Aalto University
MS-E2177 - Seminar on Case Studies in Operations Research

Semantic risk clustering

Project plan

Santeri Paljakka (Project Manager)

Essi Nikula

Henrik Purokoski

Jaakko Paavilainen

Antti Kärkkäinen

March 27, 2026

Contents

1 Introduction

Inclus is a growth company specialized in participatory, visual, and interactive cloud software designed for collaborative and complex risk management. Inclus helps its clients to understand their risk landscape by mapping different risk scenarios and providing related software tools for risk analysis and management. One of the used tool in the process is risk registers. Risk registers are databases that maintain an accurate and current list of possible risk scenarios the customer has identified. The risk registers contain data entries, such as risk labels and textual descriptions of the risks.

In qualitative risk management, risks are primarily documented in written form. The unstructured textual data creates challenges for data analysis with traditional measures. Additionally, contribution by multiple stakeholders easily leads to several inconsistencies in the data, such as redundancy, and varying formats. When conducted manually, perceiving the whole risk landscape and understanding closely connected risks becomes an unfeasible task as the risk register can have thousands of entries. Furthermore, traditional lexical clustering or keyword matching often fails to capture the informal information and linguistic themes within textual data (Kuhn et al., 2007). Grouping similar risks, identifying duplicates and creating meaningful visualizations would have real value for the usability of risk register data.

The recent development of Large Language Models (LLM) and Natural Language Processing (NLP) has enabled new possibilities for processing textual input and qualitative data. Methods such as Information Retrieval (IR), transformer-based technologies and, most recently, LLM embeddings can be used to process textual input to vectors, enabling comparison with different vector similarity measures. Clustering based on the semantic similarities of textual documents has been proposed as a part of topic modeling pipelines, for example by Grootendorst (2022) and Mersha et al. (2024). However, these advanced NLP methods are not yet studied in detail using actual risk data and risk registers.

As presented in a recent thesis work by Westergård (2025), Inclus has already been utilizing these NLP techniques with risk register data by developing LLM capabilities, such as AI agent using Retrieval-Augmented Generation (RAG). Additionally, an initial trial of semantic risk clustering pipeline has proven some potential in finding structure in the qualitative risk data. The initial trial shows that semantic clustering offers a way to tackle the current issues with large and unstructured risk registers while providing new kinds of insights about the risk landscape. For the development of this new tool,

comprehensive background study and rigorous testing is required to design a tool that employs the best practices in the field

2 Objectives

The objective for this project is designing a semantic risk clustering pipeline for risk descriptions in risk registers, while comparing the performance of different technologies and design choices. At the end, we are presenting results, limitations and capabilities, and recommendations for building an usable and reliable semantic risk clustering tool.

The process is composed of two main components: Building an extensive literature review to find industry best practices, and a practical testing using provided risk data set, selected embedding models and clustering algorithms. Our study focuses on the comparison of the technologies and identification of optimal design choices rather than software development or coding.

3 Tasks

It is beneficial to divide the project into tasks to structure the workflow and to help achieve the project goals. The first task is to familiarize with the problem and identify the underlying methods and theory. Subsequently, the following tasks presented below are completed, while working towards the final report.

3.1 Literature review

The theoretical background forms a significant part of the entire project. Therefore, the aim is to allocate sufficient time for researching articles related to each topic and using 4-6 most relevant sources for each topic when writing the literature review. These five key topics discussed are risk registers, semantic clustering, LLM embeddings, vector databases, and clustering algorithms. Responsibilities for the literature review are divided by topic, but each member is also assigned another subject to be reviewed. The literature review also aims to identify possible and suitable methods for implementing project's upcoming embedding and clustering metrics.

3.2 Embedding implementation and testing

This task involves converting risk descriptions into vectorized formats and comparing the performance of different embedding models. We evaluate various data preprocessing and homogenization options to determine how different formatting and levels of detail in risk descriptions affect the results. Additionally, we examine and utilize the code provided by Inclus and use it as the basis for empirical tests.

