CSC 405/605 Fall 2019

**Budget Text Analysis**

Project Review Document

Version 3.0

Prepared By;   
 Naseeb Thapaliya

Akash Meghani  
 Unnati Khivsera  
 Sultan Al Bogami  
 Miguel Gaspar Utrera

**Tasks**

Sultan Al Bogami

1. Reconstructed the data hierarchy for the project:

a. Collected the budget documents from 2019-2008.

b. Compressed them to reduce the size of the files.

c. Converted the PDF files into CSV files.

d. Standardized the naming scheme of the data files.

e. Organized the data into folders.

f. Uploaded the data to the project repository.

2. Preprocessed the data for FY2020:

1. Convert the csv files into data frames.
2. Label the data frames adding “organization” column.
3. Combine the data frames and insert a “year” column.
4. Drop unnecessary columns in the combined data frame.
5. Perform text normalization.
6. Produces a single csv file for the year in question.

3. Started the exploratory data analysis.

Miguel Gaspar Utrera

1. Did some analysis on the budget for Durham city for 2019 and 2020
2. My goal was to see how similar the two budget documents were
3. Using NLTK, genism and TFID I was able to use the most frequent words from each document and see their similarity
4. They are 57% similar, I can conclude the budget documents are reused and that is why they are very similar