MMU

MULTIMEDIA UNIVERSITY

Final Year Project

Time Series Based Summarization

Mak Yen Wei

Bachelor of Computer Science

(Data Science)

Jan 2023

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

90 percent of engineering is searching on google! Bugs solving being one of the crucial task for a developer. While it's very important, solving bugs is tedious when you have to look through forums after forums, stackoverflow after stackoverflow just to solve a simple linear equation problem in python. It is very common to see developers spending hours just to find the exact solution on stackoverflow, This can happen in possible two situation.

Solving bugs in the software world is not straight forward at all, there are way too many co-dependent libraries and packages that are used in the software world across langauges, frameworks, and platforms. At times, it's a fresh issue that occured based on the latest release of a particular software. On the other hand, it could be a typo in the code that is causing the problem. Then, there are times where the problem is caused by the local environment of the developer. There is a non-ending list of possibilities that can cause the problem.

Therefore the aim of this research is to solve these two problems using text processing methods, rankings methods and lastly summarization models to ease the process of finding the exact solution to the problem for developers.

1.1 Problem Statement

Throughout this research work, we aim to solve 2 of the major issues outlined. The first being that developers are not able to find the exact solution to the problem. The second being that developers are not using the right keywords to search for the solution.

First being that the developer is not using the right keywords to search for the solution. This can lead to miss-interpretation of the problem and the solution queried by the forum. Since there are millions of developers using stackoverflow, you can expect the amount of questions and answers to be very similar to each other when the keywords are vauge and not specific. Without using the right keywords, developers will be spending hours and hours trying to locate the exact solution that fits well to his situation and try to apply it to his problem. This can be extremely frustrating and time consuming.

Second being that the developer is not able to find the exact solution to the problem even the keyword is correct. The software world now is very much complicated, every language, packages or libraries are co-dependent to each other, this creates a problem where it's very time-consuming to find the root solution to the problem. Developers ended up going into an endless rabbit hole just to find dependencies and dependencies of dependencies to solve the problem. Such problems has been acknowledged by the software industry and they have been trying to solve this problem by creating a dependency graph for each software. However, this is not a perfect solution as it is not able to solve the problem of finding the exact solution to the problem.

1.2 Objective

The unified objective of this research is to solve the above mentioned problems by using text processing methods, rankings methods and lastly summarization models to ease the process of finding the exact solution to the problem for developers. However the objectives of this research can be broken down into 4 main objectives.

The first objective is to ease the process upon hours of researching the bug and working your way through the forum communities. This can be challenging for the developers to do it themseleves manually as you not only have to search for multiple forum pages but also have to read through the entire forum to find the exact solution to the problem. To counter that, we aim to automate this task by using scraping technologies to scrape all the related forums alongside with all of the metadatas we can find (Eg: Post Voting Count, Post Comments and much more) we can find online and save it for future use.

The second objective is to ease the process of finding the exact solution to the problem – this is the main objective of this research. String matching algorithms will be applied to find the best matching posts that will bit to the developer use case. This will be done by using the keywords that the developer has used to search for the solution.

The third objective is to rank the posts based on the relevancy of the post to the developer's use case. There a couple of determining factors that we should consider in ranking the posts. The ranking of the post will be determined by the number of keywords that matches the post, semantic similarity between the post and the developer's use case, the number of upvotes the post has received, the number of comments the post has received and the number of views the post has received and finally sentiment analysis will also be involved. The ranking will be done by using a weighted sum of the above mentioned factors.

At last, the final objective of the research work is to summarize all of the top ranking post's comments, answers and answers comments. Deep learning based summarization models will be used to summarize the comments and answers, in the hopes where the developer can get a quick overview of the solution to the problem.

1.3 Scope

By the end of this research work, we aimed to propose a framework that solves the problems stated above. A working version of the framework will be engineered and deployed for public usage by the end of the project (FYP2).

The framework will be able to scrape the stackoverflow forums, rank the posts based on the relevancy of the post to the developer's use case and summarize the top ranking posts comments, answers and answers comments. Furthermore a web application will be developed to allow developers to interact the framework to do inferencing on the stackoverflow forums.

The technology used are as follows:

- **Web Application**: Python, FastAPI, AWS, ReactJS
- **Scraping**: Selenium
- **Text Processing**: NLTK, Spacy
- **Summarizing**: Tensorflow, Keras