Project Proposal for 2023 Fall CS410: Course Project
Topic: Implementation of Compression Text Classification Method
Team member: Kiyotaka Kokubun (kokubun3@illinois.edu)

I will work on Theme: 5 free topic.
The Topic is Implementation of Compression Text Classification Method, which is recently explained in the paper of “Low-Resource” Text Classification: A Parameter-Free Classification Method with Compressors” (https://aclanthology.org/2023.findings-acl.426/). The process is very simple that they used gzip library in python to compress texts to calculate similarities. The method in the paper is very accurate comparing to recent deep learning methods with extremely low computing power.
In the correspondence of the class, it is a new possible text similarity calculation method with information compressing. Both text similarity and compression are topics covered in this class.
I am going to try to implement this method in simple Python script first with a web API or dataset with enough amount of text. Then I will create pseudo search engine environment so that we can test our query. If I have enough resource, I will implement it on AWS environment which is accessible for everyone.
I will mainly use Python for this project.
I am estimating first 5 hours for testing and choosing dataset or API for text data, then I will spend 10-15 hours to implement the ranking program. I am willing to spend additional hours for cloud implementation of this system. Also, I will test algorithm in the class, like BM25 to compare to this new method, and hopefully I will implement the comparison algorithm in the script/application as well.