Offline Recommendation System for Python Programmers

Ashutosh Chaturvedi
North Carolina State University
Raleigh, North Carolina, USA
achatur@ncsu.edu

Ayush Gupta
North Carolina State University
Raleigh, North Carolina, USA
agutpa25@ncsu.edu

Harshdeep Kaur
North Carolina State University
Raleigh, North Carolina, USA
hkaur4@ncsu.edu

Ankit Kumar
North Carolina State University
Raleigh, North Carolina, USA
akumar18@ncsu.edu

ABSTRACT
The rule of meaningful communication is using proper language. Language which is understandable to communicator as well as the receiver. Programming works on a similar rule, i.e. for writing any application we need to follow the rules as put down by the underlying programming language. This paper presents the analysis of a major problem programmers face while working with any language. The problem is analyzed using related literature and user surveys. A project idea is formulated in order to solve the identified problem.

Keywords
Recommendation System; Stackoverflow.com; Python; Programmers; Syntax; IDE; Sublime Text;

1. INTRODUCTION
A common problem that programmers have been facing since the advent of programming languages and face even today, is finding a solution to the syntactical errors while programming in different languages. Users who are new to a programming language, or migrating to different programming languages regularly face problem with correct syntax to be used while coding any feature. While many languages like Java, C++, C#, etc., have advanced IDEs like eclipse and Visual Studio which provide a solution to this problem to some extent, many other like python do not have such advanced IDEs and the programmers who program in these languages have to constantly refer to resources like tutorial documents or internet for solutions. Also, because python has an interpreter instead of a compiler, the task of finding errors at compile time becomes even more difficult and leaves no option to the user but to refer the stated resources. Documentation and Tutorial provide with information about the features of the language and how to use different features but they are incompetent when it comes to providing solution to general issues faced by the users. Generally, the sought solution is based on other user's experience with the similar issue and same is posted on different forums across internet. Most of the time people get solution for their problems by referring to these forums. Searching for the optimal solution across these platforms is pretty cumbersome and time consuming at times as the user might have to go through a number of solutions for the same problem. Also, in the absence of internet, there is pretty much no option left because of so much dependency on the internet these days. There is no dedicated offline tool to help in this cause. When a programmer moves to a new language other than his primary programming language, or a new user is learning a new language, most of the time the issue faced by them is how to express their login in proper syntax as per the rules of the programming language at hand. As we talked to the people and collected data through surveys, we found out that most of the programmers face this issue in programming and that syntactical problems are more common and cumbersome to solve than semantics. As discussed later in this report, majority of the people want an offline solution to this problem which can extract results from sites like stackoverflow.com and generate optimized results for their problems. This is what we propose to solve!

2. RELATED WORK
There are various applications being build that use the Stackoverflow.com data for different purposes [7]. In specific there is an application present to provide an Offline version of Stackoverflow.com called StackDump. This application requires the user to download the huge dump of Database from StackExchange.com and then use the application for offline search. But this application does not provide an efficient alternative for browsing through the various articles it provides in the search, i.e. it provides only basic search results as Stackoverflow.com provides. Also it asks the user that they have a lot of memory free for the data dump but it would include a lot of overhead for programmers who just need syntactical help and hence articles related to that. Our aim is to provide solutions to these unanswered problems.

3. PROBLEM ANALYSIS
Data was collected from different sources such as user observation, user survey and literature review. This data was categorized, analyzed and visualized to better understand the problem statement and provide an appropriate solution. The following sections describe the various user base observed, the data collection process, visualization process and the validation of the data.

3.1 User Base
We targeted two types of users - **Naive** programmers, who are new to python or are learning python as secondary language and **experienced** programmers, who use Python as their primary programming language (including programmers migrating to

Python 3 from Python 2 [3]). Sixty-three users participated in our online survey form while six programmers were informally interviewed and discussed with.

3.2 Data Collection

The participants were involved in informal interviews, collective discussions, self-analysis, results of online activities. Data was primarily gathered by means of report taking and collection of data generated through the online survey. Informal data was collected by observing and interviewing our users.

