**Assignment-1**

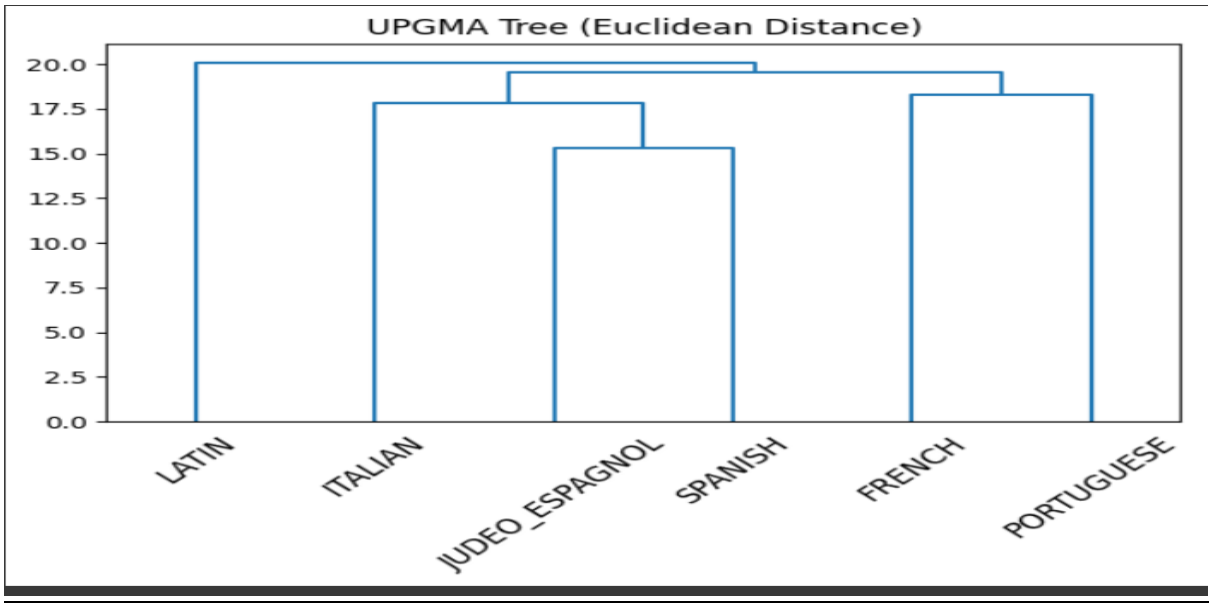
**Natural Language Processing**

**Part-1**

To provide a detailed analysis report

**Euclidean Distance Analysis:** This method, which measures linguistic disparities based on a variety of linguistic characteristics, determines the direct distance between points in a multidimensional space. The resulting dendrogram provides an overview of linguistic relationships by grouping languages according to their overall similarity, as seen through the prism of feature disparity.

This is the plot of the Euclidean distance:

