Business Case

Project Title: Trend Identification on Stack Overflow

Team 30

1 Executive Summary

Stack Overflow is the most popular programming community online. Discussions on Stack Overflow often reflect the trends in computer science within a particular time window. Trend identification is important for a number of stakeholders including educational institutions, technology companies, and developers in the IT industry. However, the trends cannot be easily observed by reading through every thread. Therefore, we aim to design a software which can extract popular concepts corresponding to user-specified time window, using natural language processing (NLP) algorithms. With this software, the users can get information about the latest trends within a few clicks.

In the following sections, we will firstly provide the background information on Stack Overflow, text mining and NLP. Next, we will describe what the project is and explain why the project is important. Finally, we will describe the goals and assess risks of the project.

2 Background Information

Stack Overflow is a popular question-and-answer website for programmers. It has over 100 million visitors every month. Since 2008, there have been over 45.1 billion times a developer got help. It has a return on investment (ROI) of 179%. There are over 5000 active instances of Stack Overflow for Teams every day (Stack Overflow, 2019). The popularity of Stack Overflow makes it a potential data source for trend identification through text analysis.

Text mining is a sub-category of artificial intelligence that uses NLP algorithms to extract structured data from unstructured text in documents and databases. The structured data usually reveal critical and useful information that would otherwise remain buried in the massive amount of textual data without text mining techniques. Among the various text-processing methodologies employed by text mining, one of the most important is NLP (Linguamatics, 2019).

NLP is an area of research which aims to enable machines to understand and manipulate natural languages such as English, Spanish or Chinese. The design of NLP algorithms relies on the way humans extract meaning from natural languages, in textual or spoken form (Chowdhury, 2003). In addition to natural language understanding, NLP also aims at natural language generation, which enables computers to create natural language (e.g. summary of information) like humans (Linguamatics, 2019).

3 Project Definition

Trend identification on Stack Overflow is of interest to a number of stakeholders. However, it is difficult for a human to identify the trends by looking at all the posts on Stack Overflow. Therefore, this project aims to design a software that allows users to define a specific time frame and return the phrases that appear most frequently in all Stack Overflow threads belonging to that time frame. The software should work in the way described below.

Once the user enters a time frame, the software should perform NLP algorithms to extract all two-word phrases from the subset of Stack Overflow data corresponding to that time frame. Those two-word phrases will then be filtered and only phrases composed of adjective-noun or noun-noun will be retained. These phrases will be sorted according to their frequency of appearance. The results can be displayed as tables or graphs, depending on the user's preference. In this way the user can easily identify the most popular concepts in that specific time frame.

In addition, the user can enter multiple time frames into the software. The software will return the most popular concepts from all these time frames and show the changes in trends over time in graphs.

The software provides intuitive understanding of the trends in a particular time frame as well as the changes in trends over time, where the time frame is flexible to users.

4 Benefits

This project could lead to a number of benefits for different stakeholders including educational institutions, technology firms, and developers.

Firstly, for educational institutions, knowing the latest trends can help design courses that prepare students for the job market. According to a survey conducted in the UK, many computer science graduates consider the content of their courses and the technologies taught are outdated and do not satisfy employers' expectations (Great Britain. Department for Business and Skills, 2016). This issue could be prevented if universities are well aware of the latest trends in computer science and are able to integrate those new concepts into teaching. For example, in the top universities in the US, the most common programming languages taught in introductory courses are Python and Java, which are the most popular programming languages these days (Guo, 2014).

Secondly, for technology companies, the latest trends can provide directions for their investment. For example, machine learning, data science, and cyber security are popular concepts in computer science these days. Investing in these areas can lead to a high return. Therefore, technology companies often prioritize these areas in their investment scheme (Kim Weins, 2020). This software will help companies to easily identify the latest trends and decide where to invest.

Thirdly, for developers, the trends can reveal which tool or programming language is best suited for a particular technology framework. For example, Python, Ruby, or C++ are

associated with less open-source security vulnerabilities compared to other languages (Tung, 2020), which makes them more favorable options in minimizing security vulnerabilities. This software will help developers to quickly identify the most suitable tool for their developing purpose.

