CONSTRUCTING AN EVOLUTIONARY TREE
BY ENHANCING A GENERAL TREE ADT

Tree of Life

Asst. Prof. Dr. Feriştah DALKILIÇ
Res. Asst. Orkun ÇINAR
Res. Asst. Muharremcan GÜLYE

2020510110 Shakhobiddin Urinov

1. Project Overview

This report explains my efforts to build a Java-based implementation of the Tree of Life (ToL) dataset. The project models evolutionary relationships between species, emphasizing clarity, scalability, and user-friendliness. It enables functionalities such as dataset loading, tree traversal, and property analysis, making it a practical educational tool.

2. Key Components

2.1 TreeNode Class

This class is fundamental to representing each species or taxonomic group in the tree. Each node encapsulates critical attributes and methods for managing relationships.

Attributes:

- **nodeId**: A unique identifier for a node, essential for referencing nodes within the dataset.
- **nodeName**: The name of the species or group, allowing users to understand the node's significance.
- **children**: A list storing child nodes, representing the node’s direct descendants.
- **parent**: A reference to the parent node, enabling navigation upward in the tree hierarchy.
- **isLeaf**: A boolean indicating whether the node is a leaf (has no children).
- **isExtinct**: A boolean showing whether the species is extinct.
- **confidence**: Represents how confident the placement of the species is within the tree (e.g., "confident", "problematic").
- **phylesis**: Describes the phylogenetic status of the node, such as "monophyletic".

Methods:

- addChild(TreeNode<T> child): Ensures the parent-child relationship is accurately established. It updates the child's parent reference and avoids adding duplicate children.
- getExtinctDescription(): Simplifies understanding by returning either "extinct" or "living" based on the node's extinction status.
- getConfidenceDescription(): Provides a user-friendly explanation of the confidence level (e.g., "confident position").
- getPhylesisDescription(): Describes the phylogenetic group, such as "monophyletic" or "uncertain monophyly".

2.2 TreeInterface

This interface defines the standard operations for the tree data structure. It ensures consistency in how the tree behaves across various methods.

Key Methods:

- getRootData(): Retrieves data from the root node, offering a starting point for tree analysis.
- getHeight(): Calculates the height of the tree, which is the longest path from the root to any leaf node.
- getNumberOfNodes(): Counts the total number of nodes in the tree, helping to assess its size.
- isEmpty(): Checks whether the tree contains any nodes, ensuring validity before operations.

2.3 TreeIteratorInterface

This interface focuses on traversal functionality, allowing the user to efficiently navigate the tree.

Key Method:

- getPreorderIterator(): Enables pre-order traversal, where the root is visited first, followed by its children recursively from left to right.

2.4 EvolutionaryTree Class

This class forms the core logic for the evolutionary tree. By implementing both TreeInterface and TreeIteratorInterface, it combines data handling with traversal mechanisms.

Features Explained:

1. **Dataset Loading:**
   - loadFromCSV(String nodesFile, String linksFile): This method reads node and link data from CSV files to construct the tree. Malformed data is logged for troubleshooting.
   - **Helper Methods:**
     - processNodeLine(String line): Parses a line from the nodes file, creating and storing a TreeNode object.
     - processLinkLine(String line): Establishes parent-child relationships based on the links file.
2. **Tree Traversal:**
   - getPreorderIterator(): Implements a stack-based traversal. By ensuring nodes are visited in the correct order, it facilitates accurate analysis of hierarchical data.
3. **Tree Properties:**
   - getHeight(): Recursively calculates the tree's height by exploring the longest path from the root to a leaf node.
   - getBreadth(): Counts all leaf nodes by employing breadth-first traversal with a queue.
   - getDegree(): Determines the maximum number of children for any node, providing insights into the tree's complexity.
4. **Data Analysis and Visualization:**
   - printSubtree(int subtreeId): Displays all nodes in a specified subtree, formatted hierarchically for easy understanding.
   - printAncestorPath(int nodeId): Shows the lineage from the root to a given node, highlighting evolutionary connections.

- findMostRecentCommonAncestor(int nodeId1, int nodeId2): Identifies the nearest common ancestor of two nodes.

