Stack Overflow Network Analysis

A Network Tour of Data Science

Project Proposal

Claas Bruess, Max Ruenz and Simon Romanski

Abstract— **This project is aiming for an intuitive understanding of collaborations and community structures in StackOverflow. Therefore, a network based on the StackOverflow dataset is created. A comprehensive analysis will be carried out focusing on the communication between users in selected communities. The structure and evolution of these communities are evaluated and compared based on the means that were taught in the lectures and exercises. Finally, the obtained structures are visualized to gain the desired intuition.**

I. INTRODUCTION

Stackoverflow is one of the most used forums for software development. It contains questions and answers for a variety of programming languages, libraries, operating systems and domains.

We want to examine whether these communities are similar in their structure and whether they are connected to other communities.

Therefore we will create a network on the StackOverflow Dataset with a strong focus on communication between users.

II. DATASET

The dataset that we are going to analyze is provided by StackOverflow itself. As it has been published in September 2017 it is reasonable to assume that the data represent the current structure of the communities. The Dataset can be found in [1]. As the entire dataset is too big and unhandy, we decided to focus on selected communities only. In order to obtain these communities we are extracting the data that is centered around certain programming languages and libraries, for instance, Pandas or D3.js.

III. ANALYSIS

We intend to classify the users in these communities into different types of user categories:

- One Time Contributer
- Super User
- Enquirer

The classification will be based on their role and behavior in the community network. We will perform the classification by considering metrics that are platform inherent:

- Reputation
- Number of Answers
- Number of Questions
- Votes

as well as the network properties:

- Giant Component Analysis
- Incoming/Outgoing Degree
- Number of Questions
- Hubs

Please note that these lists are not comprehensive at this stage of the project. They are likely to change after the proposal. In addition to the tasks mentioned above, we would like to investigate how these community networks evolved over time and whether they can be represented by the models that were presented in class.

IV. COMPARISON

Finally, we would like to compare the examined communities and investigate in their differences and similarities. We want to obtain if certain communities contain stronger hubs than others if or how many disconnected components there are in the community network and how easy is it to separate the network into clusters. This comprehensive evaluation will hopefully yield insights to get an intuition of how these communities work online.

V. VISUALIZATION

To get a good understanding of the network and to verify our results later on it makes sense to plot the networks. That way we can easily see and verify hubs, giant components and clusters. Furthermore, visualizing the networks of helps to identify differences and similiarites between communities from the plots.

Note that Claas Bruess and Max Ruenz are working on the same dataset for the data visualization course COM-480, but do not visualize the network.

REFERENCES

[1] https://archive.org/details/stackexchange