
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

❖ **Budget text Analysis for counties and cities:**

- ◆ Guilford County
- ◆ Wake County
- ◆ Mecklenburg County
- ◆ Durham County
- ◆ City of Charlotte
- ◆ City of Durham
- ◆ City of Raleigh

Goals

- ❖ **Understand the data**
- ❖ **Data Cleaning/Pre-processing**
- ❖ **Topic Modelling of the textual data**
- ❖ **Emotion and Sentiment Analysis of the budget texts to draw up public's emotional engagement.**
- ❖ **Next work recommender for the texts in budget.**

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
 3. Sultan Al Bogami
 - Team 2: Emotion and Sentiment Analysis
Members:
 1. Akash Meghani
 2. Unnati Khivasara

Data Overview

- ❖ **Primarily, 7 pdf files ranging from 400-500 pages long for each.**
- ❖ **Each pdf is converted to csv files by extracting all the relevant budget texts(words) from the pdf file.**
- ❖ **So, there are total of 638131 total words extracted from the budget files.**

Data Source



Guilford County
STATE of NORTH CAROLINA

Search...



Services

Our County

Business

Get Connected

How Do I...

Budget, Management & Evaluation

FY 2019-20 Adopted Budget

How are your Tax Dollars Spent?

Budget Amendments Reports

Budget Performance Reports

+ Budget History & Past Adopted Budget Documents

+ Capital Investment Plan & Capital Project Status

Other Financial Information

Contact Information

Our County » Budget, Management & Evaluation »

FY 2019-20 Adopted Budget

Font Size: [+](#) [-](#) [+](#) Share & Bookmark [Feedback](#) [Print](#)




[FY 2019-20 Adopted Budget Document](#)

[FY 2019-20 Adopted Budget-in-Brief](#)



Data Conversion

 PDF Upload Word Grouping Visualize Download Data

This application allows you to upload a PDF document and start exploring the text contained within. Once you upload a file, a searchable data table will render for you on this page. Upload a PDF document below to start exploring!

Choose PDF File

Browse...

MecklenburgCounty

Upload complete

Move over to the Word Grouping tab to start exploring your document's text by n-grams. [Click here to learn more about n-grams](#)

537

Document Pages

24,719

Tokenized Words

Show

10

 entries

Search:

Page Number	Word	Word Count	Association	Category
110	Debt	29	Negative	Sentiment
110	Debt	29	Sadness	Emotion
24	Debt	28	Negative	Sentiment
24	Debt	28	Sadness	Emotion
506	Debt	26	Negative	Sentiment
506	Debt	26	Sadness	Emotion
505	Debt	24	Negative	Sentiment
505	Debt	24	Sadness	Emotion
46	Budget	21	Trust	Emotion
466	Debt	21	Negative	Sentiment

Showing 1 to 10 of 15,660 entries

Previous

1

2

3

4

5

...

1566

Next

Data Transformation

```
In [54]: GC_df = pd.read_csv(r"../util/data/structured/original/GuilfordCounty_original_data.csv")
GC_df.drop(['Unnamed: 0'], axis=1, inplace=True)
GC_df['label'] = '0'
GC_df.shape
GC_df.head(5)
```

Out[54]:

	page_number	word	label
0	2	guilford	0
1	2	county	0
2	2	by	0
3	2	the	0
4	2	numbers	0

```
In [55]: CC_df = pd.read_csv(r"../util/data/structured/original/CharlotteCity_original_data.csv")
CC_df.drop(['Unnamed: 0'], axis=1, inplace=True)
CC_df['label'] = '1'
CC_df.head(5)
```

Out[55]:

	page_number	word	label
0	1	ensuring	1
1	1	an	1
2	1	equitable	1
3	1	sustainable	1
4	1	and	1

