CS 410 Project Proposal
Andrew Vamos (avamos2@illinois.edu)
Misha Ryabko (mryabk2@illinois.edu)
Benjamin De Vierno (bid2@illinois.edu)

Requirements

1) Progress made thus far, (2) Remaining tasks, (3)Any challenges/issues being faced.

1. Progress made thus far.
   We have written a python script to generate text data for the serial inverted text generator in python. The script simulates the Poisson distribution of frequency of words found in natural language in terms of frequency of words. The script creates a dataset with a normal or even distribution of words. A key criteria of the script for this project is to create a lexicon mimicking the distribution of words in a language.

   The script for simulating the dataset can be found at:
   generate_data.ipynb

   We have made and tested the serial text inversion process. The script can be found at inverted.py in the repo.

2. Remaining tasks for this project will be
   - Performing the benchmarking and testing.
   - Writing the final report.
   - Writing the text inversion process in Triton.

3. (3)Any challenges/issues being faced.
   The dataset they used in the paper used as reference for our project required a waiver to get access. https://ir.dcs.gla.ac.uk/test_collections/access_to_data.html The dataset required us to pay a fee for usage of the dataset which was beyond the teams stipulated budget of this project. Instead we decided to use a randomly generated dataset which should be similar in frequency to the paid dataset.