3.2.1 Observations and Informal Interviews

The new programmers were working on introduction module of Python Google Tutorial [1] under one of the course's prerequisite [2]. Programmers who use Python on regular basis were observed during their programming activities and were also informally interviewed. Our main focus was to analyze the activities of the users while they code and their different approaches for solving the encountered programming errors. We wanted to observe the difference in the approach of naive and experienced towards the issues and solution methods for these issues. We tried focus more on new programmers as the scope of variety in data is more since they tend to make more mistakes and refer to resources for finding the solution. We have documented the common errors that user make and procedures that users adopt to solve the problems. Comparisons were made between different resources and tools used as issue resolution by the users.

3.2.2 Online Survey

Online survey was used to collect participant observation at larger scale. Total 63 users participated in our survey. Data was taken only from users who have used both python and StackOverflow.com on regular basis and understand the context of the survey correctly. The users were asked to rate and provide their opinion regarding various parameters related to the subject matter. Few of the main points asked in the questionnaire include:

Q: Which programming problem do you face the most?

This was the base question to find out the type of problem that programmers face more frequently. The response of this question helps us in determining the scope of the application to be implemented.

Q: Which resource do you use most for the above problem's resolution? If using Internet, do you use StackOverflow.com for solutions?

The idea behind asking this question was to list all the resources commonly referred by the programmers for seeking solutions, as internet is the main source, we also wanted to record the user base of programmers using StackOverflow.com for answers.

Q: Would it be useful if there is an offline StackOverflow.com resource that can help you with your queries?

A common issue faced by all of us is paucity of offline resources while dealing with errors. Though books and tutorials help us to guide about the basics of a language, they are not much of a help when dealing with variety of errors a user face during developing an application. The response to the above asked question was inline with this thought process.

Q: Would you like to have more optimized results or general results from StackOverflow.com?

Everyone likes to find the best solution with shortest number of searches. Referring too many resources, opening number of tabs after google search just to find an optimal solution takes a lot of time.

3.3 Data Visualization

The data collected by observing our users and through informal interviews with the experienced programmers gave us a thorough view of the problems faced by them in coding and their approaches for finding solutions to them. Observing them showed us that as a programmer there is so much to know and memorize. But in reality programmers only memorize a little and rest is assisted by the IDE, tutorials, and internet. Stack Overflow [5] (A privately held website, the flagship site of the Stack Exchange Network which features questions and answers on a wide range of topics in computer programming) has been one of the most popular resources for programmers trying to solve programming problems. Majority of the programmers we were dealing with also tend to use and rely on Stackoverflow.com for their solutions. We noticed that in 80% of the cases programmers were preferring only Stackoverflow.com solutions over others. We even tried to get an idea of their reason for using Stackoverflow.com.

We have collected the online survey result data [3] and have represented the same in graphical format. The graphical representation shows us the differences in type of errors encountered by Naive and Experienced users. It also shows the similarity in internet usage as primary resolution resource among programmers.

From the graph, it can be easily inferred that StackOverflow.com is the primary approach for resolutions.

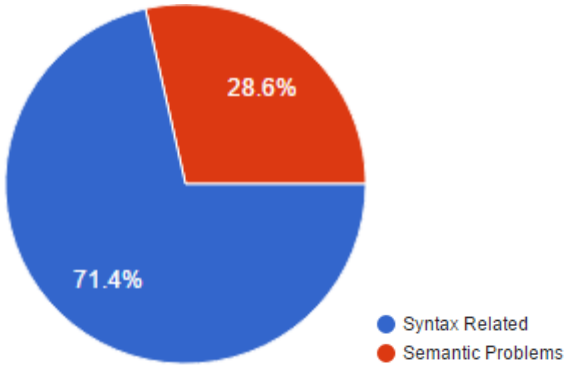

Figure 1: The percentage of the two types of problems programmers face.

