**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.

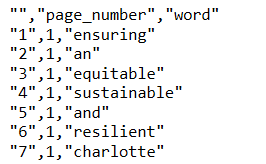
1. **Data:**

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

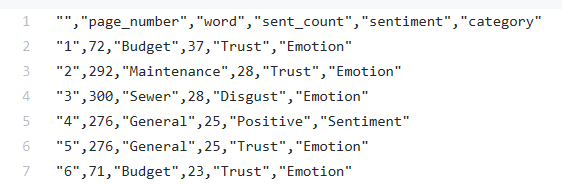
* City of Charlotte
* Mecklenburg County
* Wake County
* City of Raleigh
* Guilford County
* City of Durham
* Durham County

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.

1. 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.

Sentiment analysis and emotional analysis are two key methods experts use to quantify audiences’ emotional engagement. Lets take an example of a blog/content writer.

As content marketers, they tend to care quite a lot about their readers’ feelings. Did they make it to the end of the article? Did they laugh to that joke? Were they bored at some point?

Those are all questions they always wish to get answers to.

Sending surveys to the readers with those questions at the end of every blog post is not possible.

Fortunately, there are a few tricks that can help to bring them a bit closer to their customers’ heads: emotional and sentiment analytics. If they can manage to uncover how people feel about their content, they can easily make it perfect. Do more of what they like, and do less of what they hate.

1. Next words recommender for the texts in the budget.

Recommendation engine: is a software tool that generates recommendations to users based on their past preferences.

So, by analyzing the frequently used group of words, it’s possible to suggest a word that a user may enter along with the already input text.

1. **Tasks**
   1. Data segregation: Akash, Sultan
   2. Data Cleanup: Naseeb, Unnati, Akash, Miguel
   3. Data Organization: Akash, Miguel
   4. Research on finalizing suitable approach /techniques used for analysis: Naseeb, Unnati, Miguel
   5. Preparing Documents: Unnati, Sultan
   6. Preparing Document Term Matrix: Sultan, Naseeb
   7. Analysis : Everyone