5. **Error Handling:**
   - Errors during dataset processing are logged to an error_log.txt file, ensuring traceability for debugging.

2.5 Main Class

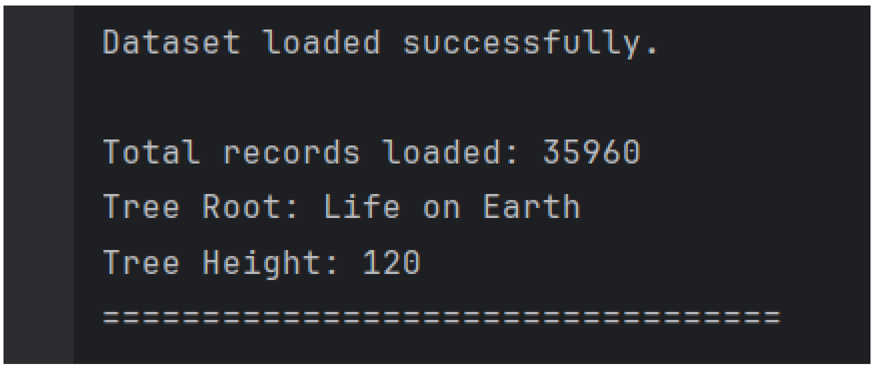
This class serves as the interface between the user and the evolutionary tree. It uses a menu-driven approach to simplify interactions.

Menu Options:

1. Load dataset: Reads and builds the tree from CSV files.
2. Search for a species: Finds a species by its ID and displays detailed information.
3. Pre-order traversal: Saves the traversal result to a file for further analysis.
4. Print subtree: Visualizes a subtree from a specified root node.
5. Display ancestor path: Shows the lineage of a node from the root.
6. Find most recent common ancestor: Determines the shared ancestor of two nodes.
7. Calculate tree properties: Displays metrics such as height, degree, and breadth.
8. Longest evolutionary path: Outputs the longest path from the root to a leaf.
9. Exit: Ends the program.

