DATA MINING

Project Proposal-Data Mining on Stack Overflow Data

Team members -Tarun Sunkaraneni and Garima Chhabra

Introducing the idea

Google is a behemoth of analytics and data storage. From offering unlimited photo storage for recognition purposes, to using search data to show retail options, it is becoming less and less of a joke when people say “Google knows me better than I do.”

Along the same lines, Stack Overflow can be thought of as the google for all kinds of software engineers, being the number one resource to get help diagnosing problems in configuration, logic, and functionality for people with all degrees of exposure.

As Software engineering is booming rapidly in this age which will most likely be known as the “technological revolution,” and analysis of Q/A data is arguably the most viable way to get solid statistics on the general population ranging from discipline distributions, skills that carry across disciplines, and information on the structures of ideal questions and answers. That is why we want to work with Stack Overflow analytics data to understand Software engineering branches that exist by tendencies of questions being asked and answered.

About the data

10% of questions and answers from the Stack Overflow programming Q&A website on Kaggle.

This is organized as three tables:

- The Questions table contains the title, body, creation date, closed date, score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.
- The Answers table contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
- The Tags table maps every question and answer to a category which they pertain to. (e.g. Java or Angular)

What we will be doing?

- We will try to extract features of developers for example by performing singular value decomposition on the data. This is supposed to yield out developer niches frequently spoken of such as Front-end development, Back-end development, Database Engineering etc, by analyzing strong links and tendencies in questions asked and answered by users.

- Understanding languages which are often complemented with one another, and what we can be deduced about the roles each of these languages serve.

Optional Questions:

- Understanding What is the general structure of a good answer, and what features make an answer good using rudimentary NLP.

Using this information, it is possible to do various kinds of analytics: ranging from the learning patterns of developers through the progression of time, understanding both popular and hidden links between languages to understand the niche these languages serve to different kinds of software engineers and environments.

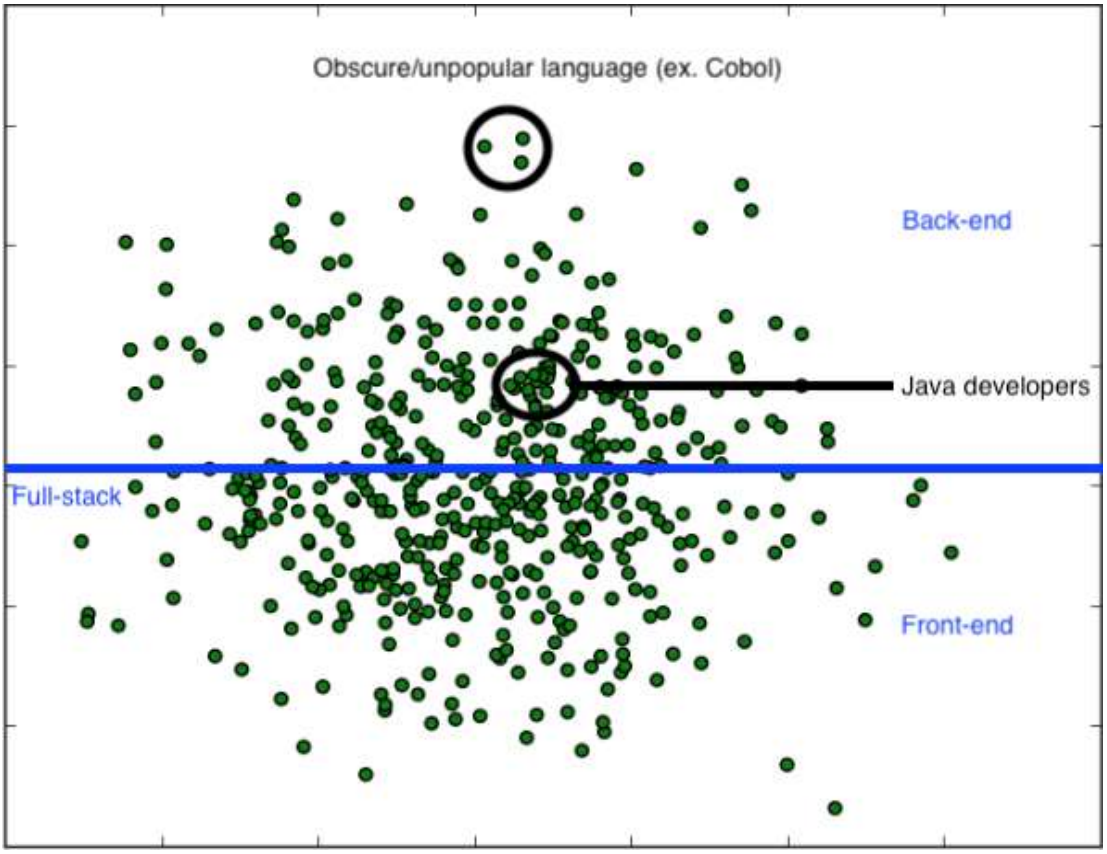
Optionally, we can come up with the ways that can show the properties of ideal format of answers, and understand the general public's proficiency in languages based on keywords that appear frequently in questions for topics.

The biggest design question in this project will be determining the best way to convert question and answer data of users into high-dimensional subspace coordinates. This may require many factors such as score, reputation and other prioritization, but consistency will be required across all categories to plot this data accurately. The coordinates will consist of sparse data, since it is unlikely that someone has either asked a question or answered one in all 2000+ categories.

End Structure:

The end product addressing the central question (what categories exist) should be scatter plots representing users on a 2-dimensional axis (projected down from a couple thousand dimensions (tags)). Assuming this is done correctly, in a hypothetical outcome we would be able to observe that all front-end developers were distributed above the axis while back-end developers would be distributed below the axis. People placed in the middle would be considered full stack, and we would see that most abundant occurrence of points would be the Java developers portion of the chart with which we can derive the borders of the specific discipline (ex. C# near the top right corner of the scatter plot distribution)

Example of "web-development" distribution (no offense back-end Cobol developers)



More than Just Homework?

In this project we are diving deeper into the science of analytics, unraveling the chosen dataset to analyze what embedded information can be recovered. We are trying to use concepts such as PCA, clustering and NLP to examine this data. As outputs of each segments rely on one another, it will require careful and compatible decisions to be made when using SVD, dimensionality reduction and clustering so that these techniques will build on each other and produce correct and intuitive results.

When this project is finished, the end results should be able to show how many users are in a certain discipline of Computer Science, which is more significant than knowing what tag (language) someone is using. Simply put, there's more substance knowing that person X uses Java for image-processing while person Y uses it for swing applications than only being told that both person X and person Y use Java. With a large timeframe of such data, it will be possible to deduce the *roles* of programming languages and how their uses are changing constantly.