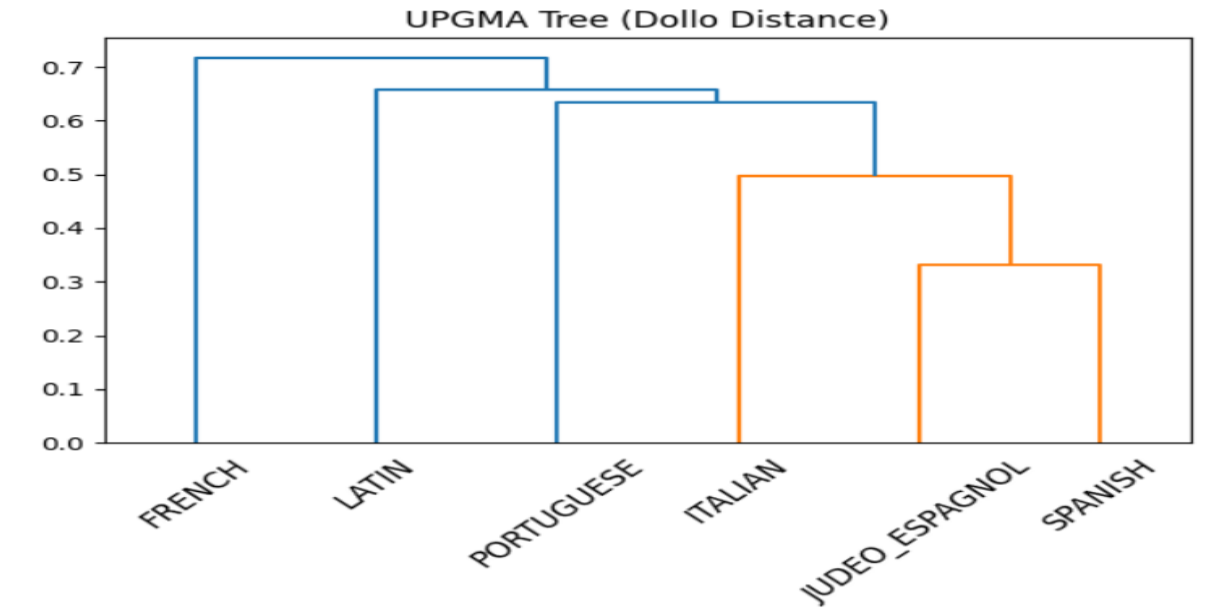



Here is the detailed report of the report through analysis:

| Language | FRENCH | ITALIAN | JUDEO_ESPAGNOL | LATIN | PORTUGUESE | SPANISH |
|---|---|---|---|---|---|---|
| FRENCH | 0 | 12.4498996 | 14.93318452 | 13.89244399 | 12.92284798 | 13.49073756 |
| ITALIAN | 12.4498996 | 0 | 13.19090596 | 12.72792206 | 12.88409873 | 11.53256259 |
| JUDEO_ESPAGNOL | 14.93318452 | 13.19090596 | 0 | 15.09966887 | 14.49137675 | 10.63014581 |
| LATIN | 13.89244399 | 12.72792206 | 15.09966887 | 0 | 14.35270009 | 14.03566885 |
| PORTUGUESE | 12.92284798 | 12.88409873 | 14.49137675 | 14.35270009 | 0 | 13.15294644 |
| SPANISH | 13.49073756 | 11.53256259 | 10.63014581 | 14.03566885 | 13.15294644 | 0 |

**Dice Distance with Dollo Principle:** Specifically designed for binary data, this metric assesses the degree of similarity between languages by concentrating on common traits, considering both the presence and absence of particular attributes. This analysis is based on the Dollo model, which suggests that linguistic traits can exist once but can also vanish more than once. This analysis's dendrogram groups languages according to common characteristics, highlighting distinctive characteristics in particular.

This is the plot of the Dolllo(Dice)




Here is the detailed report of the report through analysis:

| Language | FRENCH | ITALIAN | JUDEO_ESPAGNOL | LATIN | PORTUGUESE | SPANISH |
|---|---|---|---|---|---|---|
| FRENCH | 0 | 0.424657534 | 0.51981352 | 0.509234828 | 0.473087819 | 0.473958333 |
| ITALIAN | 0.424657534 | 0 | 0.3625 | 0.376744186 | 0.410891089 | 0.305747126 |
| JUDEO_ESPAGNOL | 0.51981352 | 0.3625 | 0 | 0.461538462 | 0.448717949 | 0.226452906 |
| LATIN | 0.509234828 | 0.376744186 | 0.461538462 | 0 | 0.492822967 | 0.438752784 |
| PORTUGUESE | 0.473087819 | 0.410891089 | 0.448717949 | 0.492822967 | 0 | 0.408983452 |
| SPANISH | 0.473958333 | 0.305747126 | 0.226452906 | 0.438752784 | 0.408983452 | 0 |