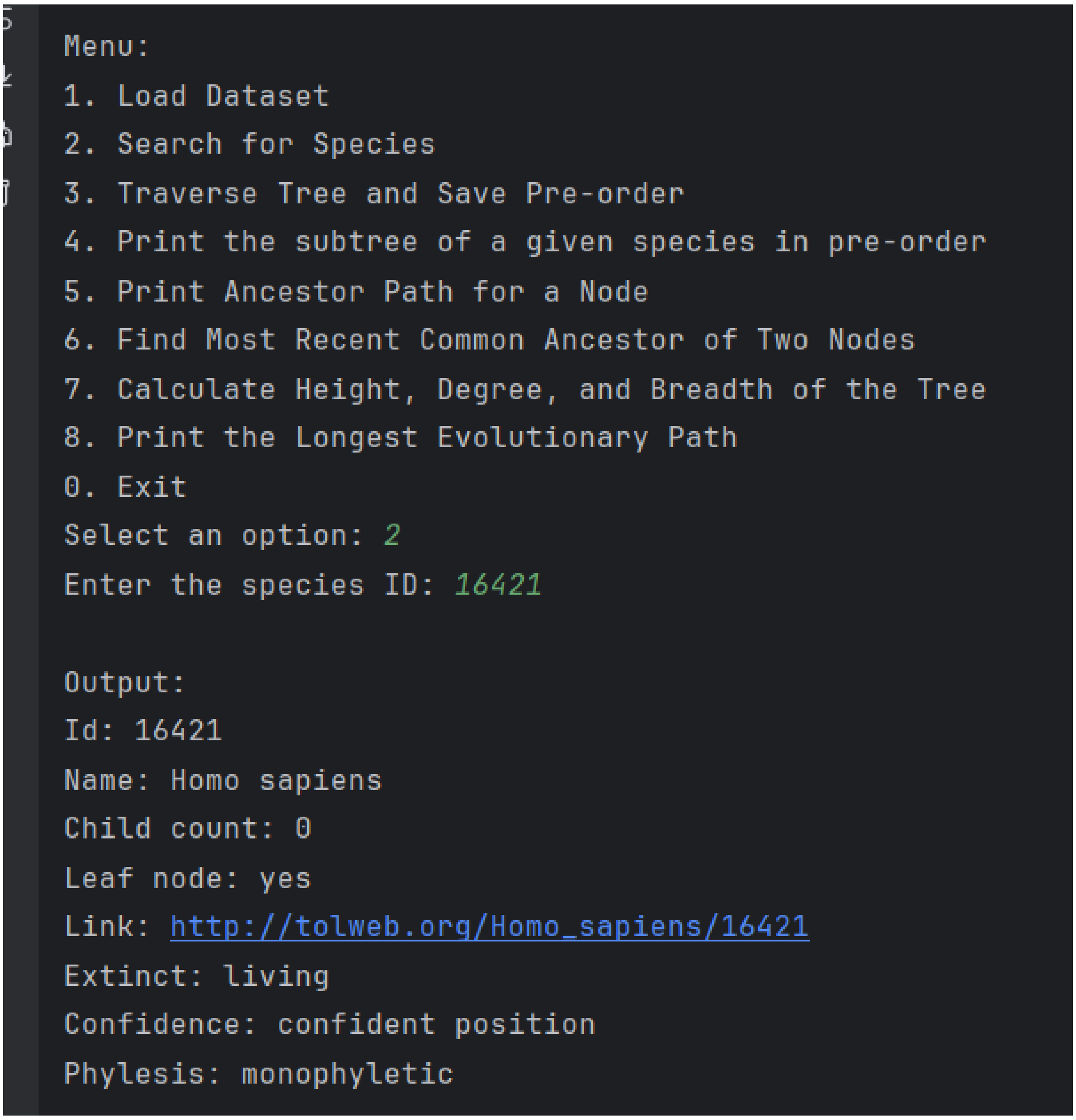
3. Example Outputs

3.1 Dataset Loading



```
Dataset loaded successfully.

Total records loaded: 35960
Tree Root: Life on Earth
Tree Height: 120
==================================
```

