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?

I chose this project because I was not confident which skills are actually important to become a good software engineer.

At the beginning of my studies, I was afraid of spending time learning outdated or irrelevant technologies.

Instead of relying on assumptions or general advice, I decided to analyze real job vacancies.

By working with real market data, I want to understand which skills are truly required in practice and how professions are formed through combinations of skills.

This project helps me better understand the current job market and gives me a more realistic view on which skills I should focus on as a student.

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 is implemented as a Python-based data analysis and machine learning pipeline.

Job descriptions from Kaggle datasets are processed using natural language processing techniques to extract skill-related information.

Clustering and classification models are applied to group job roles based on extracted skills.

The result is a model where a job title can be provided as input and the system outputs the top required skills for that profession.

During Iteration 1, the focus shifted from planning to implementation.

More than ten job vacancy datasets were explored and compared.

Selected Russian-language datasets were translated into English and unified into a single structure.

The data was cleaned, skills were extracted, clustered into profession-based skill sets, and a classification model was built to predict skill sets from job titles.

The results were evaluated and form the basis for improvements in Iteration 2.

Domain Understanding

This project is situated in the domain of the online job market, specifically IT and technical job vacancies.

In this domain, job roles are mainly described using free-text job descriptions that include responsibilities and lists of required skills.

A key problem in this domain is that skills are described in an unstructured and inconsistent way.

The same skill can appear under different names, abbreviations, or formulations.

Because of this, it is difficult to compare job roles or clearly understand which skills define a profession.

Several actors are involved in this domain.

Employers publish job vacancies and describe required skills.

Job platforms host and distribute vacancy data.

Job seekers, especially students and junior developers, use this information to decide which skills to learn.

The scope of this project is limited to analyzing job roles through required skills and comparing professions based on skill composition.

The project does not include salary prediction, labor market forecasting, real-time job scraping, or personalized career advice.

Criteria of Success:

• Job descriptions can be transformed into structured and normalized skill sets.

• Similar job roles are grouped together based on their skill composition.

• The clustering results show clear and interpretable relationships between professions and skills.

• Given a job title, the model can output a reasonable and relevant list of top required skills.

• The results provide useful insights for students and junior developers about skill requirements in the job market.

Analytic Approach

The analytic approach of this project is based on machine learning applied to job vacancy data.

The goal is to represent job roles using structured and normalized skill sets extracted from job descriptions.

The project follows an iterative AI and machine learning pipeline:

1. Data exploration and selection: to compare multiple job vacancy datasets and select relevant sources.

2. Data preprocessing: to clean, translate, and unify job titles and job descriptions.

3. Skill extraction: to identify technical and soft skills from unstructured text.

4. Skill normalization: to reduce duplication caused by synonyms and inconsistent naming.

5. Skill clustering: to group related skills into skill sets representing professions.

6. Skill set classification: to predict the most likely skill set based on a job title.

7. Evaluation and iteration: to assess results and improve data processing and models in a second iteration.

Data & Label

Primary Sources:

- Hh.ru vacancies dataset (Kaggle)
- Yandex vacancies 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.