Project 2: HetIONet

Course: Big Data Technology

Student: Fabiola Li Wu

Instructor: Arezoo Bybordi

Due Date: Monday, April 28th, 2025

Data: HetIONet (nodes.tsv and edges.tsv)

– Q1: For each drug, compute the number of genes and the number of diseases associated with the drug. Output results with the top 5 number of genes in descending order.

1. Filter all the edges with source = 'Compund'
2. Filter all the genes and diseases from step 1.
3. Map every compound <compoundID, 1>
4. Group the maps by the compoundID e.g. <compundID, genesAmount> and <compundID, diseaseAmount>
5. Sort the map by its values.
6. Prints out the first 5 compoundID, amount of genes, and amount of diseases associated with the drug(compound)

```
Query 1: Top 5 compounds with most genes associated, and their associated diseases
Query 2: Query 2: Top 5 diseases with drugs associated
Query 3: Names of top 5 compounds with most genes associate.
Press 0: EXIT
Enter the query number: 1




Query 1: Top 5 compounds with most genes associated, and their associated diseases
Compound ID         Genes     Diseases
----------------------------------------
Compound::DB08865   585       1
Compound::DB01254   564       1
Compound::DB00997   532       17
Compound::DB00570   523       7
Compound::DB00390   522       2
```

– Q2: Compute the number of diseases associated with 1, 2, 3,..., n drugs. Output results with the top 5 number of diseases in descending order.

1. Filter all the compounds that have disease in their target.
2. Map <disease, drugsAmount> e.g: disease1 has 4 drugs related to it -> <disease1, 4>
3. Filter out the diseases with amount greater than n.
4. Map <drugAmount, drugAmountAmount>. For example: <disease1, 4>, <disease4, 4>, <disease10, 4> -> <4, 3>
5. Print out.

```
Query 1: Top 5 compounds with most genes associated, and their associated diseases
Query 2: Query 2: Top 5 diseases with drugs associated
Query 3: Names of top 5 compounds with most genes associate.
Press 0: EXIT
Enter the query number: 2
Enter the number of drugs: 100




Query 2: Top 5 diseases with less than 100 drugs associated
------------------------------------------------------------
1 drugs -> 10 diseases
2 drugs -> 7 diseases
11 drugs -> 6 diseases
9 drugs -> 6 diseases
3 drugs -> 6 diseases
```

– Q3: Get the name of drugs that have the top 5 number of genes. Output the results.

1. Filter all the edges with source = ‘Compund’
2. Filter all the genes and diseases from step 1.
3. Map every compound <compoundID, 1>
4. Group the maps by the compoundID e.g. <compundID, genesAmount>
5. Sort the map by its values.
6. Print out the first 5 compound name by getting it from the nodes.tsv file, and the amount of genes.

```
Query 1: Top 5 compounds with most genes associated, and their associated diseases
Query 2: Query 2: Top 5 diseases with drugs associated
Query 3: Names of top 5 compounds with most genes associate.
Press 0: EXIT
Enter the query number: 3




Query 3: Names of top 5 compounds with most genes associated
------------------------------------------------------------
Crizotinib -> 585
Dasatinib -> 564
Doxorubicin -> 532
Vinblastine -> 523
Digoxin -> 522
```