36-651 Project proposal

Ruiyang Gan

January 2019

Project overview

For this semester's statistics computing project, I will conduct a network analysis of GitHub users' collaboration and repositories networks. GitHub is currently the biggest web-based hosting service for version control using the version control system git. Currently, there are around 20 million users registered at GitHub. Around 60 million repositories are hosted remotely on GitHub.

To process graph formed from data set of such large scale, specific computing tools will be needed. In this project, I will use Apache Spark and its graph processing package GraphX to conduct collaboration network analysis. The package comes with an implementation of Graph using the fundamental structure RDD in Spark. GraphX also provides several graph computing algorithm, such as triangle counting and pageRank. A Python interface (binding) is provided for Python Programmers.

Project goals

1. Develop an understanding about Apache Spark and its graph processing package GraphX. Learning about the pipeline of distributed computing and statistical analysis on large data set.

2. Constructing GitHub's (public) collaboration network in Spark using Spark and GitHub's REST API

3. Conduct exploratory data analysis on GitHub's collaboration network. EDA would include degree distribution, calculation of centrality measures in the collaboration network, detection of subgraphs with specific structure (such as triangles and stars).

4. Implement community detection algorithm in Spark and applies the algorithm to the GitHub collaboration network.

5. (If longitudinal network data available) Analyze the dynamics of GitHub collaboration network.

Expected Timeline

Week 1: Project Proposal
Week 2: Learning about fundamental data structure in Spark and GitHub's REST API for publicly available user information
Week 3: Construct the GitHub's collaborator network in Spark using PySpark and GitHub's REST API
Week 4: EDA of collaboration network
Week 5: Implementation and application of Community Detection algorithm
Week 6: Analysis of Dynamics in GitHub's collaboration Network
Week 7 & 8: Wrap-up; Writing of project presentation and tutorials

Previous Experience with software tools

In 36-650, I used Python as my main programming language. I have no prior experience with Spark and Scala. I have some experience with Java.

Data Set Description

The data set is composed of public user information on GitHub (The interface is provided by GitHub). We can either regard users or repositories vertices. The connection between users will be formed by GitHub's features: Star, Watch, and Follow. The connection between repositories will also depend on the aforementioned features. In addition, the dependencies features in a GitHub repositories (if a set up file is found in the repository) will be used to determine the connection between repositories. The formulation of edges will depend on whether we define users or repositories as vertices. If we were to consider users as vertices, then the edges will determined by whether the users have worked on the same project (And if so, how many project have they worked on together, thus leading to a formulation of weighted graph). If we were to use repos as the vertices, then we can use the dependency feature to connect the repos together.

Reference

Programming

Documentation of github's REST API: `https://developer.github.com/v3`
Python Bindings for Spark GraphX: `https://issues.apache.org/jira/browse/SPARK-3789`
GraphX programming Guide: `https://spark.apache.org/docs/latest/graphx-programming-guide.html`
Tutorial to Scrapy (Web Scraping package in Python): `https://doc.scrapy.org/en/latest/intro/tutorial.html`

Statistics

Cosma's course on Statistical Network Analysis: https://www.stat.cmu.edu/~cshalizi/networks/16-1/, https://www.stat.cmu.edu/~cshalizi/networks/16-2/
Mark Newman's book on Networks: Newman, M.E.J, *Network: An Introduction*, Oxford University Press, New York, 2013