Katarzyna Michalska, 244577IV

1 Exercise 1

In this task, a dataset [2] containing text articles from different categories was analyzed to identify classes. The data was cleaned by removing stop words, punctuation, numbers, and short words. Stemming was applied to reduce words to their root form. The dimensionality was reduced by Rtsne, then the Support Vector Machine was employed to classify the data points and visualize the decision boundaries. The plot below illustrates both the classified data points and the corresponding decision boundaries.

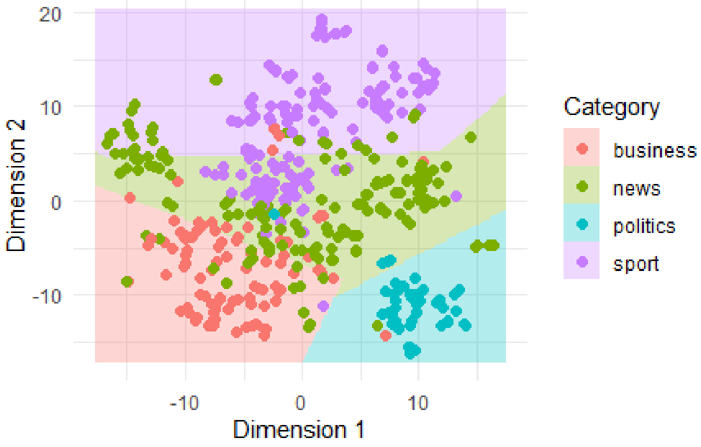

Figure 1: Decision boundary and classified data points

Next, the centroids for each category were calculated and plotted, as shown below.

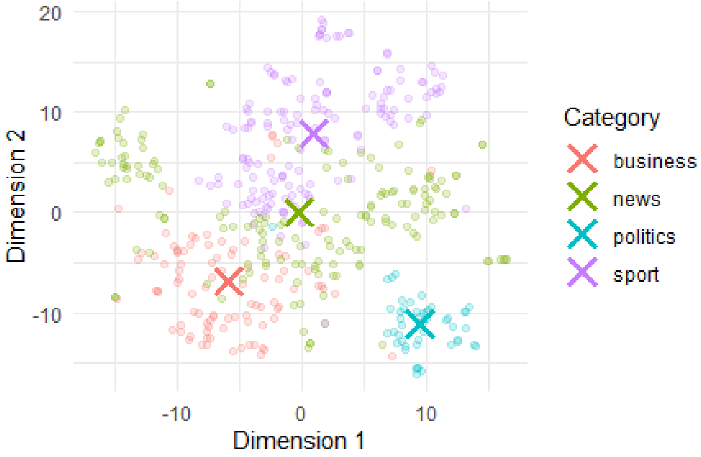

Figure 2: Centroids of classes.

2 Exercise 2

The goal was to create functions that compute metrics such as local clustering coefficient, degree centrality, degree prestige, gregariousness, closeness centrality, betweenness centrality, common neighbor measure, and Jaccard measure. The metrics that had their built-in counterparts were compared with them. In each such case of comparison, the results of the custom method and the built-in one were found to be consistent. The implemented functions were tested on sample data [1] and on the custom graph. The plot below presents visualization of the graph [1].

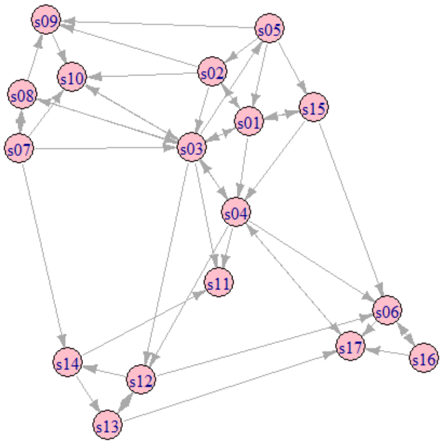

Figure 3: Graph visualization

When analyzing the output of the metrics, the first to consider is degree centrality, with the highest value assigned to node s03, which is not surprising, as this node is visually the best connected to the others. Then, the output of the local clustering coefficient shows that node s16 has the highest value of 1, indicating it forms a complete triangle with its neighbors. This metric was tested after converting the graph to its undirected form. The highest values for prestige, betweenness centrality and gregariousness belong to node s03, indicating that it holds a significant influence and central role in the network compared to other nodes. The proximity prestige has the greatest value for s04, which means that node s04 is highly accessible from other nodes.