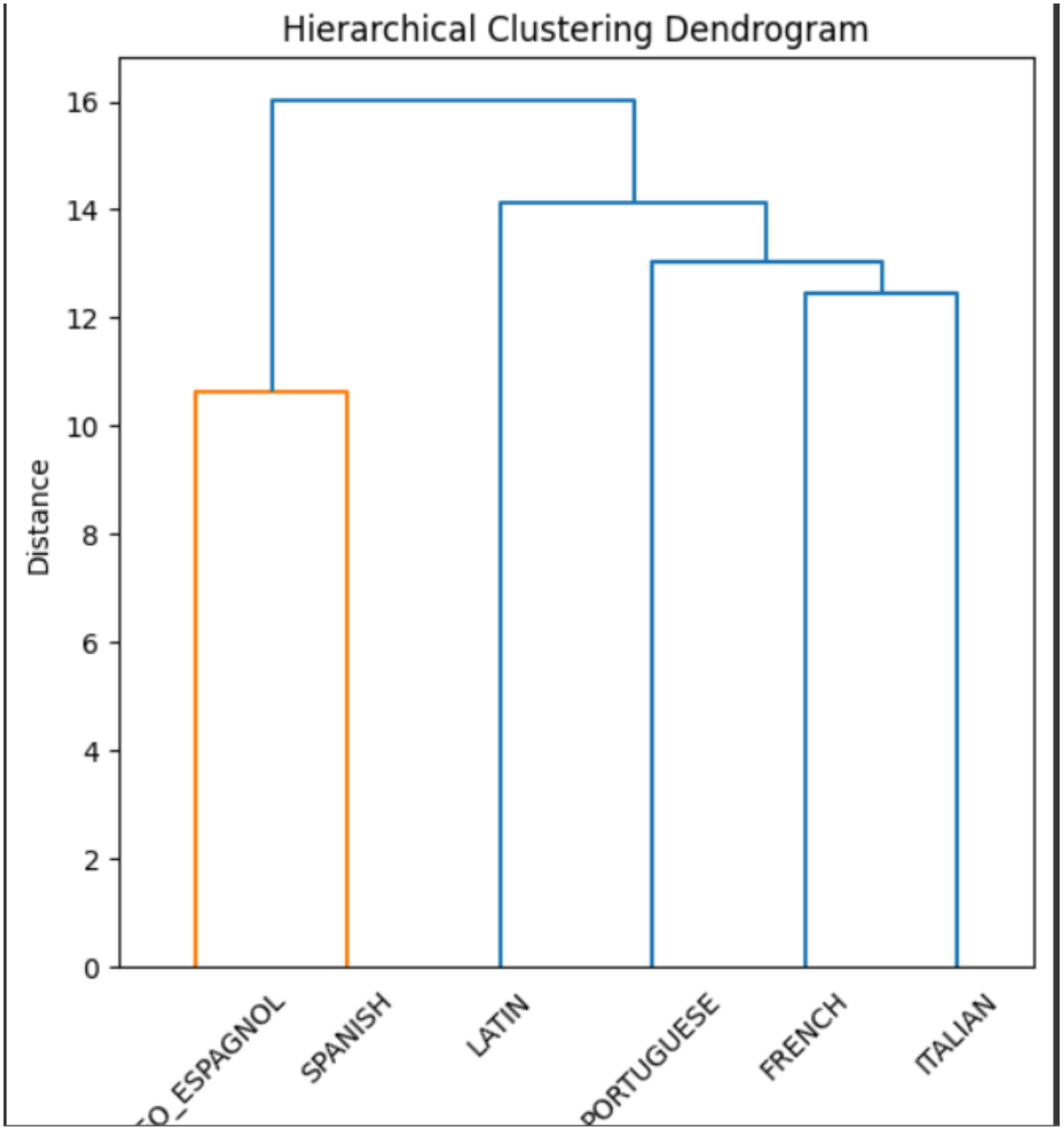
**Insights of the analysis:**

**Cluster Dynamics**: The dendrograms provide a clear picture of how languages are grouped together, highlighting the subtle parallels and discrepancies between them. More similarity between languages is shown by their closer grouping, which may indicate underlying linguistic or historical relationships. The dendrograms provide a clear picture of how languages are grouped together, highlighting the subtle parallels and discrepancies between them. More similarity between languages is shown by their closer grouping, which may indicate underlying linguistic or historical relationships.

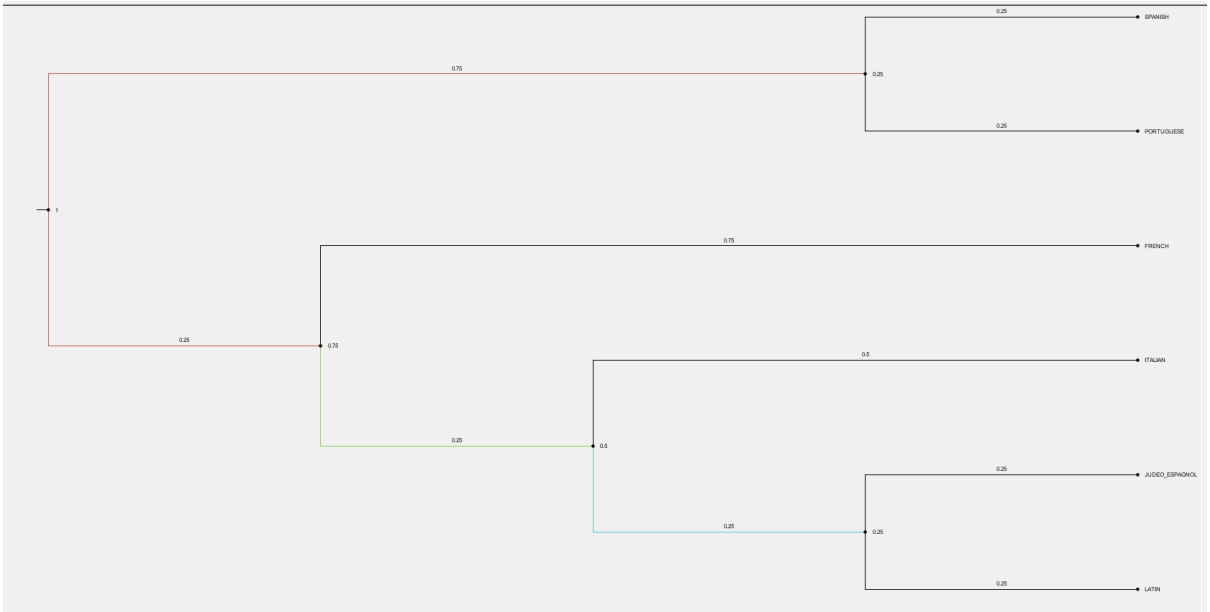
**Metric-Driven Differences:** Different clustering results are obtained depending on the metric selected (Euclidean vs. Dice). While the Dice metric, particularly when combined with the Dollo assumption, focuses on common linguistic characteristics, it can lead to a variety of interpretations regarding language relationships. In contrast, the Euclidean metric provides a broad review of feature differences.

**Linguistic Relationships:** These dendrograms' grouping patterns may reveal more intricate genetic or historical connections between languages. Tightly connected communities could point to a common ancestor or past exchanges, which would be interesting to explore further in terms of language study.

Additionally, this Hierarchical dendrogram:



**Tree of Neighbouring tree or network from the dice distance matrix and plot the results:**




This dendrogram depicts a hierarchical clustering of languages or dialects, showcasing how Spanish and Portuguese form a closely related cluster distinct from French and Italian, which are also closely linked to each other. Another grouping includes Judeo-Espagnol and Latin, with Judeo-Espagnol being somewhat more distant from Latin than the others in their clusters. The structure and branch lengths indicate varying degrees of similarity and dissimilarity among the languages, suggesting that French and Italian share a close relationship due to their Latin roots, similar to the close connection between Spanish and Portuguese within the Ibero-Romance category. Latin's foundational position highlights its historical impact on Romance languages, while Judeo-Espagnol's unique placement points to its distinct characteristics. Although dendrograms visualize relationships and do not directly imply evolution or ancestry, this analysis, based on linguistic data, offers insights into the connections among these languages, which might differ if analyzed using another method, such as a UPGMA tree, due to different underlying assumptions.

**The difference between UPGMA and Neighbouring tree for Dollo is:**

Dice distances under the Dollo model-constructed UPGMA dendrograms and Neighbor-Joining trees, respectively, that are visual representations of cluster analyses. The UPGMA dendrogram's lack of discernible branches refers to problems with the underlying distance matrix, such as inadequate variance between data points or incorrect scaling. A blank Neighbor-Joining tree Technical differences between the are probably due to how the two algorithms handle distance calculations differently: Neighbor-Joining allows for different rates across lineages and typically produces an additive or semi-additive tree, whereas UPGMA assumes a constant rate of evolution (hence its ultrametric tree).

