Artificial Intelligence (CSC3400) Project

Objective

For this group project you will present a research paper submission to the 2025 MSR Mining Challenge focused on analyzing software dependencies and ecosystems using the **Goblin framework**. The challenge provides a hands-on opportunity to explore questions related to **software dependency graphs**, ecosystem evolution, risk assessment, and more using **real-world datasets** from Maven Central.

Your project should use some of the NLP, ML, and AI techniques discussed in class. However, you can use any other algorithms, coding, data analysis, statistics concepts you deem necessary.

Important Dates

- **Proposal Deadline**: 09/23/2024
- **Project Deadline**: 12/03/2024
- **Abstract Submission**: 12/03/2024
- **Paper Submission**: 12/06/2024

Background

Using package managers is a simple and common method for reusing code through project dependencies. However, these direct dependencies can themselves rely on additional packages, resulting in indirect dependencies. It may then become complex to get a grasp of the whole set of dependencies of a project. Beyond individual projects, a deep understanding of how software ecosystems work and evolve is also a critical prerequisite for achieving sustained success in software development.

Challenge Overview

The analysis of a software ecosystem graph presents numerous research opportunities for you to explore, allowing for the investigation of various questions in areas such as structural analysis, community detection, dependency optimization, and risk assessment within the Maven Central ecosystem.

See the mining challenge page for a comprehensive list of potential project ideas `https://2025.msrconf.org/track/msr-2025-mining-challenge`

Getting Started

1. Familiarize Yourself with the Goblin Framework

- The Maven Central Neo4j dataset is available on this Zenodo archive.
- To import this dump into Neo4J, please use a dbs version 4.x.
- The Weaver project is available in our GitHub repository.
- This paper gives a more in-depth presentation of how the Goblin framework works.
- A project for simply setting up a Neo4j database and Weaver API using Docker is available here: GitHub Docker Setup.

2. Project proposal

In groups of 2-3 students, select or come up with a research questions or questions to answer. Then write a 1-page proposal containing:

- The composition of your team. All team members need to be listed with their full name
- A description of the project. Including the research questions to be answered and the concepts form the class being applied. Also list concepts from other classes that you will apply.
- A project plan and team responsibilities: This must include a detailed set of steps your team will follow in order to answer the research questions. Include as many details as possible about what you will do and how, and who will be responsible for each step of the project.

Proposal Due: September 23rd @ 11:59 pm

The proposal must be a 1-page .pdf conforming to the IEEE formatting instructions IEEE Conference Proceedings Formatting Guidelines (title in 24pt font and full text in 10pt type), use the LaTeX format provided **(See mining challenge website for more details)**.

3. Conduct Your Analysis

Write code, use the dataset to answer your research questions.

4. Project Report

You final deliverables for the project will be a challenge paper and a presentation. The challenge paper should should describe the results of your work by providing an introduction to the problem you address and why it is worth studying, the version of the dataset you used, the approach and tools you used, your results and their implications,

and conclusions. Make sure your report highlights the contributions and the importance of your work.

To ensure clarity and consistency in research submissions:

When detailing methodologies or presenting findings, authors should specify which snapshot/version of the Goblin dataset (and the Weaver version if used) was utilized. Given the continuous updates to the dataset, authors are reminded to be precise in their dataset references. This will help maintain transparency and ensure consistent replication of results.

The paper must conform to the IEEE formatting instructions IEEE Conference Proceedings Formatting Guidelines (title in 24pt font and full text in 10pt type), use the LaTeX format provided **(See mining challenge website for more details)**.

Deliverables

When submitting the project you must submit:

Your code

Either a GitHub repository will all your code or a Google Colab notebook with all your code. Your code must be properly documented ad structured into functions.

Challenge Paper

An Overleaf project with the Latex version of your paper.

- **Length**: 4 pages, with 1 additional page allowed for references.
- **Formatting**: Follow **IEEE Conference Proceedings Formatting Guidelines**.

Note: The papers will be submitted to the Mining challenge and will undergo a **double-anonymous review** process. If your paper is accepted, you will present your results at MSR 2025 in Ontario, Canada

Evaluation Criteria and Grading

Your project paper will be graded based on the following criteria:

- **Proposal (10 points):** The project proposal is complete and follows the format guidelines.
- **Analysis (30 points):** The analysis demonstrate a solid understanding of NLP, AI, or ML techniques.
- **Quality and accuracy of the results (10 points):** Presents clear, well-supported results based on the data.

- **Adherence to submission guidelines (10 points):** Follow the formatting and submission guidelines strictly.
- **Writing (10 points):** The paper is written clearly.
- **Code Quality (10 points):** The code supporting the analysis is properly documented and structured.
- **Presentation (10 points):** The oral presentation is given and clear, concise, and timely.
- **Abstract and paper submitted to MSR (10 points):** Ensure that your paper is submitted and includes a clear, concise abstract that outlines your research question and findings.

The project grade is based not only on completion of the project but on individual contributions of each members. All team members need to make significant contributions to all the items of the project. Lack of contribution from a team member will negative impact the individual grade regardless of overall completion of the project.

Use of Generative AI

The use of generative AI tools and technologies to create content and code is permitted to copy-edit and improve the writing and code. However, generative AI tools should not be used to generate full sections of the paper itself. Any use of generative AI must be fully disclosed in the Work (See ACM Policy on Authorship).

For example, the authors could include the following statement in the Acknowledgements section of the Work: During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Resources

- Goblin Neo4J Dataset on Zenodo
- Goblin Weaver GitHub Repository
- MSR 2025 Mining Challenge Website
- Look at previous years' challenges and the papers accepted: You can find the 2022 Challenge here and the papers that were accepted at: `https://conf.researchr.org/track/msr-2022/msr-2022-mining-challenge?#event-overview`

  You can find the 2021 Challenge here and the papers that were accepted at: `https://conf.researchr.org/track/msr-2021/msr-2021-mining-challenge?#event-overview`