**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.

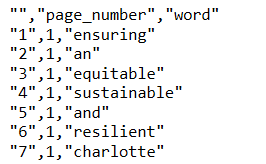
1. **Data:**

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

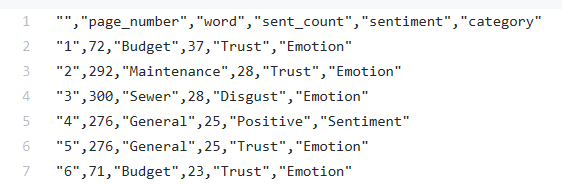
* City of Charlotte
* Mecklenburg County
* Wake County
* City of Raleigh
* Guilford County
* City of Durham
* Durham County

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. **Goals**
2. Understand the budget text data and analyzing its scope.
3. Data Pre-processing (Clean data to required format so that only the relevant data is used).
   * **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
   * Removing words that have fewer than 3 characters.
   * Removing **stop words**.
   * **Lemmatizing** words— words in third person are changed to first person and verbs in past and future tenses are changed into present.
   * **Stemming** words— words are reduced to their root form.
4. Topic Modelling of the Budget Texts.

In natural language understanding (NLU) tasks, there is a hierarchy of lenses through which we can extract meaning — from words to sentences to paragraphs to documents. At the document level, one of the most useful ways to understand text is by analyzing its topics. The process of learning, recognizing, and extracting these topics across a collection of documents is called topic modeling.

1. Emotion and Sentiment analysis of the Budget Texts.
2. Next words recommender for the texts in the budget.