3.2 Search for species.

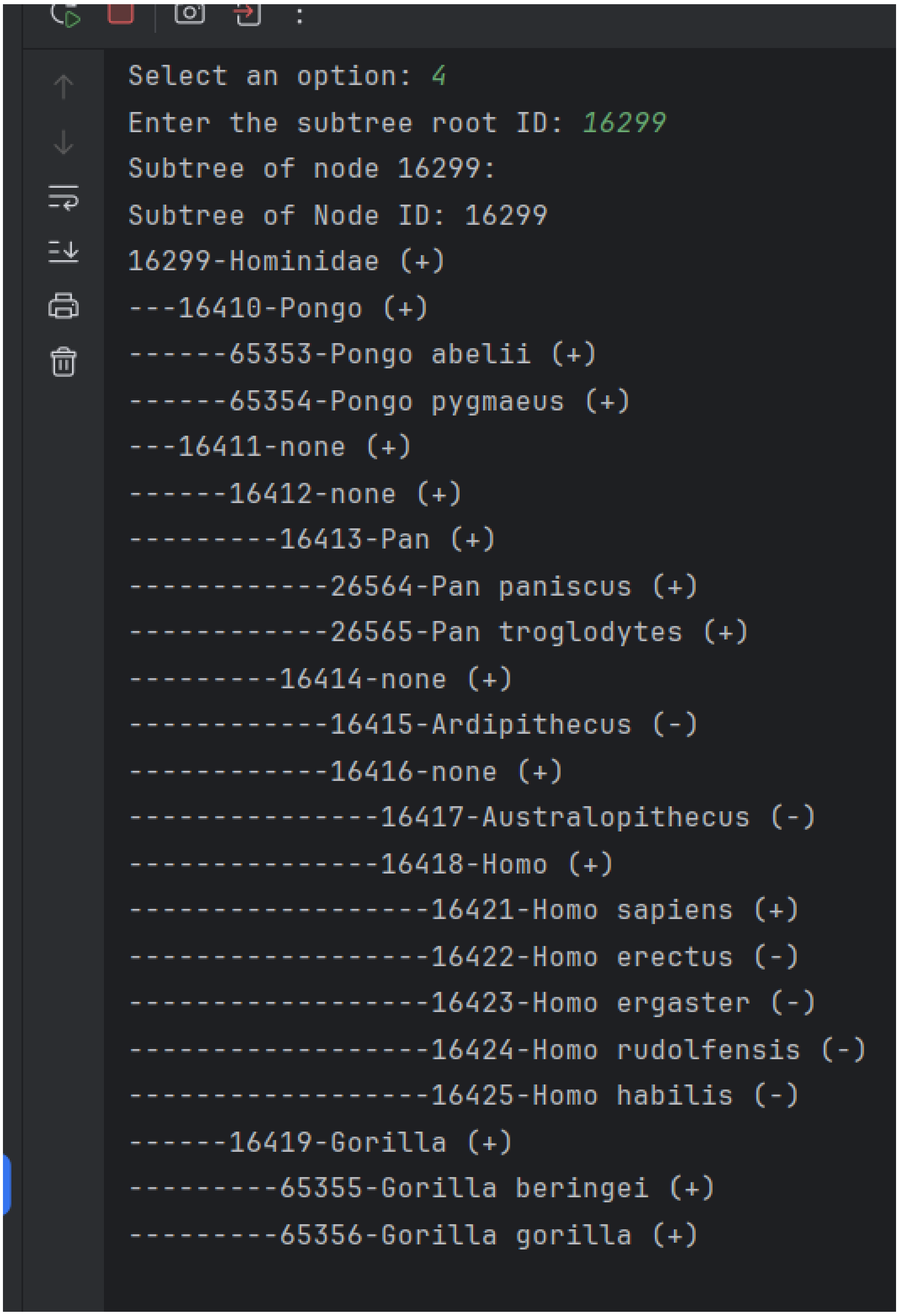

```
Menu:
1. Load Dataset
2. Search for Species
3. Traverse Tree and Save Pre-order
4. Print the subtree of a given species in pre-order
5. Print Ancestor Path for a Node
6. Find Most Recent Common Ancestor of Two Nodes
7. Calculate Height, Degree, and Breadth of the Tree
8. Print the Longest Evolutionary Path
0. Exit
Select an option: 2
Enter the species ID: 16421

Output:
Id: 16421
Name: Homo sapiens
Child count: 0
Leaf node: yes
Link: http://tolweb.org/Homo_sapiens/16421
Extinct: living
Confidence: confident position
Phylesis: monophyletic
```

3.3 Pre-order Traversal

Node ID: 1, Data: Life on Earth
Node ID: 2, Data: Eubacteria
Node ID: 2285, Data: Aquificae
Node ID: 59615, Data: Aquifex
Node ID: 59616, Data: Calderobacterium
Node ID: 59617, Data: Hydrogenobacter
Node ID: 59618, Data: Thermocrinis
Node ID: 59619, Data: Hydrogenothermus

```
Node ID: 59620, Data: Persephonella
Node ID: 59621, Data: Sulfurihydrogenibium
Node ID: 59622, Data: Balnearium
Node ID: 59623, Data: Desulfurobacterium
Node ID: 59624, Data: Thermovibrio
Node ID: 2286, Data: Thermotogae
Node ID: 59625, Data: Thermotoga
Node ID: 59626, Data: Fervidobacterium
Node ID: 59627, Data: Geotoga
Node ID: 59628, Data: Marinitoga
Node ID: 59629, Data: Petrotoga
Node ID: 59630, Data: Thermosipho
Node ID: 2287, Data: Thermodesulfobacteria
Node ID: 59633, Data: Thermodesulfobacterium

  .

.

.

.

.

Node ID: 57983, Data: Tobamovirus
Node ID: 57984, Data: Tobravirus
Node ID: 57985, Data: Hordeivirus
Node ID: 57986, Data: Furovirus
Node ID: 57987, Data: Pomovirus
Node ID: 57988, Data: Pecluvirus
Node ID: 57989, Data: Benyvirus
Node ID: 57990, Data: Bromoviridae
Node ID: 57991, Data: Ourmiavirus
Node ID: 57992, Data: Idaeovirus
Node ID: 57993, Data: Tymoviridae
Node ID: 57994, Data: Closteroviridae
Node ID: 57995, Data: Flexiviridae
Node ID: 57996, Data: Barnaviridae


Pre-order traversal saved to preorder_output.txt.
```

