Big Data Technology

CSCI 493.76: Project 2 – HetioNet

Spring 2024

Grade = minimum(total score, 100).

Coding Section (70 points)

HetioNet is a hetnet with multiple node and edge (relationship) types, which encodes biology. The hetnet was designed for Project Rephetio, which aims to systematically identify why drugs work and predict new therapies for drugs.

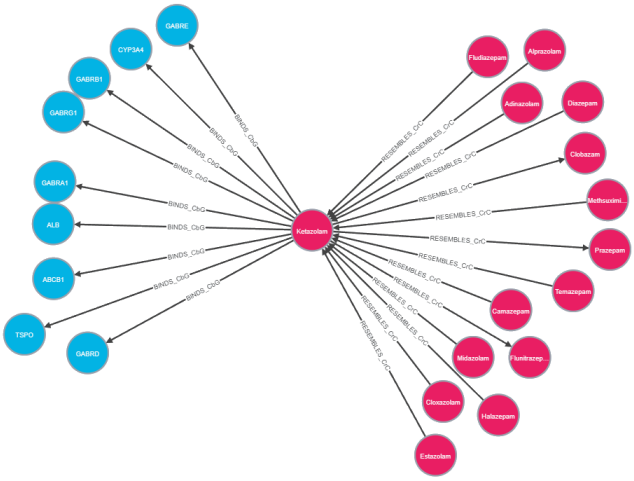

Figure 1: A compound (RED NODE) that are BINDS (CbT) to several genes (BLUE node) and is RESEMBLES to other compounds.

1. Readin nodes.tsv and edges.tsv. Find "tsv" is short for Tab Separated Values, where each values (text or numerical values) are separated by a TAB. You can choose the data structure type to store the values (pyspark or pandas).

2. For each compound, compute the number of genes that are BIND (CbG) to it using **MapReduce method**. Output results with top 5 number of genes in a descending order. See example for Figure 1.

3. For each DISEASE, compute the number of GENE(s) that are UPREGULATES (DuG) using **MapReduce method**. Output results with the top 5 number of GENE(s) in a descending order. See example for Figure 2.

   Downregulation is the process by which a cell decreases the production and quantities of its cellular components, such as RNA and proteins, in response to an external stimulus. The complementary process that involves increase in quantities of cellular components is called upregulation.

4. (Teams with 3 people) For each compound, compute the number of other compounds that are DOWNREGULATES (CdG) it using **MapReduce method**. Output results with the top 5 number of compounds in a descending order.

5. For item number 2 to 4 above, compute the hash tables using mid-square method (with $r = 3, 4$) **OR** Folding Method (digit-size = 2 and 3). Experiment with 10 hash tables for the selected method. Which method (which digit-size?, $r = 3$ or $r = 4$) require least number of storage? Use sys.getsizeof(.) to find the size of the hash tables (10) with its associated link list(s).

Each Team will prepare a report. The report should include:

- Each question in "Coding" section and the **output** to the question.
- A pseudo code for each question above.
- References. A list of all references used. Example: links to websites where you found the codes. If you read some articles, please reference them. If you used Chatgpt, write down each question you asked.
- Files of your Source Code.

Each Team will submit the report and all codes related to the project.

Presentation (30 Points)

Each team is required give a presentation on their Project. During the presentation, each team will answer all the questions in the Coding section. Also helps to include your pseudo code. **Please be prepare to answer questions about your code and how it works.**

Code Organization (7 points)

Please Organize your code. Your code should be inside Class(es) or function(s). Ex. If you reuse the same code twice, it would be useful to put those lines into a function. Your code should also include some comments, but it should not be excessive.

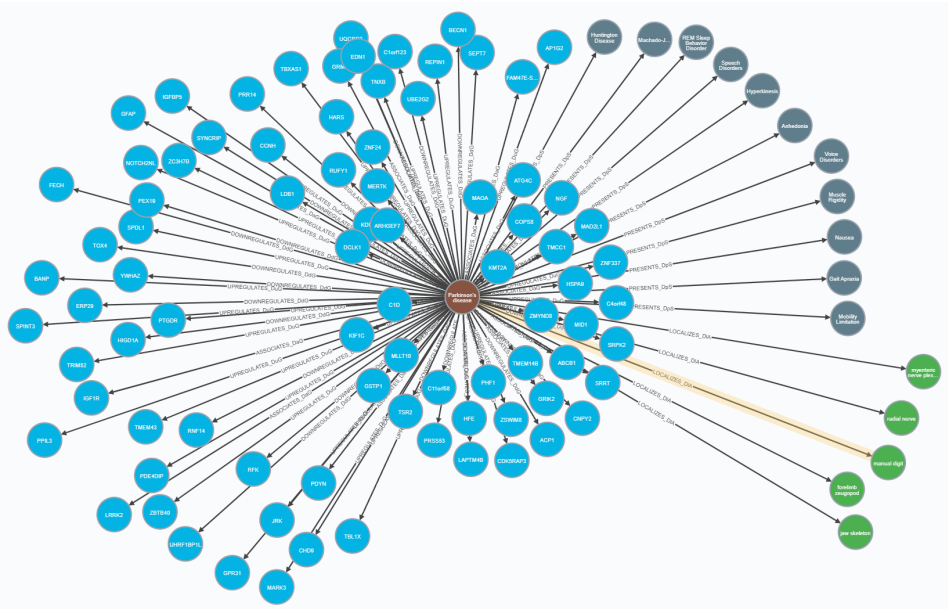

Figure 2: A disease (BROWN NODE) that are UP-REGULATES (DuG) / DOWN-REGULATES (DdG) / ASSOCIATES (DaG) to several genes (BLUE node). A disease (BROWN NODE) PRESENTS a symptom (GRAY NODE). A disease (BROWN NODE) LOCALIZE in an anatomy (GREEN NODE).

References

- https://neo4j.het.io/browser/
- Files: nodes_test.tsv, edges_test.tsv
- https://github.com/hetio/hetionet/tree/main/hetnet/neo4j
- https://github.com/cielavenir/procon/blob/master/hackerrank/map-reduce-advanced-count-number-of-friends.py