Budget Text Analysis

- Datatopian Visionaries

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

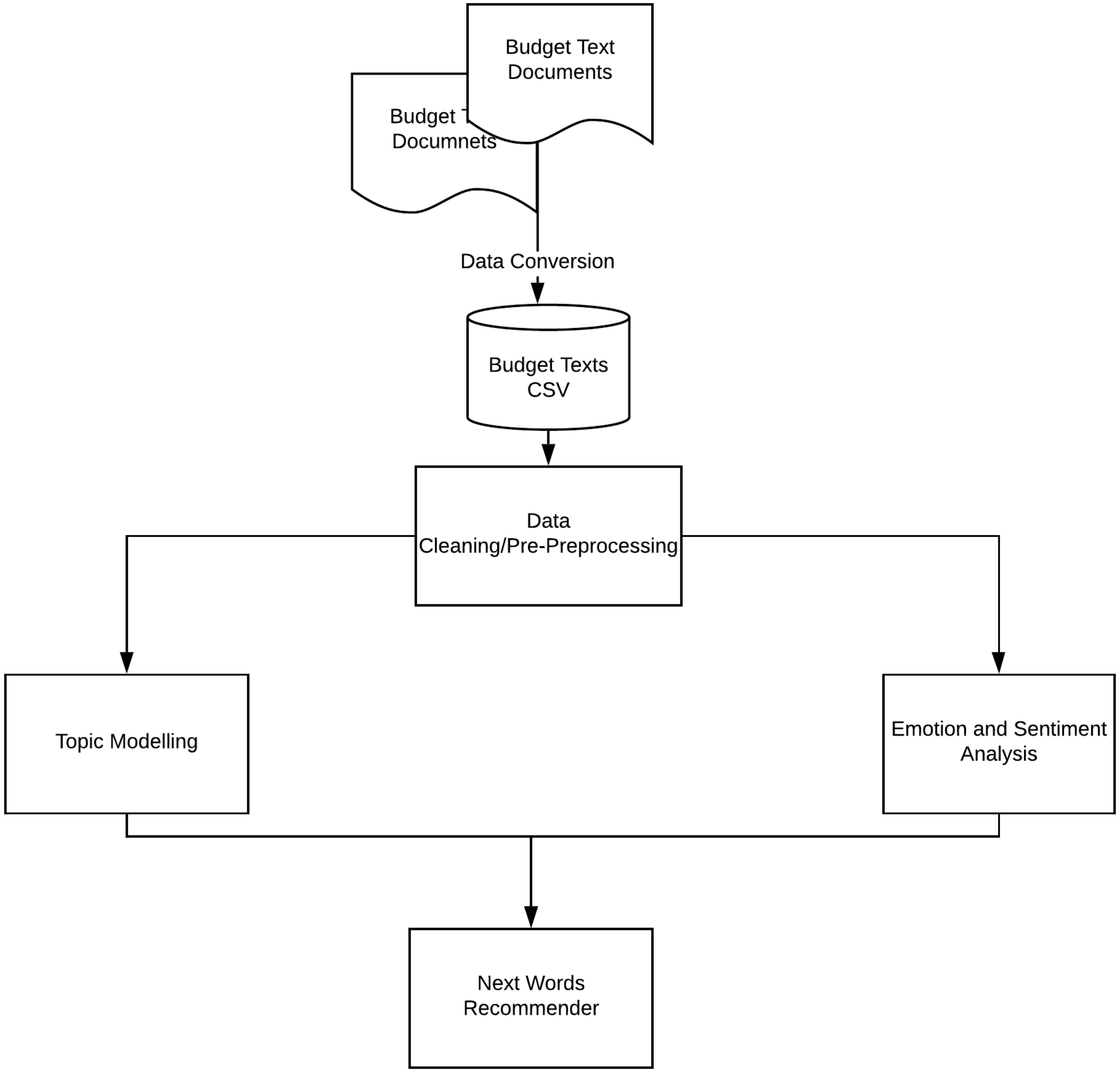
Mentors: Dr. Soumya Mohanty
Jason Jones (Guilford County)