To evaluate the closeness centrality, a new custom graph was generated to facilitate easier verification of the outputs. The graph is presented below. The highest values for closest centrality belong to node 2, which means that this node has the shortest average distance to all other nodes in the graph, making it a key point in terms of connectivity and influence.

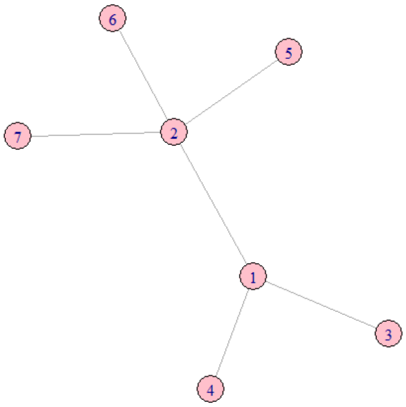

Figure 4: Graph visualization

The common neighbor measure and Jaccard measure were calculated between s01 and s02 on the first graph [1]. In the first case, we obtained a value of 2, which can be easily verified as correct, and in the second case, the value was 0.25.

3 Exercise 3

In this task, the reservoir sampling method was implemented to simulate the processing of streaming data. A dataset consisting of 3000 points was created, divided into 10 time periods, introducing concept drift by gradually changing the mean values across periods. Three sampling scenarios were implemented: the first, where the reservoir size matches the entire data stream, the second, where the reservoir is smaller than the data stream without adjusting for drift, and the third, where the reservoir is smaller, and the sampling adjusts to favor newer points due to concept drift. Visualizations were created to illustrate the distribution of points in both the data stream and the reservoir, reflecting changes over time. The plot below presents the first scenario, where the reservoir size matches the entire data stream.

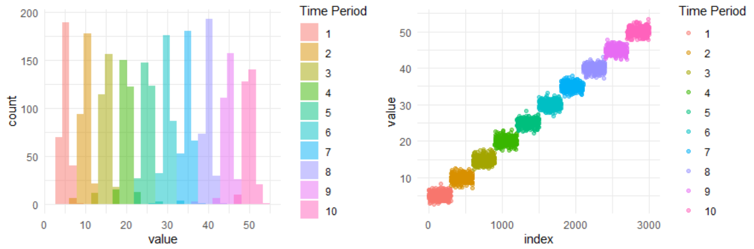

Figure 5: Full Reservoir

The next plot below presents the second scenario, where the reservoir is smaller than the data stream without adjusting for drift. As we can see, the data points are uniformly distributed.

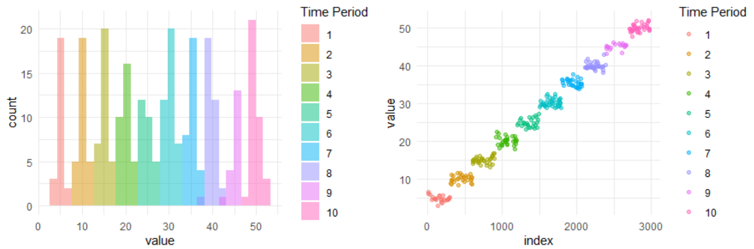

Figure 6: Small Reservoir (No Drift Adjustment)

The last plot below presents the third scenario, where the reservoir is smaller, and the sampling adjusts to favor newer points due to concept drift. As we can see, the majority of the chosen data comes from the most recent period.

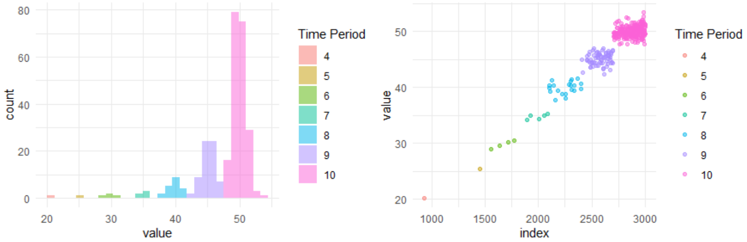

Figure 7: Small Reservoir (With Drift Adjustment)

References

[1] kateto. Network Workshop NetSciX. https://github.com/kateto/R-igraph-Network-Workshop-NetSciX, 2016.

[2] Hadas Unger. CNN News Articles from 2011 to 2022. https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning?resource=download-directory, 2021.