THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Isolation of Business Logic Represented by ETL Processes by Machine Learning Algorithms**
Author's name:	**Žilt Juraj**
Type of thesis :	bachelor
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Cybernetics
Thesis reviewer:	Yuanhong Wang
Reviewer's department:	Faculty of Electrical Engineering

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The objective of this thesis is to develop a method for identifying Business Intelligence (BI) tasks from SQL source code, with the ultimate goal of extracting business logic from these scripts. This is a challenging task due to several factors. Firstly, SQL scripts are highly diverse, implemented on different platforms and written by numerous programmers using various methodologies. Secondly, there is limited or no prior information available for these scripts. Finally, the complex structure of these scripts adds to the difficulty of analysis.

Fulfilment of assignment **fulfilled with minor objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The main outcomes of this thesis are the clusters of SQL snippets and the decision features used to classify each cluster. These decision features are represented by a set of COMPARISON and IDENTIFIER commands in SQL that are shared by each cluster. While these features may contain implicit business logic, they are not presented in an explicit form, e.g., we still don't know the exact functionality of an SQL snippet.

Methodology **outstanding**

Comment on the correctness of the approach and/or the solution methods.

The methodology of this thesis is based on several techniques in machine learning: The Uniform Manifold Approximation and Projection (UMAP) is used for dimension reduction later analyzes the identified structures; The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) then performs clustering on the reduced data from UMAP; The Decision Tree is finally utilized to interpret the business logic with each cluster produced by HDBSCAN. The chosen pipeline of dimension reduction, clustering, and interpreting is appropriate for the task at hand. The selected techniques for each step are also well-suited to the task.

Technical level **B - very good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

Most aspect of the method employed in this thesis is explained clearly. The only part confusing me is the "point shifting" in UMAP on Page 11.

Formal and language level, scope of thesis **B - very good.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The thesis is organized in a logical way. The thesis is well-presented, with clear text, appropriate headings and subheadings, and well-designed figures and tables.

Selection of sources, citation correctness **D - satisfactory.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

The primary contribution of this thesis is the development of an integrated pipeline for extracting business logic from BI scripts. Each step in the pipeline is thoroughly reviewed and discussed by the student, including an elaboration of the reasoning behind the selection of specific methods. While the techniques are cited, some of the sources provided, such as the Decision Tree and HDBSCAN, only offer online tutorials rather than original papers.

Regarding the statement in Chapter 3 that no public solution to similar works was found and therefore no comparison or inspiration could be drawn from other papers or theses, I find it unsatisfactory. This is because there is a vast amount of literature available on the analysis of SQL queries, such as "Text Mining Applied to SQL Queries: A Case Study for the SDSS SkyServer," which is not discussed in this thesis.

Therefore, I suggest rephrasing the statement in Chapter 3 to acknowledge the extensive literature on the analysis of SQL queries that exists, while highlighting the unique contribution that this thesis makes to the field. Additionally, I recommend providing more complete and relevant citations for the techniques used in the pipeline, including the original papers where possible.

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

Please insert your comments here.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

The grade that I award for the thesis is **B - very good.**

The thesis addresses a challenging and significant problem in recognizing the business logic from BI transformation codes, specifically SQL scripts. The proposed pipeline of tokenizing, discovering common snippets, dimension reduction, clustering, and interpreting is well-developed and thoroughly discussed. The techniques used for each step of the pipeline are compared and evaluated, and the final implementation and empirical results demonstrate the effectiveness of the proposed method in finding business logic from SQL scripts.

While the related work of SQL query analysis is not discussed, I believe that the novelty of the proposed method is sufficient for a bachelor's thesis.

Overall, this thesis provides a valuable contribution to the field and demonstrates the student's ability to develop and evaluate a complex pipeline for addressing a challenging problem.

Questions and other comments:

- Is GST code in Figure 2.3 correct? maxMatch and max_Match are supposed to be the same.
- Display the equation just after the reference, e.g., 2.3, and remove the citations on equations.
- Is Sec 4.2 necessary? It seems that the subsequent process does not require the similarities between tiles.

Date: **13.6.2023** Signature: