-----------|-----------|-----------|-----------|-----------|----------|----------|
| 6221      |           |           |           |           |          |          |
| sentiment | 0         | 1         | 2         | 3         | 4        | 5        |
| Year      |           |           |           |           |          |          |
| 2008      | 16.993464 | 28.540305 | 21.241830 | 10.130719 | 3.594771 | 7.625272 |
| 2019      | 8.730907  | 38.148218 | 26.758439 | 12.370356 | 0.999434 | 5.091458 |

|           |          |          |          |
|-----------|----------|----------|----------|
| sentiment | 6        | 7        | 8        |
| Year      |          |          |          |
| 2008      | 4.793028 | 4.793028 | 2.287582 |
| 2019      | 5.883462 | 1.470866 | 0.546860 |

<matplotlib.axes.\_subplots.AxesSubplot at 0x20cf2b25780>



---

# Emotion And Sentiment Analysis

---

- ❖ We have used multiple methods like NLTK, Text blob and Vader to figure out what make sense with our data.
- ❖ We have applied NLTK on multiple sections of the document but we have only presented interesting things.

# Topic Modeling

```
[(0,
 '0.315*"total" + 0.056*"commissioner" + 0.052*"park" + 0.051*"property" + '
 '0.044*"security" + 0.044*"resource" + 0.035*"policy" + 0.032*"economic" + '
 '0.027*"performance" + 0.026*"amend"'),
 (1,
 '0.196*"program" + 0.153*"provide" + 0.106*"major" + 0.064*"grant" + '
 '0.063*"exist" + 0.053*"operation" + 0.039*"information" + 0.037*"change" + '
 '0.035*"work" + 0.034*"care"'),
 (2,
 '0.115*"fund" + 0.110*"summary" + 0.108*"fire" + 0.078*"area" + '
 '0.062*"current" + 0.060*"solid" + 0.048*"state" + 0.041*"level" + '
 '0.040*"percent" + 0.039*"estimate"'),
 (3,
 '0.187*"fiscal" + 0.086*"debt" + 0.074*"unit" + 0.068*"water" + '
 '0.060*"infrastructure" + 0.050*"issue" + 0.044*"goal" + 0.042*"remain" + '
 '0.042*"government" + 0.041*"base"'),
 (4,
 '0.206*"adopt" + 0.107*"replacement" + 0.090*"support" + 0.083*"increase" + '
 '0.075*"number" + 0.044*"charge" + 0.041*"planning" + 0.038*"additional" + '
 '0.038*"require" + 0.038*"site"'),
 (5,
 '0.219*"capital" + 0.134*"expenditure" + 0.100*"management" + '
 '0.072*"equipment" + 0.050*"balance" + 0.044*"vehicle" + 0.040*"begin" + '
 '0.036*"improve" + 0.030*"identify" + 0.026*"law"'),
 (6,
 '0.242*"include" + 0.164*"community" + 0.095*"school" + 0.075*"impact" + '
 '0.037*"rate" + 0.033*"maintain" + 0.027*"recommend" + 0.027*"associate" + '
 '0.026*"pay" + 0.024*"resident"'),
 (7,
 '0.259*"year" + 0.150*"funding" + 0.117*"public" + 0.080*"development" + '
 '0.077*"actual" + 0.044*"plan" + 0.029*"annual" + 0.024*"life" + '
 '0.021*"address" + 0.019*"help"'),
 (8,
 '0.047*"service" + 0.019*"system" + 0.013*"building" + 0.012*"improvement" + '
 '0.012*"operate" + 0.011*"transfer" + 0.010*"cost" + 0.010*"source" + '
 '0.010*"complete" + 0.009*"future"'),
 (9,
 '0.260*"budget" + 0.207*"project" + 0.180*"facility" + 0.133*"revenue" + '
 '0.052*"tax" + 0.024*"appropriate" + 0.024*"control" + 0.015*"specific" + '
 '0.014*"population" + 0.011*"food"')]
```

