CSC 405/605 Fall 2019

Budget Text Analysis

Progress Report Document

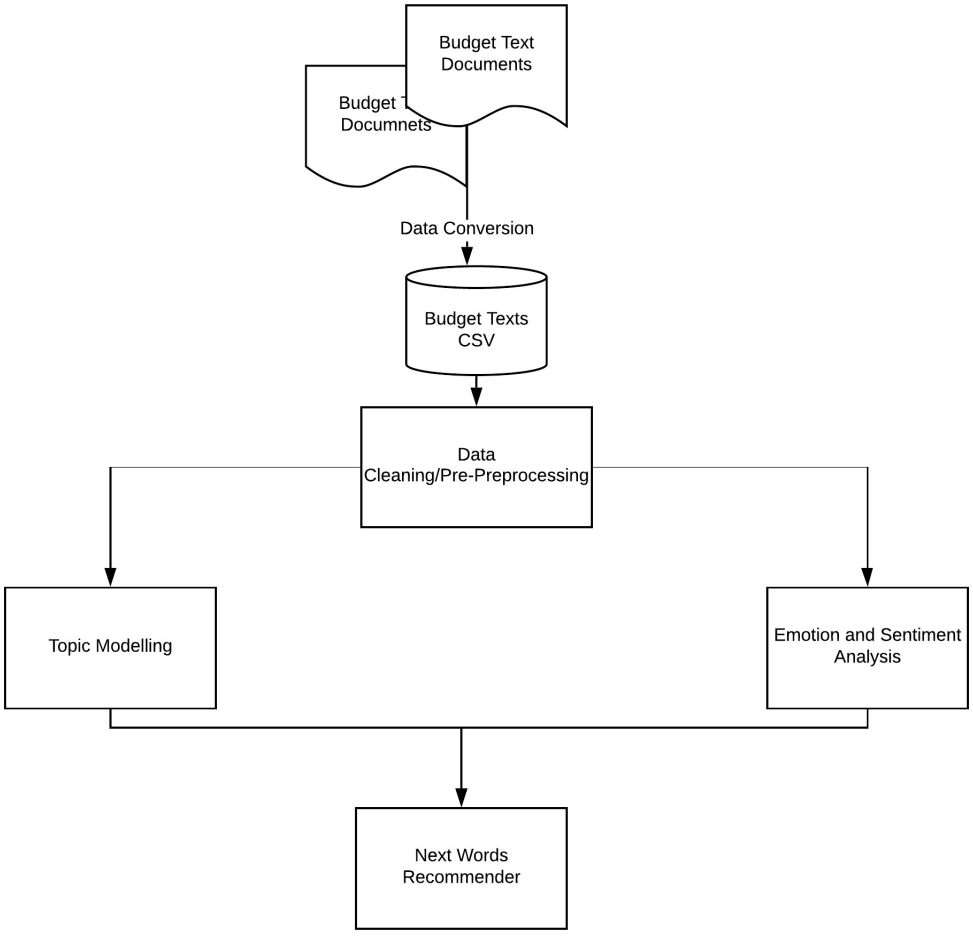
Version 1.0

Prepared By:
Akash Meghani
Unnati Khivsera
Naseeb Thapaliya
Sultan Al Bogami
Miguel Gaspar Utrera

I. Introduction

i. Project Overview:

In this project we will be analyzing Adopted budget text from the different counties of the State of North Carolina. The 'Adopted Budget Plan' is the **annual budget approved by the Board of Supervisors for the fiscal year which runs from July 1 through June 30**. By applying advanced text analysis methods, such as Topic Modelling and Sentiment Analysis, our team is hoping to extract meaningful information from each budget document individually and collectively. The team is also aiming to build a recommendation engine that may assist with the auto recommending next texts of these budget documents in the future. Below, we can see the brief project overview, which includes converting the budget text documents to pdf version, data cleaning/preprocessing and then analyzing them with topic modeling and sentiment analysis.



ii. Goals

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

II. Data Overview

i. Data Source:

The budget texts will be fetched from the following counties and cities:

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

The Budget documents (.pdf) obtained from respective counties/cities websites and is converted to two types of csv files as shown below for City of Charlotte:

a. Simple tokenization

```
"","page_number","word"
"1",1,"ensuring"
"2",1,"an"
"3",1,"equitable"
"4",1,"sustainable"
"5",1,"and"
"6",1,"resilient"
"7",1,"charlotte"
```

b. Emotion categorization

```
1  "","page_number","word","sent_count","sentiment","category"
2  "1",72,"Budget",37,"Trust","Emotion"
3  "2",292,"Maintenance",28,"Trust","Emotion"
4  "3",300,"Sewer",28,"Disgust","Emotion"
5  "4",276,"General",25,"Positive","Sentiment"
6  "5",276,"General",25,"Trust","Emotion"
7  "6",71,"Budget",23,"Trust","Emotion"
```

ii. Data Dictionary

As shown above, the csv data set is of 2 dimensions when loaded into a pandas data-frame. The data is row x column format, with three columns of index, page number and words.

The index consists of only integer values and is of type 'int', as well as page number. The words extracted are of "string" type. And, the analysis will be carried out on the word's column. When all the datasets from all the counties were combined it was observed that the total number of rows i.e. words is 638,131.

II. Tasks Completed

a. Sultan Al Bogami

1- Documented business requirements.

2. Established, organized and maintained communications between contributors using Discord platform.

3- Collected the budget documents from all organizations.

4- Converted the budget documents from pdf to csv (tokenization) and had them ready on GitHub.

5- Helped assigning issues to the team members as well as creating milestones for the project.

6- Started integrating the project using Travis CI, which is a continuous integration tool.

b. Naseeb Thapaliya

1. Initialized and set up the GitHub structure with all the necessary components.

2. Combined all the csv datasets from all the counties and assign labels to identify the counties.

3. Analyzed the combined data sets to identify data dictionaries and volume.

4. Updated the readME.md of GitHub, with all the requirements.

5. Worked on the project presentations, with including all the necessary Figures, also adding my part of tasks in presentation.

c. Miguel Gasper Utrera

1. Analyzed the Datasets individually and keep the consistent data structure for all the counties.

2. Started looking into how topic modeling works and find resources for topic modeling. (Gensim)

3. Started looking into how the next word recommender will work and how it can be implanted in python.

d. Unnati Khivasara

1. Organized and Coordinated data and documents for all the team members to access them when required.

2. Researched on finalizing suitable approach /techniques used for Emotion and Sentiment analysis.

e. 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.

3. Researched about different python packages for Natural language processing (NLTK, TextBlob, Spacy).

4. Text classification using spacy python package.

5. Removed all stop words and found a filtered list for one file.