Hadoop-Rs - Programming Assignment

Goals

There are five tasks in all. I have modified the original tasks from the programming assignment to fit my datasets.

Task 1

Study centrality metrics: degree centrality, PageRank centrality, k-core centrality.

Task 2

Compute the most highly cited *paper* using degree centrality with focus on in-degrees, i.e. most influential paper.

Task 3

Compute the most influential nodes using the PageRank algorithm.

Task 4

Compute the most influential nodes using the k-core centrality metric.

Task 5

Repeat tasks 2-4 with Apache Giraph.

Overview

The project uses the Rust programming language to create mappers and reducers for Hadoop. The Streaming tool provided by the Hadoop project is used to call the mappers and reducers (see more in `tools/hadoop.sh`). This project uses a small dataset (5.6MB of edges) and a standalone installation of Hadoop. Two datasets are used, the larger one from http://snap.stanford.edu/data/cit-Patents.html and the smaller one from http://snap.stanford.edu/data/cit-HepPh.html. Both are citation networks.

Accomplishments

I completed the first three tasks, i.e. studied centrality, and calculated InDegrees and PageRank for the datasets.

Interesting Aspects

- ** Uses the streaming Jar instead.
    - ** Simply a wrapper allowing Hadoop to call executables directly.
    - ** Meaning we can use any language we want to write our mappers and reducers.

- **Hadoop** Installing Hadoop is very different depending on the PC. I tried installing it a number of times (back in october, but that install died), and again in january, where I did it with group mates. My installation is not running just yet.
- **Debugging** - Debugging distributed applications is difficult. Finding out how the MapReduce framework works like Unix pipes helped isolate where things were going wrong.
- **Data Formatting** - In order to calculate the PageRank for the dataset, I needed to change its format. I was able to do this using a simple IdentityMapper + AppendReducer to quickly reformat the dataset into an adjacency matrix from an adjacency list.
- **Sorting and Shuffling** - The PageRank algorithm requires repetition until convergence, i.e. that certain steps be repeated until the output is stable. As far as I'm aware, it only does this for transfers from Mappers to Reducers, which means I can't just feed the output of one reducer into another.
- **Small data-set** - means that hadoop would only start a single node anyways.
- ** The data we used is instead just a list of nodes, with a list of nodes that are related to that key-node so: 1 2 3 4 2 1 2 Is

Antallet af delta skulle gerne være lig med det totale antal nodes.

Key Items

The relevant files are listed in `src/mappers` and `src/reducers`.

Alt i bin

Filer i bin er filer der er brugt til at interface med Hadoop

Results

Use the provided tools and view the results in the output directory.

Makefile:

Simply describes how the project should be run should run. Is called by doing: "make " The makefile designates which mapper and which reducer to use.

Pagerank

Preprocessing is done by combining keys of the same value, and simply adding the related values to that "master" key Should be run a number of times. To run pagerank

1. Ensure file is in data folder and that file is updated in makefile
2. Open terminal in project
3. run "make pagerank-preprocess", wait for it to finish
4. run "make pagerank-first"
5. run "make pagerank-intermediate", can be run a number of times, to close approximate the exact pagerank
6. run "make pagerank-sum", to sum up the deltas.

Preprocessing is only be used to page-rank

Is only used to join keys of the same value, so that key

Combiner

Simply combines all the deltas, reducing the complete

Degree centrality:

Simple counts the number of nodes that point to a specific node. Is actually only wordcount, there is nothing to find the max 😦

Hadoop.sh

Used to check for all the relevant requirements, hvis hadoop er installeret til usr/local behøves path ikke blive sat (følg digital ocean guiden.) https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-in-stand-alone-mode-on-ubuntu-16-04

Reducer modtager en liste af key, med en liste af values til hver key.