SDS 3786 Lab 6: L'apprentissage automatique et les règles d'association

Professor: Patrick Boily
Anthony Le, John Bute, Kelly Gao

Introduction:

In order to provide a proper understanding of the datasets and the association rules between all data columns, we had decided to approach this task by using our three derived datasets, which are aggregated based on the employee's work, the region and the associated embassy. Working through these datasets allows us to approach the association rule analysis much more easily in order to determine the strong association rules for the datasets. Our goal is to create these association rules using the apriori algorithm to do so, however, due to some issues, we were not able to complete the objective of this task.

Approach:

Before approaching the task, the main question came down to whether we use the original datasets (PIMENTO_CASES and PIMENTO_PROGRAMS) or our derived datasets to perform the Association Rule Analysis, which by choosing either one, would have a variety of pros and cons associated with it. If our derived datasets were not perfect, using the original datasets to perform the Association Rule Analysis would allow us to preprocess the data to transform it into its transaction-item matrix. This would allow us to gain another derived dataset to be used in the future. However, because our derived datasets very closely match what we need for the apriori algorithm, we decided simply working on the derived datasets would be a lot simpler than using the original ones.In order to create the transaction-item matrix, we would have to determine a transactionID in order to stamp each row with an indicator for the apriori algorithm to be properly implemented. This step however had some issues and was not able to be properly implemented in time for submission.

Results:

Unfortunately, due to some issues regarding the pre-processing, we were not able to complete this lab and were not able to determine the association rules for the dataset. We were not able to properly implement the needed steps to create the transaction-item matrix and were thus not able to run the apriori algorithm to determine the strong association rules for the dataset. In future labs, we hope to properly understand the preprocessing in order to not have issues with this in the future, and hope that we are able to implement machine learning tasks for the future to come.