3.3 Clustering implementation and testing

Following the familiarization with the embedding models, we implement and test various clustering algorithms for organizing risks. The primary objective is to determine the most suitable algorithm for processing high-dimensional semantic data and to define a method for finding the optimal number of clusters. Furthermore, we analyze the results to ensure that they perform as expected, focusing on how the clustering results can best be visualized and interpreted for the final tool.

3.4 Integration and model evaluation

A crucial task is to investigate how the choice of embedding model affects the final clustering outcome. The aim is to identify and compare combinations of vectorization and clustering algorithms that produce the most stable and logically coherent results. In addition, the scope of the problem will be adjusted as necessary based on the results found. We also assess whether the risk clustering tool could be usable without extensive knowledge of data science and what happens when a new risk is added.

3.5 Finalization and reports

The final report will be prepared and developed consistently throughout the project, with documentation taking place simultaneously alongside technical tasks. The aim is to ensure, that the final report provides comprehensive answers to all key research topics and identifies the most effective strategies for semantic risk management. We document computing resource requirements, record observations and provide actionable recommendations that could be useful for Inclus for the further development of the risk clustering tool. In addition to the key tasks of the project mentioned above, we produce three comprehensive report and presentation combinations, which will be presented during the course. These are this initial project plan at hand,

an interim report, and the final report.

4 Schedule

The planned schedule of the above tasks is presented in Figure 1. This schedule is subject to change as we gain a better understanding of the actual time required for each phase. It serves as general guideline for our progress and planning.