Data Analysis

```
In [47]: Combined_df.shape
```

```
Out[47]: (638131, 3)
```

```
In [45]: Combined_df.describe()
```

```
Out[45]:
```

	page_number
count	638131.000000
mean	213.602262
std	137.058241
min	1.000000
25%	100.000000
50%	203.000000
75%	305.000000
max	537.000000

```
In [50]: Combined_df.to_csv("Combined_Counties.csv", sep='\t', encoding='utf-8')
```

```
In [ ]:
```

Data Analysis

```
In [47]: Combined_df.shape
```

```
Out[47]: (638131, 3)
```

```
In [45]: Combined_df.describe()
```

```
Out[45]:
```

	page_number
count	638131.000000
mean	213.602262
std	137.058241
min	1.000000
25%	100.000000
50%	203.000000
75%	305.000000
max	537.000000

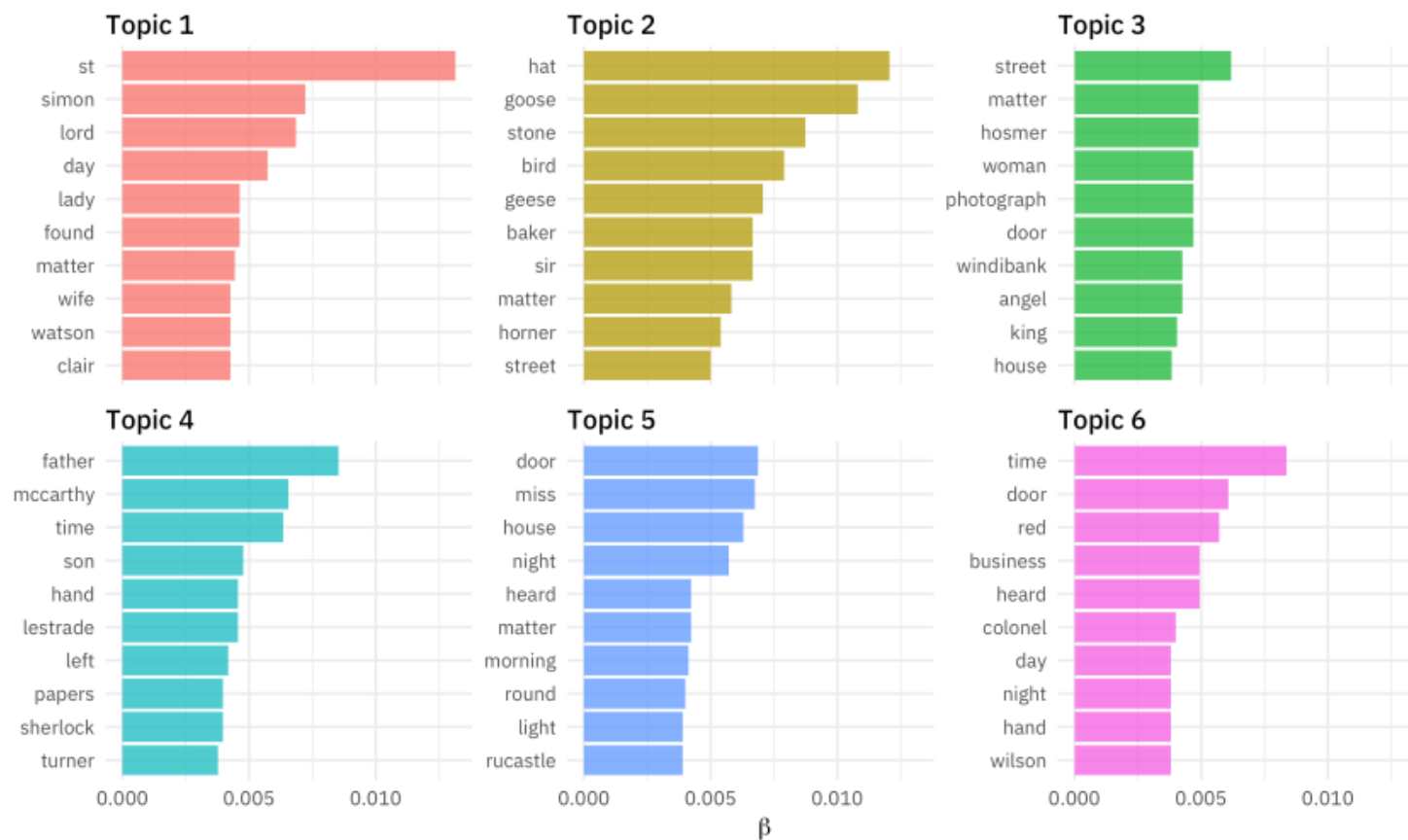
```
In [50]: Combined_df.to_csv("Combined_Counties.csv", sep='\t', encoding='utf-8')
```

```
In [ ]:
```