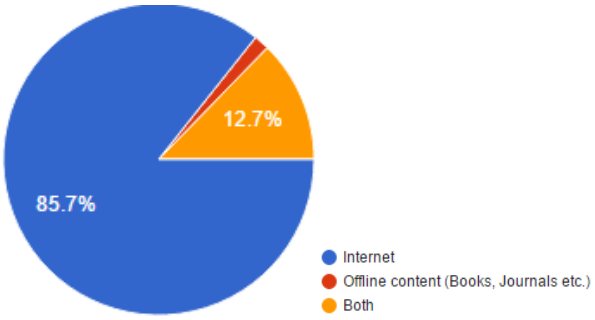

Figure 2: - Percentage of programmers using Internet, Offline Content and Both for finding solutions to their problems

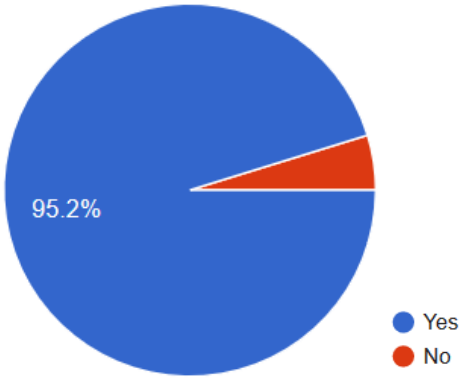

Figure 3: - Percentage of users using StackOverflow.com to solve their problems

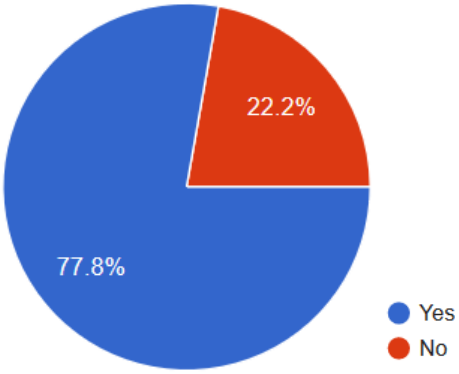

Figure 4: - Percentage of programmers in favor of offline StackOverflow.com

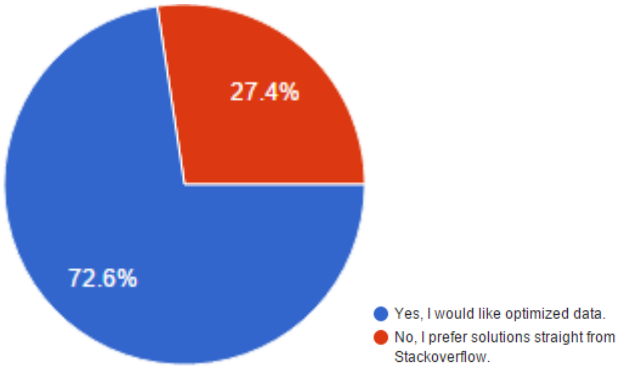

Figure 5: - Percentage of programmers depicting whether they would like optimized results over simple search results from Stackoverflow.com

3.4 Data Analysis Conclusion

The analysis of the data - problems faced by the programmers, approaches for solving the problems, dependency on the internet etc., inspired us to build a solution focused on the issues we found. Since dependency on the internet was high and users preferred having offline tool, we came to an idea of providing some offline recommendation of an idea to provide offline Stackoverflow.com recommendation system. As articles on Stack Overflow were often among the first search results in any programming related Google search among our user base. And since significantly large number of python programmers are facing problems because of lack of well-defined IDE, we decided to focus our recommendation system for python programmers. Stack overflow has currently 526,656 articles related to python. Therefore, the data scope is also big for our recommendation system.

3.5 Literature Review

We also came across one research paper [6] which involves a study of what kind of programming errors occur and how much time programmers spend on debugging them. The paper shows that nearly 50% of the time of programmers is spent on debugging. This means that programmers spend 50% of their time trying to find out errors, solution to those errors and then resolving those errors. We aim at reducing time by improving upon the component of finding solution to errors efficiently and providing an offline system so that the programmer need not delay his work in case a stable internet connection is unavailable. Further the paper shows that the result of their experimentation was that about 30% of the errors and the cause of about 22% of the breakdowns was due to language constructs i.e. syntaxes and it came in 2nd just after errors due to algorithms. This in coherence with our survey shows that a majority of errors programmers face today is syntax related and hence this inspired us to focus on providing solutions to the syntax related errors the programmers face.

