More Pipelined Data Processing using Spark

Chan-Ching Hsu

Purpose

Write programs with pipeline jobs to analyze the GitHub and graph data using Apache Spark.

Task 1

The data is “github.csv”. For each language, find out how many repositories using it, one repository that has the highest stars number.

The generated list is the following with the format:

<language> <num_of_repo> <name_of_repo_highest_star> <stars>

num_of_repo: total number of projects on GitHub using a specific language
name_of_repo_highest_star: name of a repository that has highest stars number

This list is sorted by the num_of_repo in descending order.

```
Java 462182 elasticsearch/elasticsearch 15698
Ruby 363801 rails/rails 29825
Python 331883 jakubroztocil/httpie 21283
PHP 273999 zurb/foundation 22636
C++ 159831 rogerwang/node-webkit 27350
C 145354 neovim/neovim 17395
C# 116155 dotnet/corefx 9176
```

Task 2

A directed graph G=(V,E) consists of a set of vertices V, and a set of edges E such that each element e in E is an ordered pair (u, v), denoting an edge directed from u to v. In a directed graph, a directed cycle of length three is a triple of vertices (x, y, x) such that each of (x, y), (y, z), and (z, x) is an edge in E.

The data is “patents.txt”. The graph is in the form of an edge list. Every line of the file has information about a single edge. A line contains information in the format <from vertex> <to vertex>, which means the patent <from vertex> has a citation to patent < to vertex>.

The program calculates the number of all directed cycles of length three in a graph, but for this particular dataset, the number is 0 since it is a graph of citations, meaning the edges have chronological meaning. More precisely, a directed cycle of length three exists when there is a loop among three nodes, and the loop goes from the first node to the second and to the third and finally back to the first node. Such kind of loop, however, is not a nature thing in a citation network since a patent A filed in an early time has no way to cite another patent B that is filed in a later time and also have a citation to patent A. For a directed cycle of length three to exist, there must be a loop among three nodes in existence. Due to the nature of the graph formed by the dataset, there is no such loop.