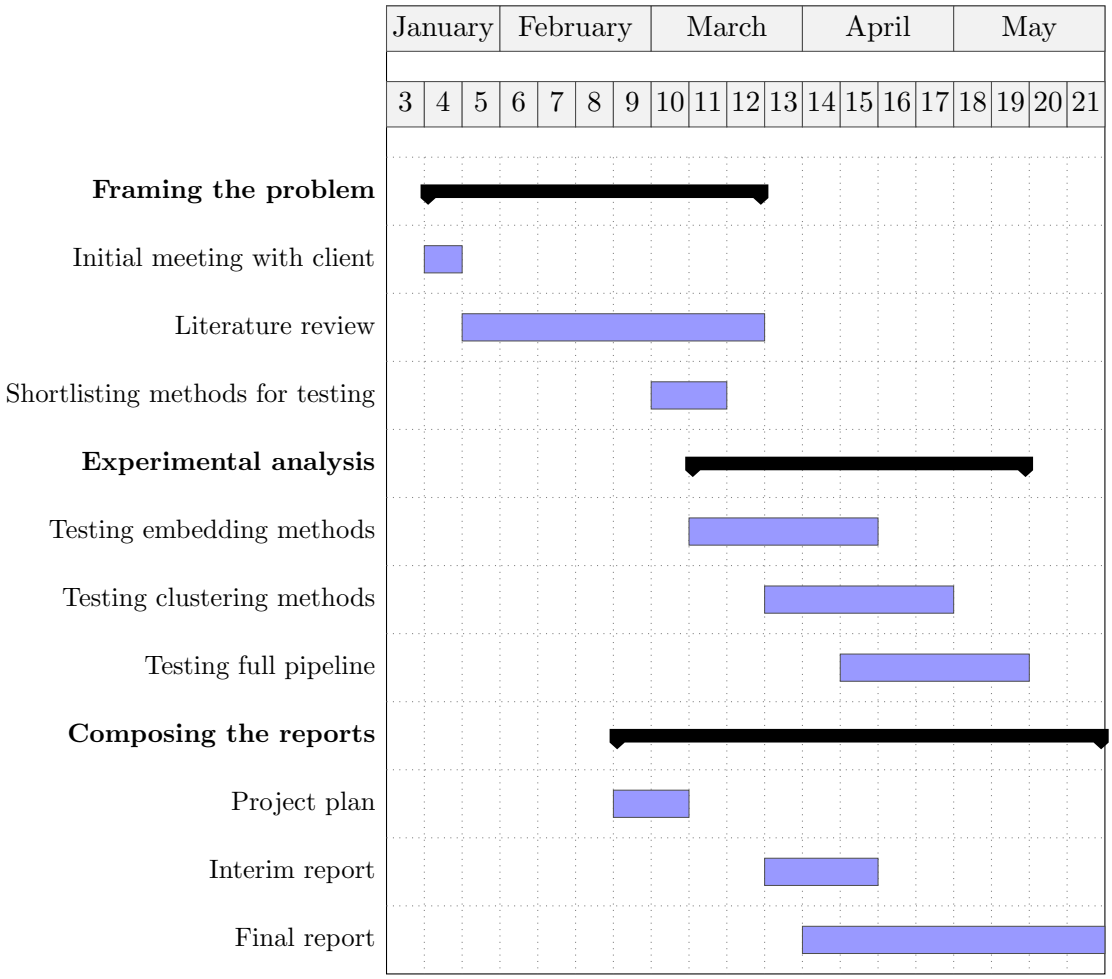

Figure 1: Project schedule

5 Resources

Our team consists of five master's level students majoring in systems and operations research providing a strong mathematical background as well as relevant knowledge in risk analysis and machine learning. We have got an orientation on the topic and the objectives by our Inclus contacts, Juha Törmänen (CTO) and Alexander Westergård, who both have prior experience on the seminar course which is particularly beneficial for receiving targeted guidance and aligning the project with course expectations and best practices. We will maintain regular communication with them throughout the project to clarify specifications and ensure an iterative workflow. Furthermore, professor Ahti Salo will be our general supervisor on the course guiding with project execution and following overall progress.

The project will be heavily relying to recent literature which is expected to be relatively easily available since the topic has attracted interest in academic research for the last years. From Inclus we have received a risk register dataset for the project along with access for RAG-based embedding models. Additionally, a demo pipeline was provided which has been very helpful for understanding the structure of the solution. This will serve as a starting point for our implementation, but will of course be adapted and extended when testing different approaches suggested in the literature.

To ensure steady progress, we decided to have weekly meetings on campus. In these meetings the goal is to plan the upcoming tasks, distribute workload and share our findings to keep everyone on track what has been done and what will the next steps be. The primary tools utilised for the project include Teams, Overleaf, Zotero and GitHub.

6 Risks

To ensure proper completion of the project while presenting satisfactory results for the client, it is important to identify possible risks, most likely of which are listed in Table 1. Since additional risks are possible, constant risk monitoring is conducted during meetings, and any new risks are addressed.

Risk	Effect	Likelihood	Impact	Mitigation
Problem definition too broad	No effective solution for the client; workload and resources stretched	Medium	High	Clear planning and communication with the team; clear goal definition in the planning phase
Clusters not intuitive for the risk context	Found solution not useful for the client	High	Medium	Use clustering and validation methods that encourage intuition in addition to cluster correctness
Scheduling problems	Some features might not have been implemented	Medium	High	Clear schedule and task allocation; regular meetings with the team
Communication challenges with the client	Solution does not meet the client's expectations	Low	Medium	Clearly define project scope and goals with the client; regular check-ins
Communication challenges within the team	Workflow not efficient / responsibilities not properly distributed	Low	Medium	Regular intra-team meetings; encourage questions; project manager monitoring and task assignment
Insufficient risk data	Model does not properly identify real-world risk clusters	Medium	Medium	Conduct exploratory data analysis on the given data; identify possible issues
Inadequate cluster validation	Model performance cannot be validated; model not satisfactory for the client	Low	High	Broad literature review of clustering validation methods

Table 1: Risk table describing the risks that are most likely

References

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022. URL http://arxiv.org/abs/2203.05794. arXiv:2203.05794 [cs].

Adrian Kuhn, Stéphane Ducasse, and Tudor Gîrba. Semantic clustering: Identifying topics in source code. *Information and Software Technology*, 49(3):230–243, 2007. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2006.10.017. URL https://www.sciencedirect.com/science/article/pii/S0950584906001820. 12th Working Conference on Reverse Engineering.

Melkamu Abay Mersha, Mesay Gemeda Yigezu, and Jugal Kalita. Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms. *Procedia Computer Science*, 244:121–132, 2024. ISSN 18770509. doi: 10.1016/j.procs.2024.10.185. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050924029867.

Alexander Westergård. Improving risk registers with an AI assistant. Master's thesis, Aalto University, 2025.