UNC
GREENSBORO

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

Overview of the Project



Goals

- Understand the budget text data according to different counties, and their relationships, similarities/dissimilarities.
- Data Cleaning/Pre-processing: Removing stopwords, unwanted words, and lemmatizing the texts for further analysis.
- Topic Modelling of the textual data. Compare how the important topics in budget documents has changed with time (From 2009 to 2019).
- Emotion and Sentiment Analysis of the budget texts to draw up public's emotional engagement over the years.
- Next words recommender for the texts in budget when searching.

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

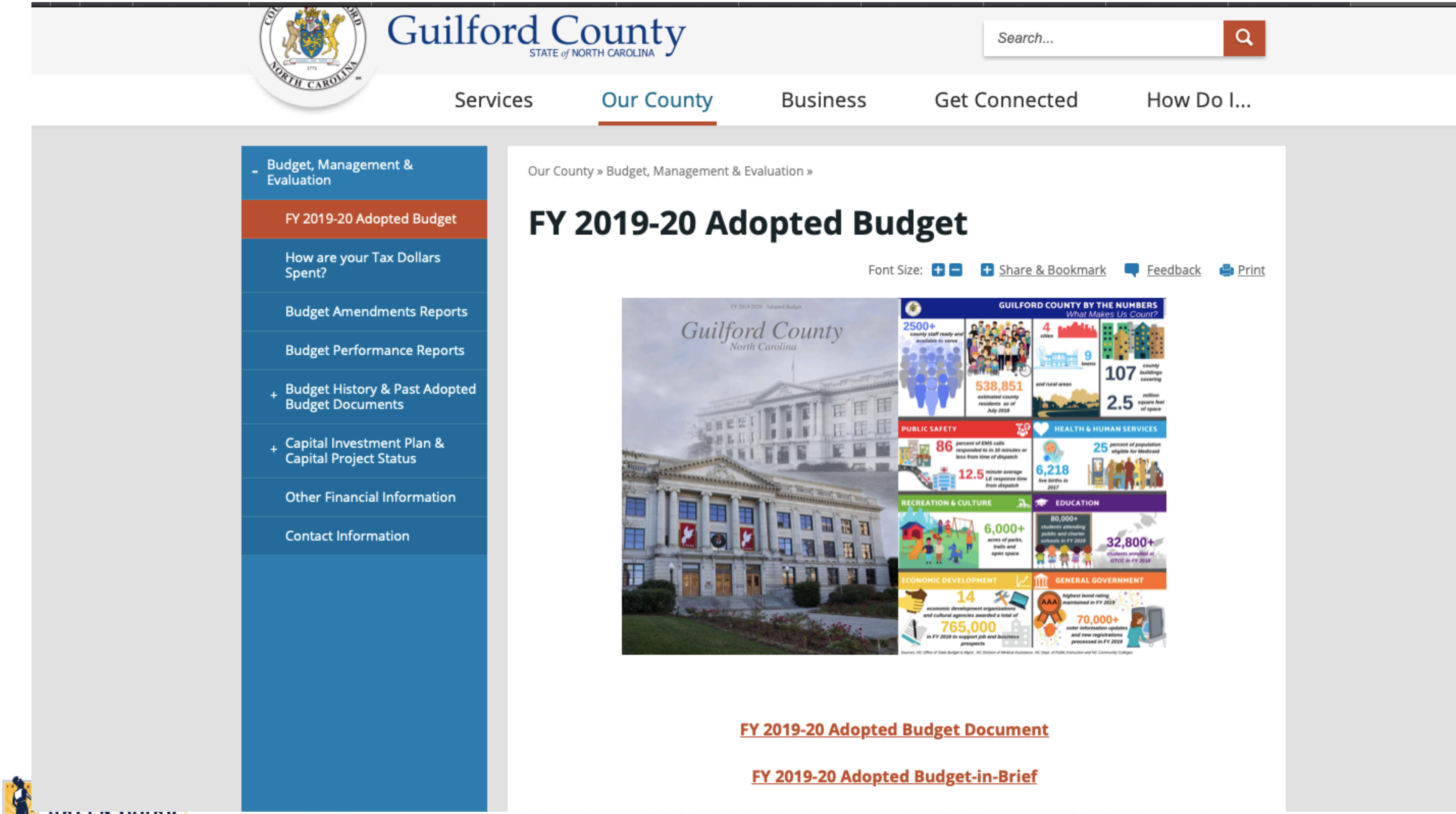
Individual Tasks Done

- Sultan Al Bogami
  1. Collected Budget Documents from all the different Counties websites and other sources.
  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. Combine all the csv datasets from all the counties, and assign labels to identify the counties.
  3. Analyze the combined data sets to identify data dictionaries and volume.
- Miguel Gasper Utrera
  1. Analyze the Datasets individually and keep the consistent data structure for all the counties.
  2. Started looking into how topic modelling works, and find resources for topic modelling.
- Unnati Khivasera
  1. Organize and Coordinate data and documents for all the team members to access them when required.
  2. Research on finalizing suitable approach /techniques used for Emotion and Sentiment analysis:
- Akash Meghani
  1. Collect Emotions csv data from the budget text documents.
  2. Carry out individual analysis of the county documents to discover emotions in words.

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 »

