**Budget Text Processing**

1. **Introduction**

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**. The scope of this project is Adopted Budgets for FY 2019-2020 each.

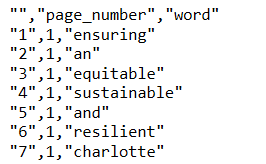
1. **Data:**

The budget texts will be fetched from the following counties or cities as mentioned below:

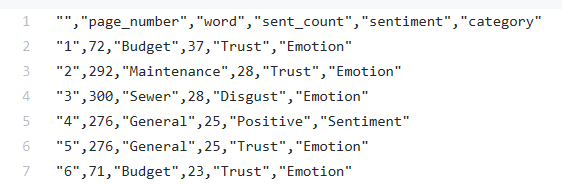
* City of Charlotte
* Mecklenburg County
* Wake County
* City of Raleigh
* Guilford County
* City of Durham
* Durham County

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

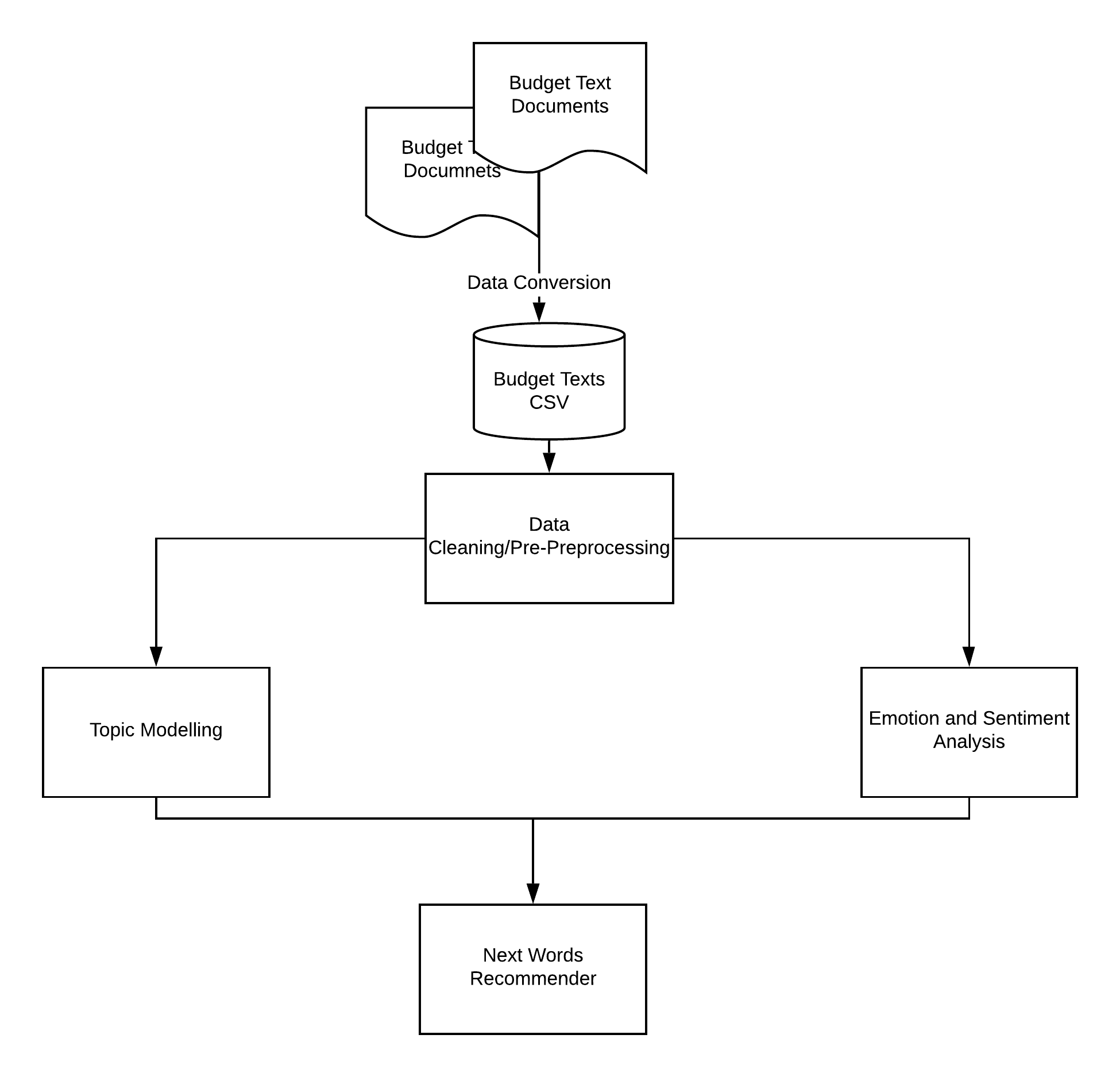
1. Simple tokenization



1. Emotion categorization



1. **Project Overview**



1. **Goals**
2. Understand the budget text data according to different counties, and their relationships, similarities/dissimilarities.
3. Data Cleaning/Pre-processing: Removing stopwords, unwanted words, and lemmatizing the texts for further analysis.
4. Topic Modelling of the textual data. Compare how the important topic in budget documents has changed with time (From 2009 to 2019).
5. Emotion and Sentiment Analysis of the budget texts to draw up public’s emotional engagement over the years.
6. Next words recommender for the texts in budget when searching.
7. **Tasks**
8. 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.
9. Naseeb Thapaliya  
    1. Combine all the csv datasets from all the counties, and assign labels to identify the counties.  
    2. Analyze the combined data sets to identify data dictionaries and volume.
10. 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.
11. Unnati Khivasara  
     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.
12. 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. Text classification using spacy python package.

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