COSC 320 – 001

Analysis of Algorithms

2022/2023 Winter Term 2

Project Topic Number : 1

Title of project :

Keyword replacement in a corpus

Group Number 12 :

Anna Ciji Panakkal

Erfan Kazemi

Sahil Chawla

Abstract : For this milestone, we have formed a group and picked a topic. We have deepened our understanding of the topic we chose and assessed what each of us will be doing for this project and talked about important logistics such as expected complexities, difficulties and responsibilities, as well as discussed edge cases, what programming languages to use and discussed data collection strategy.

Problem Description : For this topic, we will be discussing the use of text analytics in replacing a set of keywords with a given set in documents. For Example, the use of abbreviations on Twitter, we should find and replace all abbreviations in tweets with their proper phrases or keywords. The challenge is to create an algorithm that, despite the possibly lengthy list of keywords and acronyms, can quickly and accurately locate and replace these terms in a high volume of tweets. While there is a naive strategy, that would be to replace every element in the condensed list that is found in a tweet by your algorithm, we are aiming to create a more effective algorithm in lieu of the naive technique.

Edge Cases :

- too much data and keywords
- wording of tweets that may make it difficult to distinguish and replace the keywords accurately

Expected complexities : Between O(n logn), O(n^2)

Dataset Collection : We will be using the dataset provided by the professor that includes 4 million tweets to implement and test the algorithm that we create.

Programming Language : We are planning to implement our algorithm using Java as it is better for the large volume of data that we have to work with and all three of us are knowledgeable in the language

Task Separation and Responsibilities:

- Anna - Problem formulation and pseudo code, implementation
- Erfan - Formal analysis, implementation
- Sahil - Data structure and analyzing, implementation

Unexpected Cases/Difficulties :

- Keyword not present in the documentation (Solution : if its a common keyword, add to database)
- Chosen data structure not working out. (Solution : talk to TA to see if there is something wrong in our end, otherwise try to choose a new one with their help)
- One group member not finishing allotted task on time (Solution : Have weekly meetings/check-ins to ensure everyone is completing their work on time)
- Programming language barrier - java not good for the algorithm or performance of devices (Solution : switch to another language)