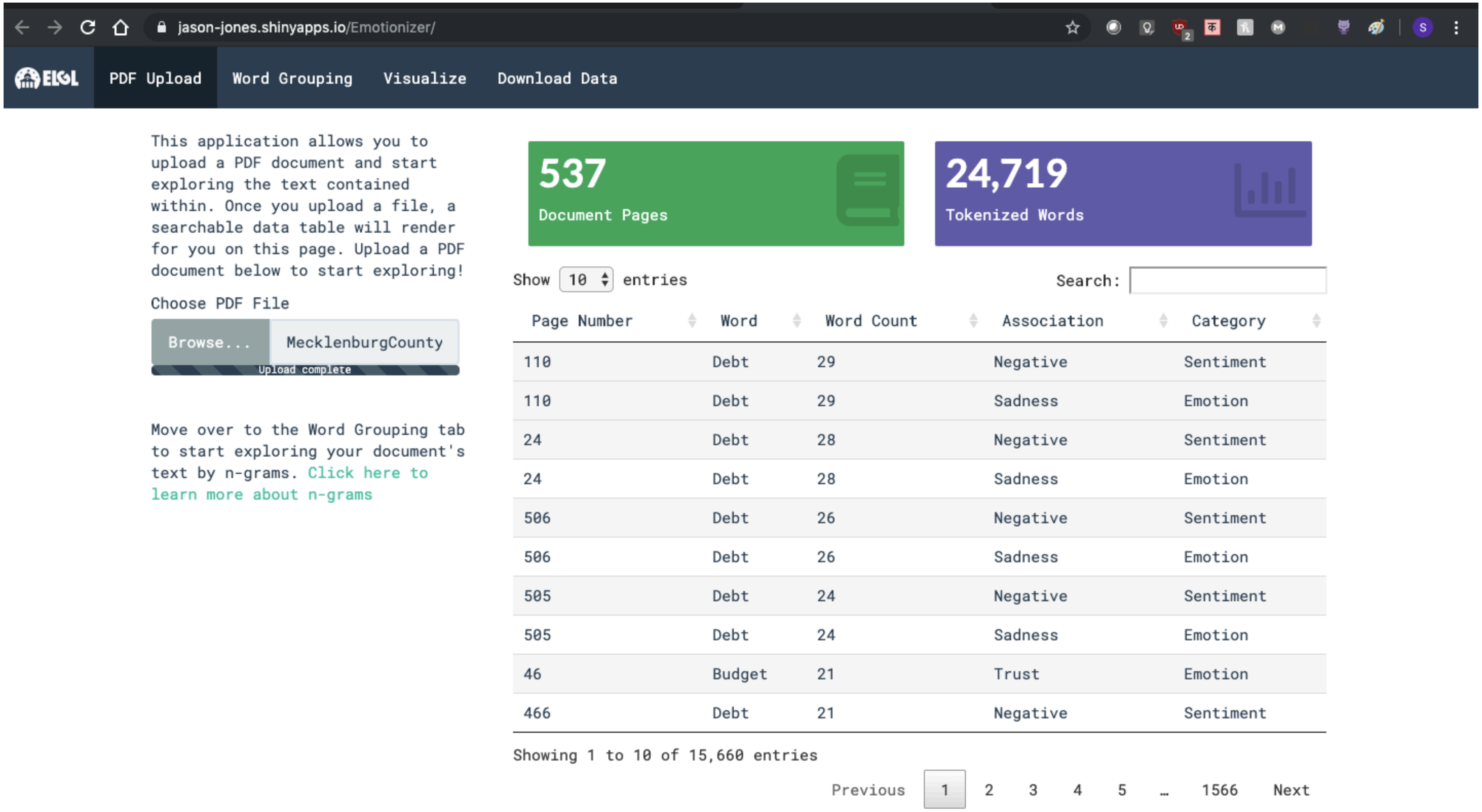
FY 2019-20 Adopted Budget

Font Size: + − | Share & Bookmark | Feedback | Print

FY 2019-20 Adopted Budget Document

FY 2019-20 Adopted Budget-in-Brief

Data Conversion