5 Goals

The goal of this project is to develop a software with a user interface where the users can specify the time window which they are interested in. By the first milestone, the software will be able to retrieve data corresponding to that time frame via Google BigQuery service. The software should perform NLP algorithms on the questions and answers of the dataset and extract two-word phrases from the text. These phrases will be filtered so that only adjective-noun and noun-noun phrases remain. These remaining phrases will then be sorted according to their popularity. By the second milestone, the software should be able to display the results as graphs. In addition, the users can specify multiple time windows and the software will perform parallel analysis on each of the datasets and return the changes in trends over time.

6 Risks

Risk	Likelihood	Severity	Management Strategies
One or more of the team members are unable to work due to illness or other unforeseen circumstances.	Moderate	Serious	At least two people will work on the same task at the same time. If one person becomes unavailable, the other can maintain the progress.
The time required to add a feature or fix a bug is underestimated.	Moderate	Serious	Start working on a feature or bug as early as possible
The back end and front end cannot be connected.	Moderate	Catastrophic	Start connecting back end and front end as early as possible, and maintain the connection with more features being added
Files and data get lost on the computers we are using.	Low	Catastrophic	Back up files using GitHub and Google Drive regularly
User attempts to process more data than the limit (1TB/month) of BigQuery (free version).	Moderate	Tolerable	Inform the user about the processing limit or purchase more processing capacity if budget allows.

Client changes the requirements, and the software needs major re-design.	Very low	Serious	Communicate with the client actively to make sure we detect changes of mind as early as possible

7 References

Chowdhury, G.G., 2003. Natural language processing. Annual review of information science and technology, 37(1), pp.51-89.

GREAT BRITAIN. DEPARTMENT FOR BUSINESS, I. & SKILLS 2016. Computer science graduate employability: qualitative interviews with graduates. London: Department for Business, Innovation and Skills.

Guo, P. (2014). Python Is Now the Most Popular Introductory Teaching Language at Top U.S. Universities. [online] cacm.acm.org. Available at: https://cacm.acm.org/blogs/blog-cacm/176450-python-is-now-the-most-popular-introductory-teaching-language-at-top-u-s-universities/fulltext [Accessed 3 Apr. 2021].

Kim Weins (2020). IT Spending by Industry. [online] Flexera Blog. Available at: https://www.flexera.com/blog/industry-trends/it-spending-by-industry/ [Accessed 3 Apr. 2021].

Linguamatics. (2019). What is Text Mining, Text Analytics and Natural Language Processing? [online] Available at: https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing#:~:text=Text%20mining%20(also%20referred%20to [Accessed 24 Nov. 2020].

Stack Overflow. (2019). Stack Overflow - Where Developers Learn, Share, & Build Careers. [online] Available at: https://stackoverflow.com/.

Tung, L. (2020). Open-source security: This is why bugs in open-source software have hit a record high. [online] ZDNet. Available at: https://www.zdnet.com/article/open-source-security-this-is-why-bugs-in-open-source-software-have-hit-a-record-high/ [Accessed 3 Apr. 2021].

