COSC 320 – 001
Analysis of Algorithms
2022/2023 Winter Term 2

Project Topic Number: #1

Keyword Replacement

Group Lead:

Sanjith Senthil

Group Members:

Issa Hashim, Cole Van Steinburg, Sanjith Senthil

Abstract

In this milestone, we got the opportunity to introduce ourselves and get to know more about each other. We chose the topic and began distributing the work between us. We completed the headers of the project proposal. Finally, we discussed and became familiar with the algorithms we would be using.

Problem Description

Keyword replacement is a crucial task for text analysis and natural language processing applications because it can improve their accuracy and relevance. The problem is to identify all the keywords in a text and replace them with their respective replacement word or phrase. It needs to be done in such a way that the meaning and context of the sentence is not changed while replacing certain words with more precise or relevant terms.

For example, consider the short paragraph below (in this case a tweet):

“Imho, I think that the earth is flat, but I can understand why some people think it is round lol. Anyone who still thinks that the earth is round should take a trip to the edge and see for themselves asap.”

The above paragraph should be converted to:

“In my honest opinion, I think that the earth is flat, but I can understand why some people think it is round laugh out loud. Anyone who still thinks that the earth is round should take a trip to the edge and see for themselves as soon as possible.”

‘Imho’ was replaced with ‘In my honest opinion’, ‘lol’ was replaced with laugh out loud, and ‘asap’ was replaced with ‘as soon as possible’ in the paragraph.

Edge Cases

There are a few edge cases to consider, as the underlying problem is effectively text matching and replacement.

- Case sensitivity of the letters of the keyword.
- Keywords followed by a punctuation sign such as a full stop, comma, question mark, etc.
- Keyword within a word.
- Irregularities in spacing between words in the paragraph text.
- Number of words in the paragraph text is zero.

Expected complexities

Using the naive approach, that is for each word in the paragraph text, the algorithm checks for all of the elements in the abbreviated list and replaces them. The time complexity would be O(n * m) where n is the number of words in the paragraph text and m is the number of elements in the abbreviation list. The auxiliary space complexity would be O(1).

A better approach is to use a hashmap for the abbreviation list. For each word in the paragraph text, we can check if it needs replacement in constant time. This means that the time complexity for replacing a whole paragraph is O(n), where n is the number of words in the paragraph text. For the auxiliary space complexity, the only additional space our algorithm uses is to store the list of replacement words in a hashmap, which would be O(m), where m is the length of the replacement word list.

Dataset Collection.

The dataset will be provided by the professor. It is said to contain approximately 400 million records, based around reviews from the Google Play store and commonly used abbreviations. The format of the data is said to be in CSV.

Programming Language

The programming language that we are going to use for implementation of our ideas is Python.

Timelines

We will use the deadlines set by the project milestones as a general timeline. The first milestone is due on February 6th, the second milestone on February 27th, the third milestone on March 20th, and the fourth milestone on April 10th. We will finish each milestone at least five days before its deadline. With potential to adjust as and when required, we will commit more than 3 hours each week to the project as a baseline.

Task Separation and Responsibilities.

We will divide the work among all group members equally, meaning that all members will do every task of each milestone together. We will collaboratively work on algorithm design, analysis, implementation, empirical evaluation, and choice of data structure. Each of us will also conduct our own research and discuss our findings together. In addition, Issa and Cole will be responsible for organization of group meetings and Sanjith will be responsible for communicating with the instructor and TA's.

Unexpected Cases/Difficulties.

We have not encountered any issues for design and implementation of our work at the moment. As with any project, there are bound to be unforeseen circumstances which can negatively affect project timelines such as team members falling sick and not being able to work on the project during a milestone. Therefore, to account for unexpected difficulties, we will aim to complete our tasks 5 days prior to the official deadlines. This will allow us some slack time to deal with unexpected cases/difficulties as they arise.