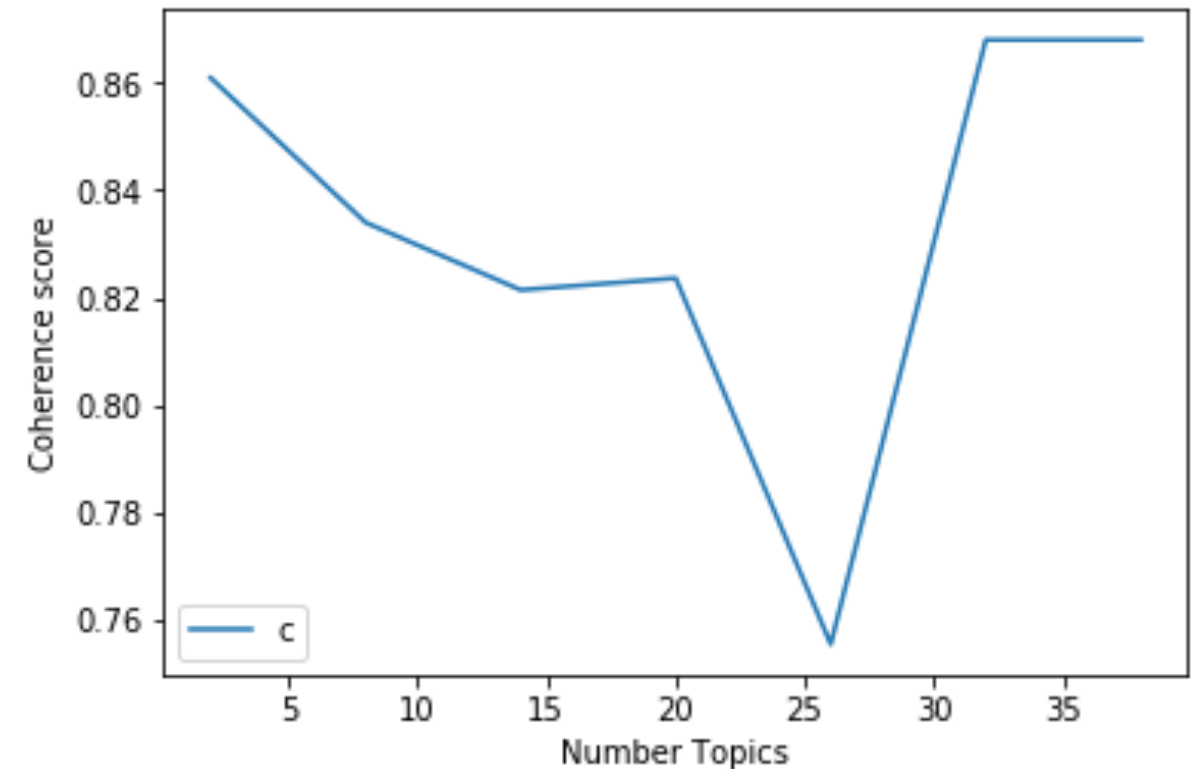
# Topic Modeling

Topic 0  
property resource  
commissioner  
park policy  
security  
**total**  
economic performance  
amend

Topic 2  
percent  
**fire**  
level  
state estimate  
current area  
solid  
**fund**

Topic 1  
work operation major  
care  
**provide**  
exist  
**program**  
information grant  
change