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

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

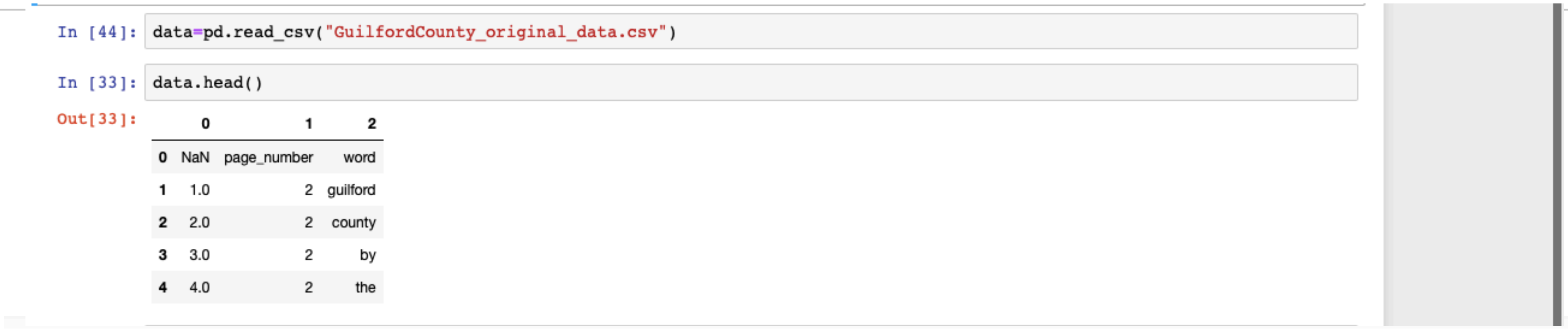
Show 10 entries