4. PROJECT IDEA

Upon analyzing the major issues programmers face while learning a new language, or during day to day programming tasks, we realized that the majority of programmers faced a heavy dependence on internet for finding solutions to a lot of syntactical errors they face. Upon looking at the resources programmers use on the internet, we found that the majority of the traffic was on a single website called StackOverflow.com. This website provides a space where people can host their questions and in turn others can answer them. Most of the programmers benefit from not asking the questions themselves but by looking at someone's question which maybe the same as theirs. As the programming community right now is quite a big one, most of the question a programmer has, specially related to syntaxes, has already been asked and answered before and hence programmers prefer to just browse through Stackoverflow.com to find similar questions which have bee answered. But a big hurdle in finding such solutions is the inefficiency of Stackoverflow.com to provide the users with better filtered results, due to which users need to browse through a huge collection of articles with simple keyword search or basic filtering to find what they need. Further our survey indicated that a majority of the programmers we surveyed would prefer a tool which could provide them with better results to reduce the time spent on their end for searching for the correct article. This motivated us to analyze the tactics humans use for finding the correct post or article on Stackoverflow.com matching their needs, and then implement these tactics in the form of a recommendation system to help out the programmers efficiently search for the desired question.

Further as our survey also indicated that a lot of programmers wanted an offline system that would offer them the capability of working even at places where a good internet connection is not available, we concluded that pulling the database of Stackoverflow.com offline through the use of existing APIs and then using this database to offer solution to the users would be a

great way of meeting their need for an offline Stackoverflow.com. Hence we concluded that we need to develop an offline recommendation system for Stackoverflow.com for programmers facing problems with syntaxes. As the whole database comprising of every language would be really big and it would be infeasible to pull the whole database offline using the existing APIs and in the time frame we have, we are focusing on one language: Python. Why Python? This is due to the following:

- Python is an interpreter and not a compiler based language
- There is no good IDE that python programmers use and instead most of them use simple text editors to code
- Python has a large community

Hence our final project idea has converged to building an "Offline Recommendation System for Python Programmers".

This project has various different components in it.

- There is a recommendation system that would recommend articles from Stack Overflow to the programmers.
- The recommendation system would only need internet to set it up and then would work completely offline
- The recommendation system aims at proving better results than simple search to the user.

Moving forward, we aim to build this project as a plugin for Sublime Text. Why Sublime Text? Because of the following:

- Sublime Text is a widely used text editor by programmers.
- Sublime Text supports plugins in python hence making it possible for us to design the recommendation system using NLTK: a great natural language processing tool kit for python.
- It is available free of cost

The functionality we aim to provide would involve the users highlighting a few keywords and then searching with them using the plugin. They would also get an option to type in additional keywords that they think are relevant, and then the recommendations system would return in an optimal manner a list of top 5 articles that it deems most fit for the user's query. Our aim is to design a system that would almost always return the desired article in these top 5 results, however due to external noise such as irrelevant keywords added in by the user or other factors sometimes this system might not return exactly what the user seeks and hence there would also be an option to fetch further more results and look at all the possible articles that might be relevant to the user's query.

Thus, through this project we aim at reducing the dependency of the users on a good stable internet to be able to resolve their programming errors and also to provide them with a more efficient tool that would help them save a lot of time of browsing through pages to solve their issues. As this issue is quite a big issue we aim at creating a "High Impact" application for the programming community!

5. REFERENCES

[1] https://developers.google.com/edu/python/

[2] http://engineeringonline.ncsu.edu/onlinecourses/coursemarketing/SPR-2016/CSC591-791.html

[3] http://python-notes.curiousefficiency.org/en/latest/python3/questions_and_answers.html

[4] https://docs.google.com/forms/d/1vjQVA1AWui84H0gsym3gjAuuf_A-nfSct-HFPHY0f0E/closedform

[5] http://stackoverflow.com/questions/tagged/python

[6] http://repository.cmu.edu/cgi/viewcontent.cgi?article=1183&context=hcii

[7] http://stackapps.com/