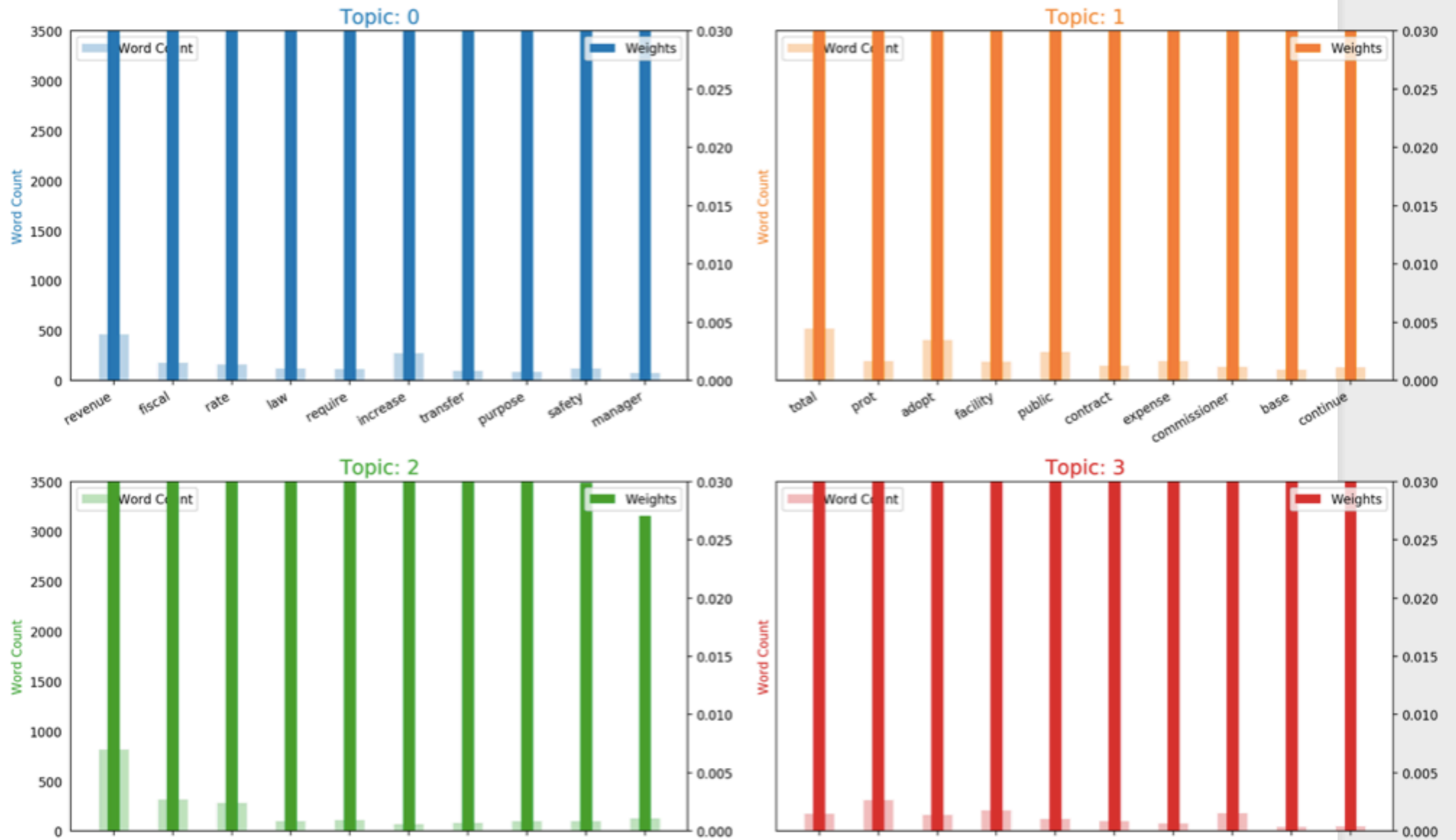
Topic 3  
infrastructure debt  
base  
unit issue  
remain goal  
**fiscal**  
water government



Coherence Score: 0.8256146597574272

# Topic Modeling

Word Count and Importance of Topic Keywords



# Topic Modeling Comparison

Topic 0  
property resource  
commissioner  
park policy  
security  
total  
economic performance  
amend

Topic 1  
work operation major  
provide  
exist  
program  
information grant  
change

Topic 2  
percent  
fire  
level summary  
state estimate  
current area  
solid fund

Topic 3  
infrastructure base debt  
unit issue  
remain goal  
fiscal  
water government

Topic 0  
adopt employee  
increase facility prior  
project  
balance information  
department risk

Topic 1  
care  
capital  
level  
transportation maintain  
student tax  
technology  
base court

Topic 2  
law economic  
staff  
area  
expenditure total  
program work  
space administration