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

Data Transformation

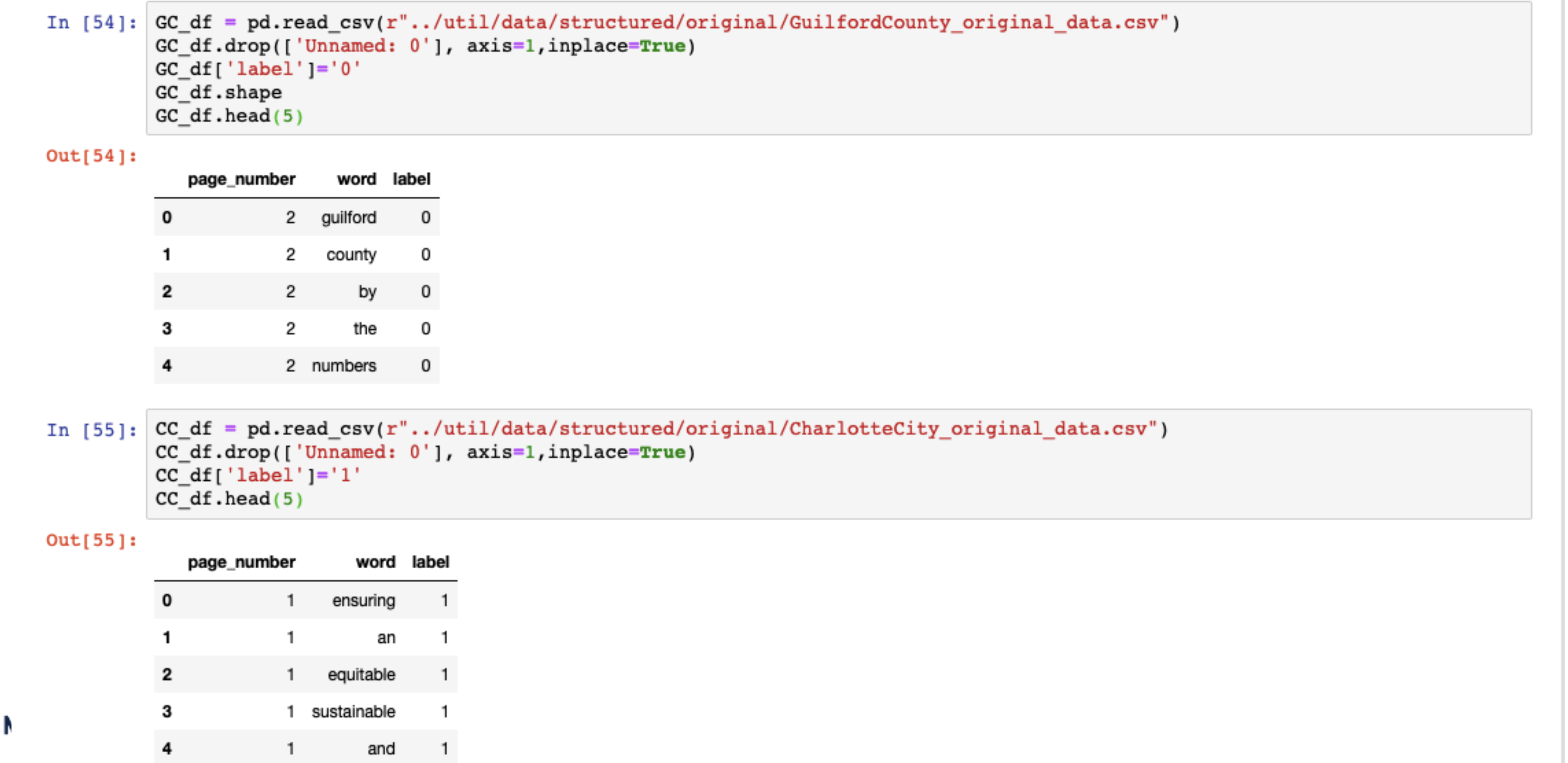


```
In [44]: data=pd.read_csv("GuilfordCounty_original_data.csv")
```

```
In [33]: data.head()
```

Out[33]:

	0	1	2
0	NaN	page_number	word
1	1.0	2	guilford
2	2.0	2	county
3	3.0	2	by
4	4.0	2	the



```
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]:

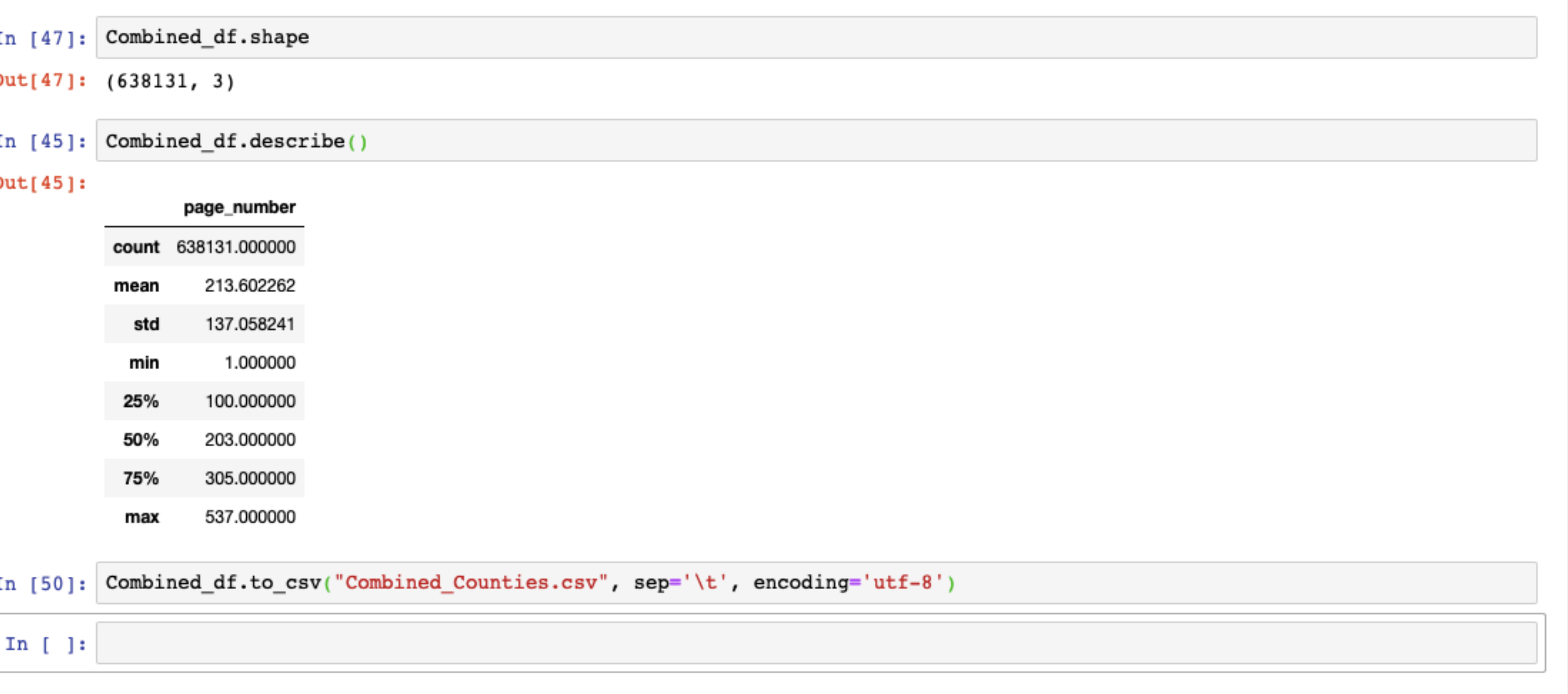
	page_number	word	label
0	2	guilford	0
1	2	county	0
2	2	by	0
3	2	the	0
4	2	numbers	0

