Project Proposal: Role2Skill

What?

This project focuses on analyzing job vacancy datasets to identify which skills are associated with different job roles. By processing job titles and descriptions, the system will extract technical and soft skills and apply machine learning techniques to cluster and classify job roles based on their skill profiles.

The outcome will be a structured overview of how different professions are defined by combinations of skills, highlighting similarities and differences between roles.

Why?

Job descriptions often list many skills, but it is not always clear which of them truly define a profession and which are optional or redundant.
Students, junior specialists, and career switchers frequently struggle to understand which skills they should prioritize when preparing for a specific role.

This project aims to provide clarity by transforming unstructured vacancy texts into data-driven insights about skill demand and role composition.

Who?

Main Developer: Andrii Matviienko – responsible for data collection, analysis, and delivery.
Supervisors/Teachers – providing academic guidance and feedback.
Stakeholder: Vlad Luzianin, 4th semester ICT student representing the target group of entry-level job seekers

How?

The project will be implemented as a Python-based data analysis and machine learning pipeline. Job descriptions from LinkedIn and Indeed Kaggle datasets will be cleaned and processed using natural language processing techniques to extract skill-related information.

Machine learning models (such as clustering and classification algorithms) will be applied to group similar job roles based on extracted skills.
The results will be presented through a lightweight web interface or analytical reports showing skill distributions per role.

Data & Label

Primary Sources:

- LinkedIn Job Postings Dataset (Kaggle)
- Indeed Job Postings Dataset (Kaggle)

Core fields include:

- Job title
- Job description

Extracted labels will consist of:

- Skill lists per vacancy
- Skill-based role clusters

Risks & Mitigations

• Unstructured and noisy text data → Apply text cleaning, normalization, and filtering techniques.
• Skill ambiguity or synonym overlap → Use skill normalization strategies and embeddings.
• Model instability or unclear clusters → Start with baseline clustering and iteratively tune parameters.

•Dataset bias between platforms → Analyze LinkedIn and Indeed both separately and combined.

Scope

In Scope:

• Skill extraction from job descriptions.

• Clustering and classification of job roles.

• Analysis of skill demand patterns across datasets.

• Visual or tabular presentation of insights.

Out of Scope:

• Real-time vacancy scraping.

• Salary prediction or market forecasting.

• Personalized career recommendations.

• Resume analysis or user profiling.

Deliverables

• Cleaned and validated dataset.

• Extracted and normalized skill lists.

• Machine learning models for clustering and classification.

• Analytical report summarizing findings.

• Optional prototype or dashboard presenting results.

Iterations Overview

Iteration	Phases Included	Goal	Notes
Iteration 1	Phase 1 → Phase 2 → Phase 3	Get a working model that predicts (accuracy not important yet)	Focus on getting a proof of concept quickly.
Iteration 2	Phase 2 → Phase 3 → Phase 4 → Phase 5	Improve accuracy, finalize documentation	More time for model training, model accuracy, and presentation.

Planning

Project Phases and Timeline

Phase	Description	Start date	Finish date	Goal
Phase 1 – Proposal & Requirements	Problem understanding, define data needs, begin data collection.	20-11-2025	30-11-2025	Clear problem definition & initial dataset ready.
Phase 2 – Data Provisioning	Collect, clean, and prepare dataset for experimentation.	01-12-2025	12-12-2025	Usable dataset ready for modeling.
Phase 3 – Modeling (Iteration 1)	Build baseline skill extraction and clustering models	12-12-2025	22-12-2025	Working model (proof of concept).
WINTER BREAK		22-12-2025	06-01-2026	
Phase 4 – Modeling (Iteration 2)	Refine features, improve models, iterate evaluation and validation.	07-01-2026	13-01-2026	Refined model with structured data flow.
Phase 5 – Delivery	Final documentation, results consolidation, demonstration, and presentation	14-01-2026	20-01-2026	Final prototype & presentation ready.

Contribution per Course

Domain Analysis (DA):
Analyze job market structure, role definitions, and skill demand patterns.

Data Analysis & Information Architecture (DAIA):
Acquire, clean, structure, and visualize vacancy and skill data.

Machine Learning (ML):
Develop and evaluate clustering and classification models based on extracted skills.