In summer 2018, I interned at IBM US in the position of Data Scientist Internship. The goal of this position is to help the team of IBM DevOps Insights verify their hypothesis about software engineering. My job task was to evaluate the frequency and impact of constant string literals in the code commit history.

The main job tasks were 1) mining GitHub projects to filter out interesting data, including commits, committed files, issues, relations along with them; 2) extracting constant strings appeared in the files before and after commits; 3) measuring the correlation between constant string changes and issue resolution time; 4) classifying the changed strings to find the types we are interested in (e.g., password, URL, configuration, file path); 5) the computing huge amount of data using Spark.

By the end of the internship, I had successfully demonstrated the high frequency of changes of constant strings in the commits, and the proof of confidential information in the explicit text. My analysis results verified the importance of working on string literal analysis in security, bug resolution, and code refactoring.

Although the output of a research paper immediately is difficult because of IBM's commercial regulations, this internship experience was very helpful for my current research topic. By using industry-adopted techniques, I can retrieve and filter out the data I want more easily and efficiently. Furthermore, my experiment speed can be accelerated by utilizing the programming tips and tools I learned during the internship.

Firstly, data cleaning criteria I learned can be applied to my current regular expression study; secondly, the extraction of constant strings can be applied to extract the changes of regular expressions; then the knowledge of manipulating Git internals can save me the time of writing code for same or similar tasks; tools like Jupyter notebook can fast prototype, visualize, and debug Python code before turning them into modules; Spark can provide fast in-memory computing for mining thousands of GitHub projects.

The implicit impact of this internship is on my career and skills development. It puts me in a better situation for job hunting by demonstrating my industrial practice and skills. More importantly, it broadened my vision of industrial research. Industrial research is highly business motivated, and it is challenging to interpret research results and conclusions into business values.

Personally speaking, this internship also educated me in the long run on professional practices, such as skills of communication, time management, work collaboration, and fast learning. Professional practice is not only about not being late to work and replying to email promptly. Asking questions is a good way to learn fast about my work there. Talking to colleagues and peers when I encountered a technical problem rewarded me a prompt answer and correct method and thus saved me tens or even hundreds of hours. Talking to them about my ideas also helped me think critically and carefully.

The most satisfying aspect of my internship is that I learned how people work in the industry. It corrected my illusion that developer positions are boring and not research-related. Some of the developer positions can have a lot of research related. And the fast-change industry can keep challenging current solutions, leaving large enough room for developers to improve.

The least satisfying aspect is on team collaboration. Due to the job goals, I am not involved in the product development procedure. It is a pity that I have not had the chance to get my code reviewed by colleagues, and thus am not be able to know the gap between industry requirements and my capabilities.

Regarding the quality and quantity of work, I find that there are no clear and stable criteria to evaluate them. My response to these two boxes is decided by comparing it to my previous internships.