```
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)
```

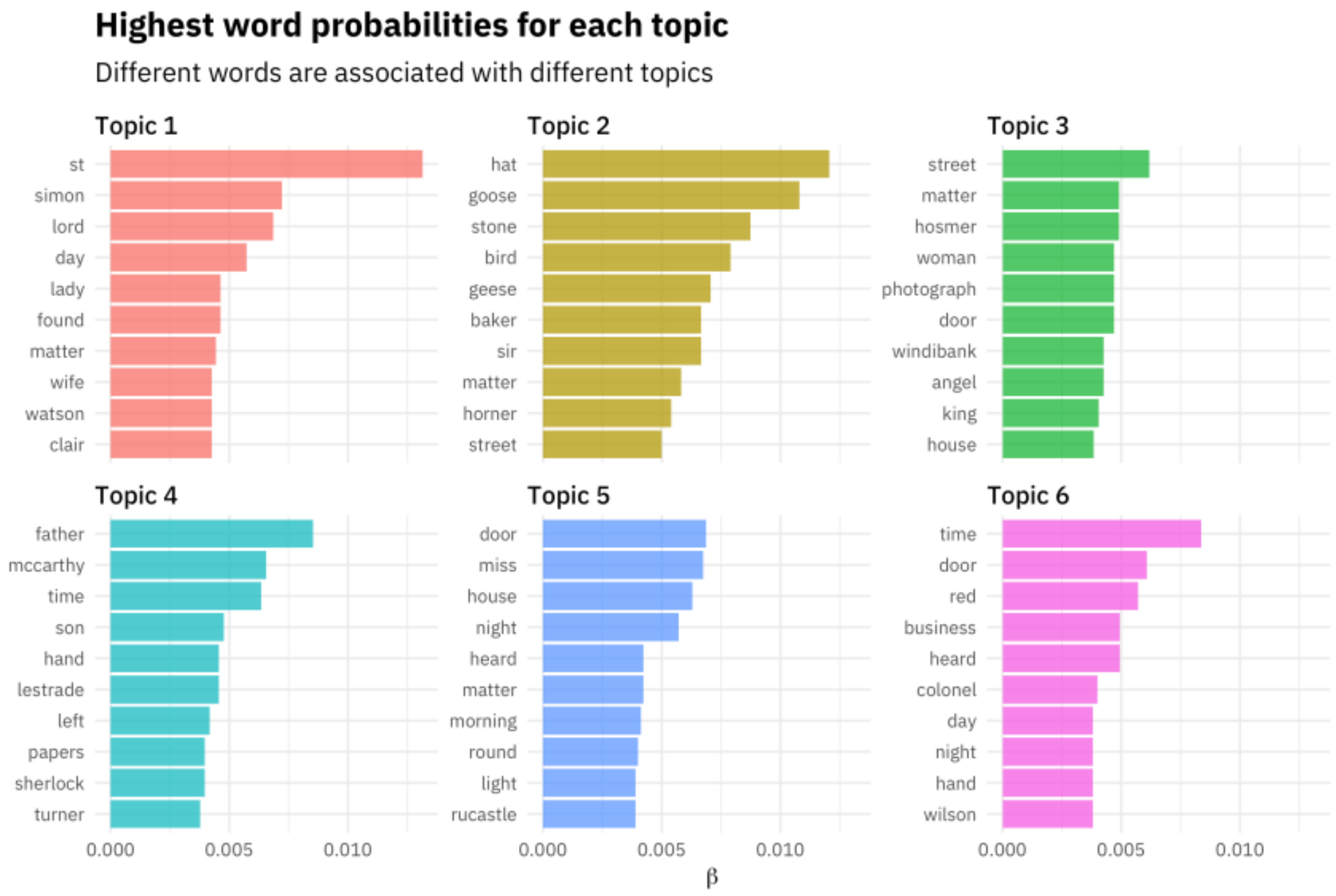
```
In [45]: Combined_df.describe()
Out[45]:
```

	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 [ ]:
```

