MCI project First Milestone Report

Team number: 30
Project Title: Trend Identification on Stack Overflow

Milestone 1	Activities	Planned Outputs	Achieved Outputs
Restate the milestone from your Draft plan .	Restate the key activities from your draft plan.	Restate the planned outputs from your draft work plan.	Outline the actual outputs compared to what was projected (or type “same as planned”)
By the end of week 7, 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.	Same as planned
	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 from BigQuery, and import the file for NLP analysis	Same as planned. But we are now using BigQuery client library in Python to request the data.
	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 in a .csv file for frequency analysis	Same as planned
	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 with its frequency of occurrence in the corresponding time frame, sort the phrases by the frequency, and save the sorted list in a .csv file	Same as planned
	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 and save the sorted phrases in a .csv file	Same as planned. But only works for smaller datasets.
	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 allow users to enter a time frame and display the phrases with highest frequency. It should also allow the user to export the entire list as a file.	Same as planned
	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 showing the most popular phrases and their associated frequency.	Same as planned
	Connect front-end and back-end	Extension, not planned before	Be able to display the bar plot of most popular phrases and allow the user to download a .csv file containing all results after the user submitted the input.

Team reflection on progress

Provide some comments below regarding the completion of this milestone specifically around:
1. How is the project progressing?
2. Are there any differences between projected and actual outputs/outcomes?

Question 1

At first the teamwork was not very efficient. We gradually became familiar with platforms such as Discord and Google file sharing system for collaboration in document writing. We increased the frequency of team meetings for better communication. We also learnt about each other’s strengths and weaknesses to assign tasks more reasonably. These measures together improved the efficiency of teamwork.

Back-end development was the major part of milestone 1 and we progressed slowly in the beginning because of lack of experience in the programming tools. Due to the slow progress in the beginning, we made our milestone 1 plan relatively conservative. However, as we became more familiar with the tools, the project progressed faster than we thought and we ended up implementing extension features.

When we first started working on the front-end, we concentrated on the essential input areas of the webpage. However, we underestimated the difficulty in connecting front-end and back-end due to lack of experience in Python Flask, which made us to spend extra efforts connecting front-end and back-end in week 7.

Overall, we are a little ahead of our original plan. We will plan milestone 2 more ambitiously and implement extension features since we are more familiar with the tools now.

Question 2

- In our plan, the projected outcome for “Coding 1” was to download a .csv file from BigQuery and import it manually with Python for NLP analysis. Although we did achieve this, it will be tedious for the user to download and import the data manually. Therefore, we switched to an alternative method which is requesting data directly at the back-end using Google BigQuery client library in Python, which is more convenient for the user. Therefore, our planned output is no longer needed for this project.
- In our plan, the projected output for “Coding 4” was to perform NLP analysis on all posts from a year. However, the number of all posts within a year is massive and it will take too long to process due to limited computing power of regular laptops. Setting up access to high performance computing infrastructure would be time-consuming, so we limited the number of posts to a few thousands in our actual output.
- Connecting the front-end and back-end is an extension of milestone 1. We implemented this instead of other possible features because it is important to have a working prototype early to avoid risks of failing to connect front-end and back-end.

Team reflection on managing problems

Have you encountered any problems to date?
If so, how have you managed them?

Technical problems

- In the beginning of the project, Google BigQuery's authentication failed because we were unable to install the google.cloud package in Python due to conflicts with existing packages. At first, we got stuck in understanding the long error messages generated by package conflicts. But later we found it was much easier to reset the virtual environment which clears conflicts between packages.
- There were some strange words which seem like code fragments in our results. To manage the problem, we located the original post producing the strange word by printing out the post’s ID number. By looking at the original post, we realized that, due to lack of knowledge in HTML syntax, our method to process the raw data did not remove the code fragments completely, so we modified the method to eliminate code fragments in posts.
- The HTML file was not styled with the CSS file when we started running the user interface, because we were not aware that Python Flask looks for “template” and “static” folders for HTML and CSS files. To manage this problem, we re-configured the file structure according to the convention of Python Flask, and stated the path of the CSS file clearly in the code.

Project management problems

- At the early stage of the project, the team did not work efficiently because we had too much overlap between tasks of different team members. The way we managed it was to work as two pairs for front-end and back-end respectively.
- We did not get in-time update of other team members’ progress in coding due to low committing frequency of some team members. We are planning to manage this problem by setting up regular times when people have to commit their code.

Supervisor assessment	Please, rate your team (1) effort, (2) project progress and (3) their self-reflection for milestone 1 Rating scale 1-10 as per standard marking scheme, ie 5 is a Pass and 7 is a credit. Add some comments to explain your rating
Effort: Progress: Reflection:	