MCI project First Milestone Plan

Team: 30
Project Title: Trend Identification on Stack Overflow

Milestone 1	Activities	Projected Outputs
Define the first milestone to be completed by end of week 7	List activities required to achieve 1st milestone	Define projected outputs from your work plan
By the end of week 7, we hope to have achieved the most basic yet systematic revision of our project, consisting of the functions that we have developed, over the 7 weeks. Amongst these, the most defining functionalities are that of phrase extraction via NLP and ranking by frequency of occurrence of certain phrases. We need to write and test code to extract and download a subset of Stack Overflow data from a specific time frame via BigQuery. The dataset should be imported and analysed with the NLP library we use. We need to extract adjective-noun (A-N) and noun-noun (N-N) phrases from the dataset and save those phrases for further analysis. Each of the phrase should be associated with its frequency of occurrence. The phrases should be sorted into a list according to their frequency. The time frame can vary largely, from a week to a year. In addition, if we progress well, we can work on a user interface to allow users to enter a time frame and display the resulting list. The phrases with the highest frequency can be displayed in graphs for visualization.	By week 6, decide which NLP libraries to use for extracting information from Stack Overflow corpus	Choose either Spacy in Python or DL4J in Java.
	Coding 1 || By week 6, design, write and test code to extract a subset of Stack Overflow data corresponding to a specific time frame (e.g. a week), via Google BigQuery service	Be able to download a .csv file containing all questions and answers from the specified time frame via BigQuery, and import the file for NLP analysis
	Coding 2 || (This is a succeeding activity of “Coding 1”.) By week 1 of mid-semester break, extract adjective-noun (A-N) and noun-noun (N-N) phrases from the subset mentioned above	Be able to save the A-N and N-N phrases in an object in the code or save them in a .csv file for frequency analysis
	Coding 3 || (This is a succeeding activity of “Coding 2”.) By week 2 of mid-semester break, sort the phrases mentioned above according to their frequency of occurrence	Be able to associate each phrase object with its frequency of occurrence in the corresponding time frame, sort the phrases by the frequency, and save the sorted list in an object in the code or save it as a .csv file
	Coding 4 || (This is a succeeding activity of “Coding 1-3”.) By week 7, perform the same analysis mentioned in “Coding 1-3” on a longer time frame.	Be able to expand the time frame to a year, i.e. be able to analyse a much larger dataset. The output should be an object or file containing a sorted list of A-N and N-N phrases corresponding to the specified time frame.
	Coding 5 || (This is a succeeding activity of “Coding 1-4”, and it may or may not be achieved depending on our progress.) By week 7, design a simple user interface that takes a time frame and outputs a list of sorted phrases.	The user interface should have one input area for time frame and display a portion of the sorted list containing the phrases with highest frequency. It should also allow the user to export the entire list as a file.
	Coding 6 || (This is a succeeding activity of “Coding 1-4”, and it may or may not be achieved depending on our progress.) By week 7, display the phrases with highest frequency in graphs	Be able to plot graphs (bar graphs or pie charts) showing the most popular phrases and their associated frequency.

Team Organization

Our team consists of four members – Tianjiao, Po-Yi, Keerthika, and Vishnu. Tianjiao, Keerthika, and Vishnu are in Adelaide. Po-Yi is studying remotely. We have one client – Dr. Christoph Treude from School of Computer Science. In the project, we need to write documents (e.g., business case, milestone report) while continuously working on the code. We decided when there is a document to write, we will split into two pairs – one pair will work on the document and the other will work on the code. Each member can choose the task to work on according to his/her preference. When there is no document to write, all of us will work on the code but we will still work as two pairs – one pair works on the back end (Tianjiao and Po-Yi) and the other works on the front end (Vishnu and Keerthika). Meeting organization for each week, including writing of agenda, communication with client, and recording of minutes, will be rotated between the two pairs. For example, Vishnu and Keerthika do it for odd weeks; Tianjiao and Po-Yi do it for even weeks.

Risk analysis: since we always work in pairs, if one person in a pair cannot keep up due to unforeseen circumstances, the other person can take over the work. If one person gets stuck on a problem, the other can help. If neither of them can keep up, which is very unlikely, one person from the other pair will take over their work.

Communication Plan

Since we have a mixture of onshore and offshore students, all the team meetings will be held online via Zoom or Facebook. The time difference is 1.5 hours (after daylight saving time ends) between Po-Yi and the rest of us. The media for communication we use include Email, Facebook, Zoom for textual or verbal communications, and Discord, Google, GitHub for file sharing.

Client meetings are held once a week via Zoom on Thursdays, where the client will review our progress, provide feedback, and specify further requirements. After the client meeting, we will have a meeting within the team to discuss the goals and objectives for the following week. While we are working on the code, if one has any technical difficulty, he/she can post the question on any of the media we use, and others will try to help resolve it. In the middle of two client meetings, we will have a second meeting to discuss our progress and make sure we can achieve the goals for that week. Before the next client meeting, we will have another meeting to gather the questions we need to discuss with the client. In total, we have three meetings between every two client meetings and written communications throughout.