Topic Modelling

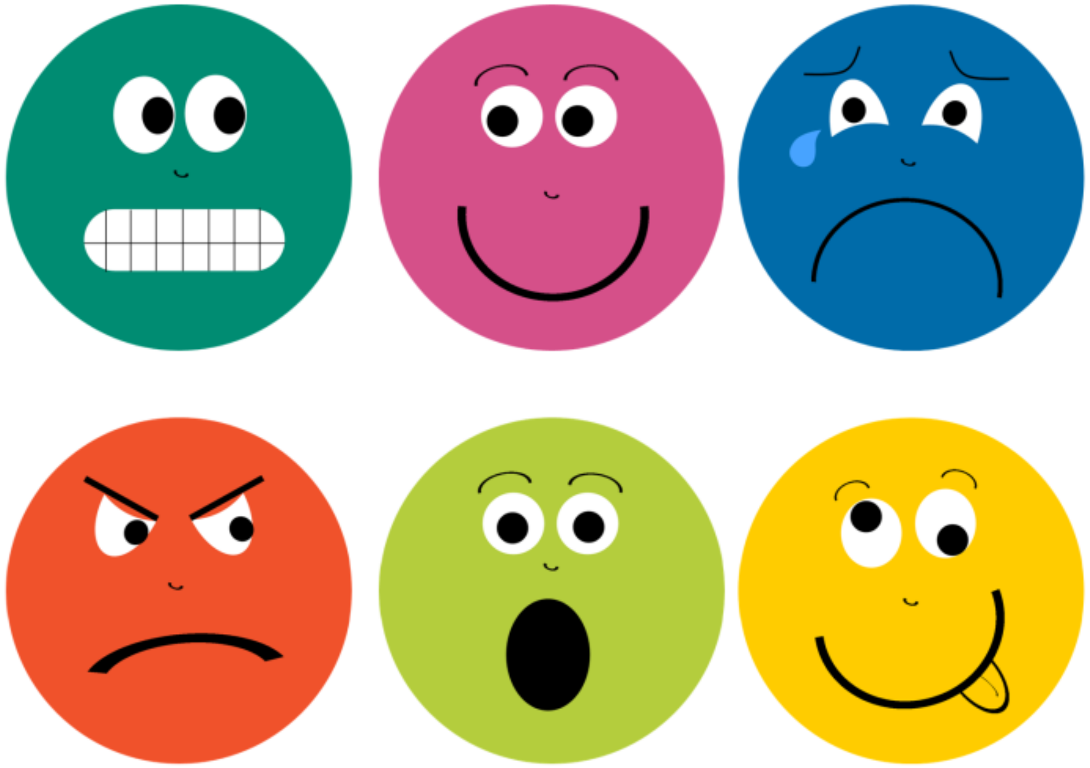


Highest word probabilities for each topic

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.

Emotion And Sentiment Analysis

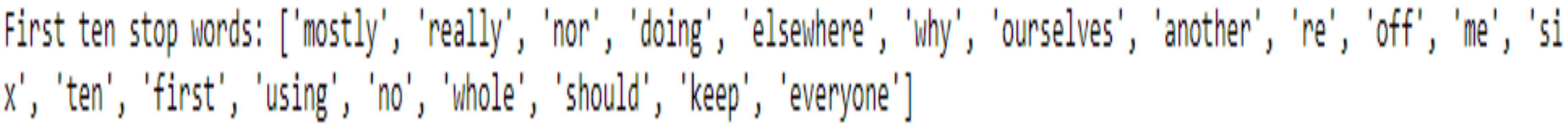
- 1) Tokenization
- 2) Cleaning the data
- 3) Removing stop words
- 4) Normalization
- 5) Supervised learning/Lexicon based approach

Emotion And Sentiment Analysis

1) **Text classification using spacy python package**

2) **Number of stop words in the list : 326**

3) **First ten stop words in Spacy:**



```
First ten stop words: ['mostly', 'really', 'nor', 'doing', 'elsewhere', 'why', 'ourselves', 'another', 're', 'off', 'me', 'si
x', 'ten', 'first', 'using', 'no', 'whole', 'should', 'keep', 'everyone']
```

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