Topic Modelling

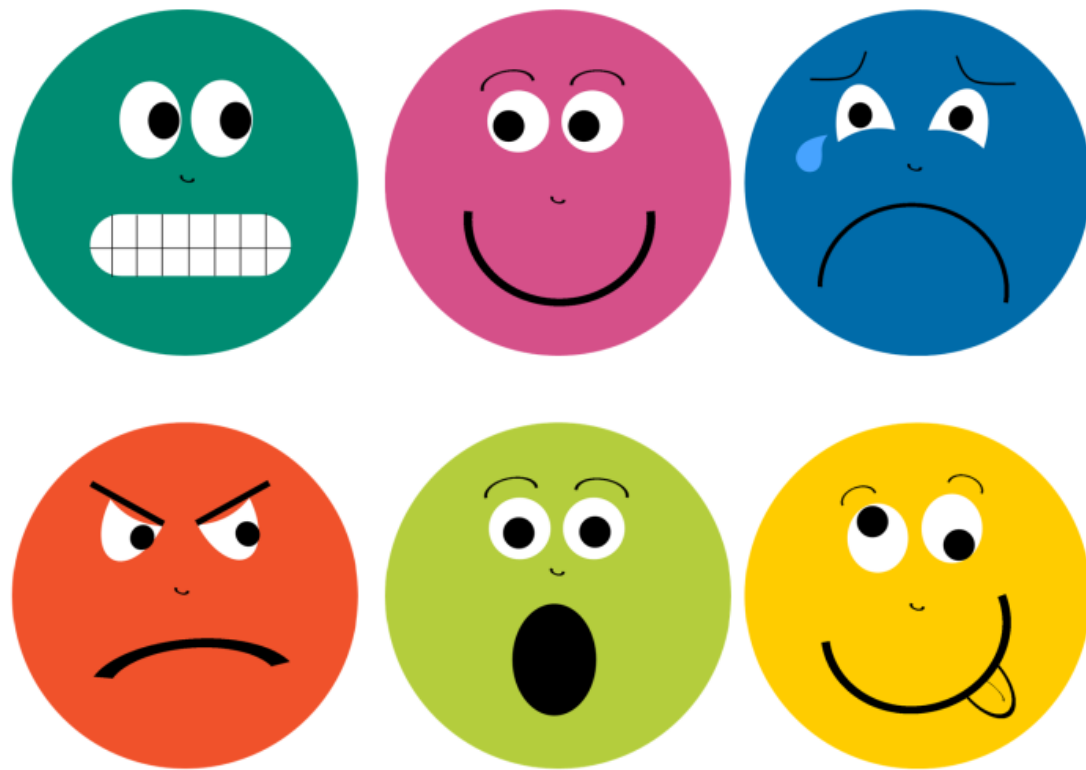
Highest word probabilities for each topic

Different words are associated with different topics



- Finding a group of words (i.e topic) from a collection of documents that best represents the information in the collection.

Emotion And Sentiment Analysis



- Sentiment analysis and emotional analysis are two key methods experts use to quantify audiences' emotional engagement.

Next Word Recommender(optional)

- Whenever a user tries to enter a word/s suggest the next word based on combination of words used as input in previous searches.

Relevant Work

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