3.4 Print the subtree of a given species in pre-order.

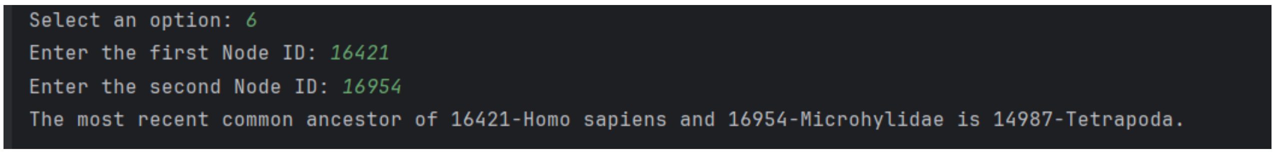


```
Select an option: 4
Enter the subtree root ID: 16299
Subtree of node 16299:
Subtree of Node ID: 16299
16299-Hominidae (+)
---16410-Pongo (+)
------65353-Pongo abelii (+)
------65354-Pongo pygmaeus (+)
---16411-none (+)
------16412-none (+)
---------16413-Pan (+)
------------26564-Pan paniscus (+)
------------26565-Pan troglodytes (+)
---------16414-none (+)
------------16415-Ardipithecus (-)
------------16416-none (+)
---------------16417-Australopithecus (-)
---------------16418-Homo (+)
------------------16421-Homo sapiens (+)
------------------16422-Homo erectus (-)
------------------16423-Homo ergaster (-)
------------------16424-Homo rudolfensis (-)
------------------16425-Homo habilis (-)
------16419-Gorilla (+)
---------65355-Gorilla beringei (+)
---------65356-Gorilla gorilla (+)
```

3.5 Print Ancestor Path

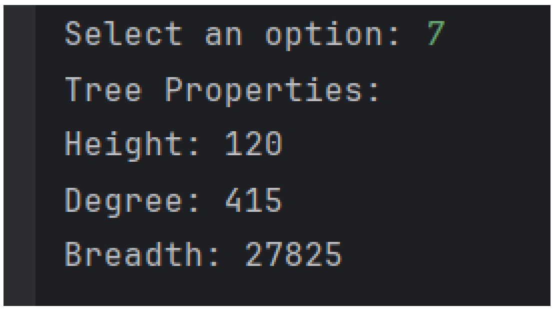
```
Ancestor Path for Node ID 16421:

1-Life on Earth (+)

--3-Eukaryotes (+)

----2372-Opisthokonts (+)

------2373-none (+)

--------2374-Animals (+)

----------2458-none (+)

------------2459-Bilateria (+)

--------------2466-Deuterostomia (+)

----------------2499-Chordata (+)

------------------14820-none (+)

--------------------14822-none (+)

----------------------14826-Craniata (+)

,

,

,


-----------------------------15963-Primates (+)

-------------------------------16290-none (+)

---------------------------------16291-none (+)

-----------------------------------16293-Catarrhini (+)

-------------------------------------16298-none (+)

---------------------------------------16299-Hominidae (+)

-----------------------------------------16411-none (+)

-------------------------------------------16412-none (+)

---------------------------------------------16414-none (+)

----------------------------------------------16416-none (+)

-------------------------------------------------16418-Homo (+)

---------------------------------------------------16421-Homo sapiens (+)

            .

                .

                 .
```

3.6. Find the most recent common ancestor of the given two species.



```
Select an option: 6
Enter the first Node ID: 16421
Enter the second Node ID: 16954
The most recent common ancestor of 16421-Homo sapiens and 16954-Microhylidae is 14987-Tetrapoda.
```