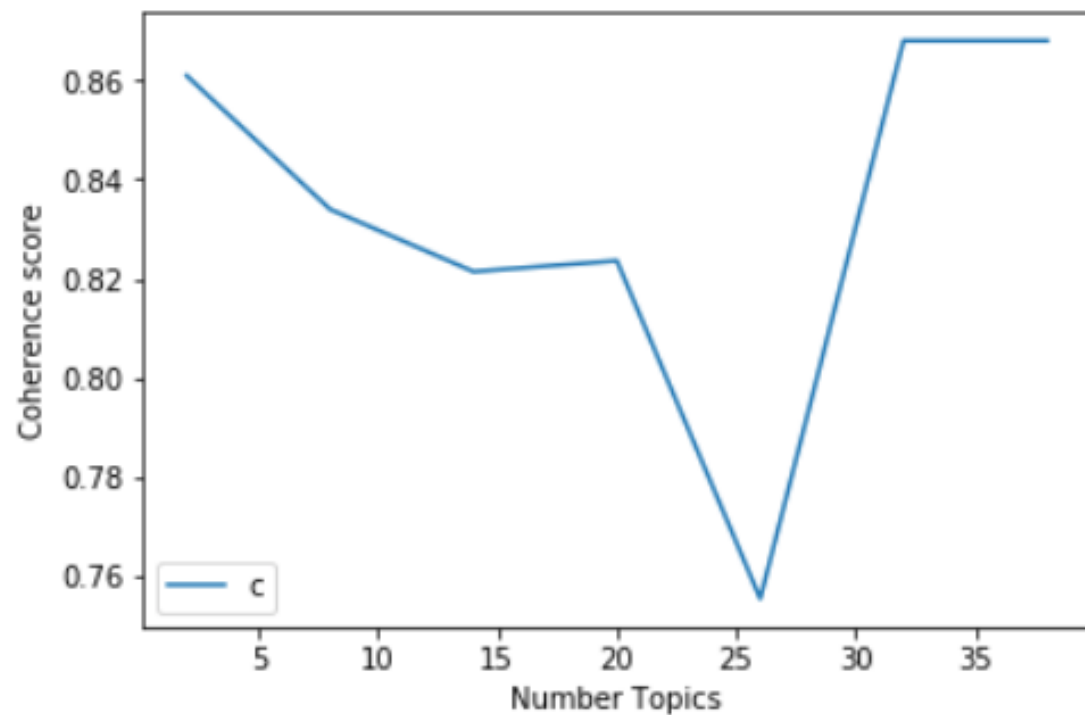
Topic 3  
cost  
ordinance  
require  
bond  
point  
fund  
system  
incentive result  
rate

2019

2008

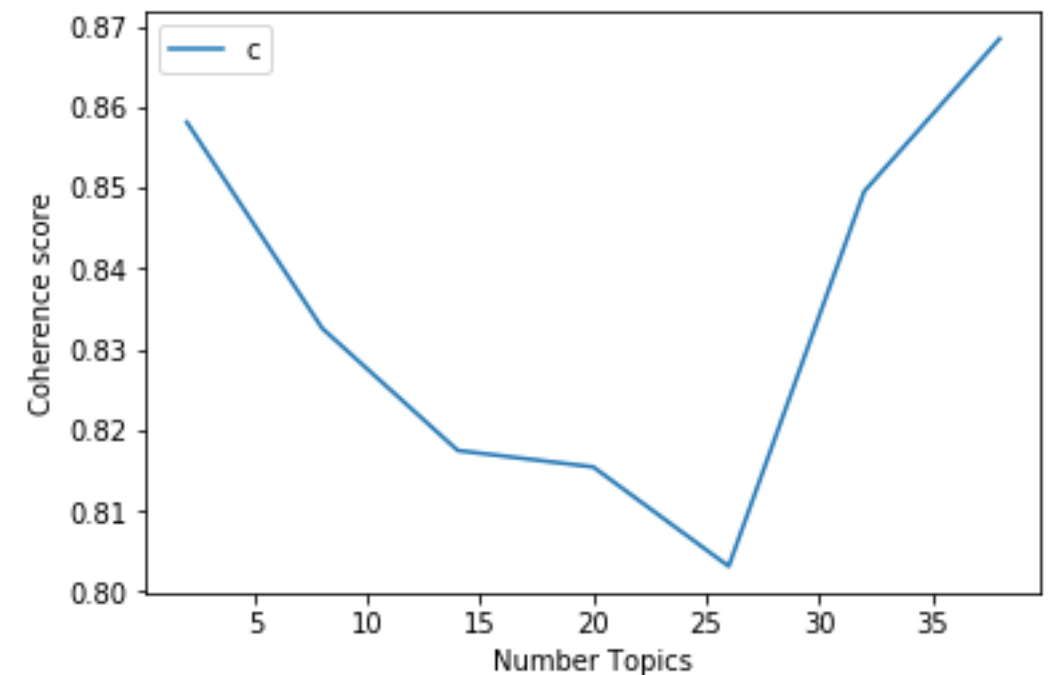


# Topic Modeling Comparison



Coherence Score: 0.8256146597574272

2019



Coherence Score: 0.8247949042506306

2008

---

# Next Word Recommender

---

- Whenever a user tries to enter a word/s suggest the next word based on combination of words used as input in previous searches.
- Use results from Topic modeling to predict the recommended word/topic which are important.

---

# Relevant Work

---

- **Emotion Sentiment Extraction Website by Jason.**(<https://jason-jones.shinyapps.io/Emotionizer/>)
- **“Peoples Opinion on Indian Budget Using Sentiment Analysis”** -Varat Nayak