Trend Identification on Stack Overflow

Team 30B
Team members: Tianjiao Jiang, Po-Yi Lee
Client: Dr. Christoph Treude

1. Introduction

Our web-based application is developed for discovering trends on Stack Overflow (SO) and visualize how the trends change over time.
The software extracts certain bigrams from specific subsets of SO posts within user-specified time windows, using natural language processing (NLP) algorithm. It calculates the frequency of each bigram and display changes in the frequency of most popular bigrams over the selected time windows.
This software allows you to learn how the technology ecosystem is changing and where it might be going in the future.

2. Platforms and Tools

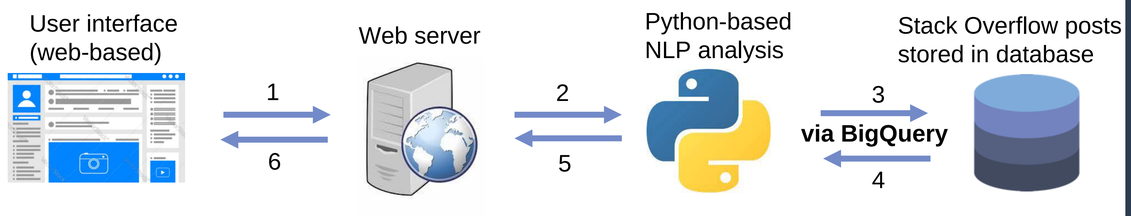
❖ **Platform**

- **Google BigQuery:** a cloud-based data warehouse that supports SQL queries. It was used to query the Stack Overflow dataset.

❖ **Tools**

- **spaCy**: a Python library for Natural Language Processing (NLP). SpaCy's rule-based matcher was used to recognize noun-noun and adjective-noun phrases from Stack Overflow posts.
- **Django:** a Python web application framework for back-end development.
- **HTML ● CSS ● JavaScript**: HTML was used to structure the content of our web pages; CSS was used for the design or styling of the website; JavaScript was used to make the website interactive.
- **Apache ECharts:** an interactive charting and visualization library for browser. It was used for trend visualization.
- **AJAX:** a technique for accessing web servers from a web page
- **jQuery:** a JavaScript library used to it easier to use JavaScript

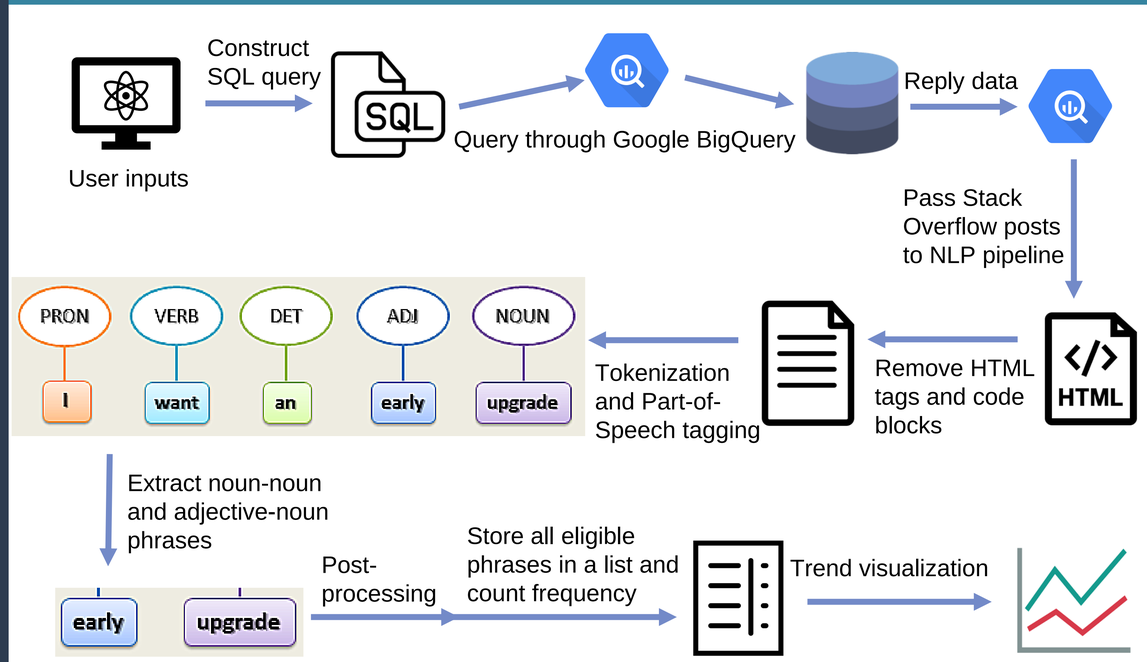
3. System Architecture



1) HTTP request
2) Pass on the user inputs (e.g., time windows, number of posts)
3) Request Stack Overflow posts from database based on user inputs, via BigQuery

4) Stack Overflow posts were sent to the NLP pipeline via BigQuery
5) Perform NLP analysis and generate data for trend visualization
6) HTTP response

4. Algorithms



5. Achievements

Planned

- ✓ The user interface accepts necessary inputs
- ✓ Inform the user about estimated processing time ~~and prevent the user from entering a number of posts that is too large~~
- ✓ Filter out incorrect results caused by defects of the NLP library
- ✓ Establish a method to select phrases to be displayed
- ✓ Enable user to search for trends related to a specific word

Extensions

- Search in titles and comments
- Make the trend visualization an interactive chart

6. Extensions

- The processing speed can be increased with access to more powerful CPU.
- The software can run on newly published browser versions.
- We could apply statistical testing to make the results more rigid.
- We could expand the range of target phrases to n-grams (n > 2).
- We could further explore the trends related to a particular programming language.

7. Conclusions

In conclusion, this software satisfied the client's basic requirement. This web-based application allows the users to search for trends within a particular time frame and will visualize changes in trends over the time frame. In the future, the software could be improved by increasing the processing speed and expanding the range of searching results.