## Part-2

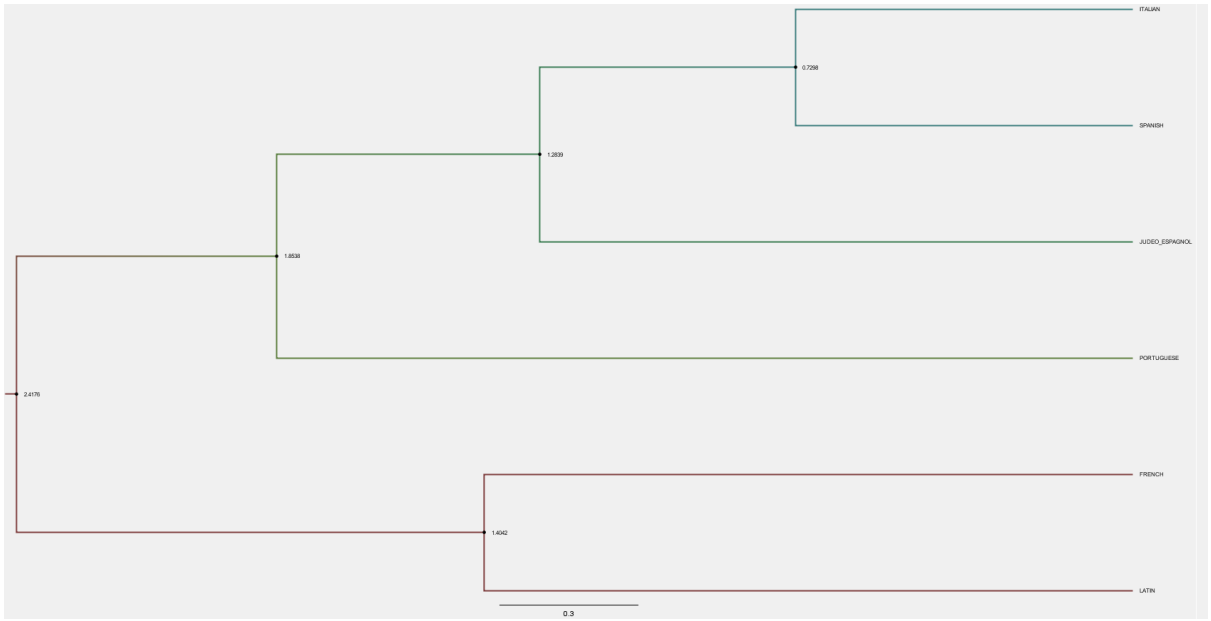
The phylogenetic tree which is used to show the evolutionary relationships among various languages within the Romance language family, with Latin as their common ancestor. These types of trees are similar to those used in biological studies to show the evolutionary relationships among species.

To analyze the tree, we look at the branching pattern and the length of the branches which often represent the amount of change or the time elapsed. The branching points, or nodes, can represent hypothetical common ancestors and the divergence times from those ancestors.

In the tree, Latin appears to be at the base, suggesting that it is the ancestral language from which the others diverged. The lengths of the branches leading to each language might indicate the relative amount of change or time since they diverged from their common ancestors. The tree also appears to have specific values at the nodes, which could represent divergence times calculated in the model, although without specific time units or scale mentioned, it's difficult to interpret these values precisely.

The tree also shows that some languages, like Judeo-Espagnol and Spanish, are closely related with a recent common ancestor, while others, like Italian and French, might have branched off earlier from the root. This pattern of branching reflects the historical and linguistic developments that have occurred since the languages diverged from Latin.

Understanding this tree requires knowledge of historical linguistics, the history of the Romance languages, and the methods used to infer phylogenies from linguistic data. This can include examining phonological, grammatical, and lexical changes that have occurred over time, and comparing these changes across different languages to establish relationships and divergence times.


ITALIAN
0.7298
SPANISH
1.2839
JUDEO_ESPAGNOL
1.8538
PORTUGUESE
2.4176
FRENCH
1.4042
LATIN
0.3