3.7. Calculate the height, degree and breadth of the tree.



```
Select an option: 7
Tree Properties:
Height: 120
Degree: 415
Breadth: 27825
```

3.8 Longest Evolutionary Path

Select an option: 8

Longest Evolutionary Path:

1-Life on Earth

--3-Eukaryotes

----2372-Opisthokonts

------2373-none

--------2374-Animals

----------2458-none

------------2459-Bilateria

--------------2466-Deuterostomia

----------------2499-Chordata

------------------14820-none

--------------------14822-none

----------------------14826-Craniata

------------------------14829-Vertebrata

--------------------------14833-Node 1

----------------------------14836-none

------------------------------14838-Node 2

--------------------------------14840-Node 3

----------------------------------14843-Gnathostomata

-------------------------------------14919-Node 1

---------------------------------------14920-Teleostomi

---14921-Osteichthyes

---14922-Sarcopterygii

---14944-none

---14948-none

---14950-none

---14952-Terrestrial Vertebrates

---14975-none

---14976-none

---14977-none

---14978-none

---14979-none

---14980-none

---14981-none

---14982-none

---14983-none

---14984-none

---14985-none

---14986-none

---14987-Tetrapoda

---14988-Reptiliomorpha

---14989-none

---14990-Amniota

---14846-Reptilia

---14862-Romeriida

---14864-none

---14866-Diapsida

---14903-none

---14905-none

---14907-none

---14913-Sauria

---14916-Archosauromorpha

---14886-none

---14888-none

---14890-none

```
------------------------------------------------------------------------------------------------------------14892-none
--------------------------------------------------------------------------------------------------------------14894-none
----------------------------------------------------------------------------------------------------------------14896-none
------------------------------------------------------------------------------------------------------------------14898-none
--------------------------------------------------------------------------------------------------------------------14900-Archosauria
----------------------------------------------------------------------------------------------------------------------14869-none
------------------------------------------------------------------------------------------------------------------------14871-none
--------------------------------------------------------------------------------------------------------------------------14873-none
----------------------------------------------------------------------------------------------------------------------------14875-none
------------------------------------------------------------------------------------------------------------------------------14877-none
--------------------------------------------------------------------------------------------------------------------------------14879-none
----------------------------------------------------------------------------------------------------------------------------------14881-none
.
.
.

-----------15829-Euornithes (true birds)
-------------15834-Neornithes
---------------26291-Neognathae
-----------------26305-Neoaves
-------------------26410-none
---------------------15868-Passeriformes
------------------------67943-none
-------------------------29222-Oscines
---------------------------67944-none
-----------------------------67945-none
-------------------------------67946-none
---------------------------------67947-none
-----------------------------------67948-none
-------------------------------------67949-none
---------------------------------------29223-Passerida
-----------------------------------------67950-none
-------------------------------------------67277-none
---------------------------------------------67278-Passeroidea
-----------------------------------------------67283-none
-------------------------------------------------67284-none
```

```
----------------------------------------------------67285-none
------------------------------------------------------67286-none
--------------------------------------------------------67287-Nine-primaried oscines
----------------------------------------------------------67288-none
------------------------------------------------------------67289-none
--------------------------------------------------------------67290-none
----------------------------------------------------------------67292-Icteridae
------------------------------------------------------------------67380-Icterus
--------------------------------------------------------------------67499-none
----------------------------------------------------------------------67444-none
------------------------------------------------------------------------67445-none
--------------------------------------------------------------------------67446-none
----------------------------------------------------------------------------67447-none
------------------------------------------------------------------------------67500-none
--------------------------------------------------------------------------------67449-none
----------------------------------------------------------------------------------67501-none
------------------------------------------------------------------------------------67450-none
--------------------------------------------------------------------------------------67451-none
----------------------------------------------------------------------------------------67452-Icterus cayanensis cayanensis
```

4. Conclusion

As a student, this project provided me with valuable experience in implementing tree data structures and solving real-world problems using Java. By modeling the Tree of Life dataset, I enhanced my understanding of hierarchical relationships and traversal algorithms. The program’s modularity and user-friendly design make it a versatile tool for learning and exploration.