



Artur Correia Romão

**Desenvolvimento de um sistema computacional para
determinação das competências ESCO associadas a
ofertas formativas**

**Development of a computational system for
determining ESCO competences associated with
training offers**

DOCUMENTO PROVISÓRIO



Artur Correia Romão

Desenvolvimento de um sistema computacional para determinação das competências ESCO associadas a ofertas formativas

Development of a computational system for determining ESCO competences associated with training offers

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica do Doutor António José Ribeiro Neves, Professor Associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e da Doutora Fabianne de Araújo Ribeiro, Investigadora do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

o júri / the jury

Palavras Chave

Processamento de linguagem natural, grandes modelos de linguagem, taxonomias, oferta formativa e educativa, competências, sistemas de informação.

Resumo

Keywords

Natural language processing, large language models, taxonomies, training and educational offers, skills, information systems.

Abstract

Contents

Contents	i
List of Figures	iii
List of Tables	v
List of Code Excerpts	vii
Glossary	ix
1 Introduction	1
1.1 Objectives	2
1.2 Document Structure	2
2 State of the Art	5
2.1 Taxonomies for Classification of Occupations, Skills and Competences	5
2.1.1 ESCO	6
2.1.2 ISCO	7
2.1.3 O*NET	8
2.2 Large Language Models (LLMs)	10
2.3 Applying NLP and LLMs to Occupation and Skill Taxonomies	12
2.3.1 Skill-Span Extraction	12
2.3.2 Zero-Shot Matching	12
2.3.3 Distant Supervision	13
3 Methodology	15
References	21
A Additional content	23

List of Figures

2.1	ESCO's Landscape in September 2022 - The taxonomy currently contains information about 3,008 occupations and 13,890 skills [1].	6
2.2	Example of an ESCO's Web Service page [9].	7
2.3	Role of ISCO-08 in the hierarchical structure of the ESCO occupations pillar [12].	8
2.4	The O*NET® Content Model schematic representation [14].	9
2.5	ESCO-O*NET crosswalk methodology [13].	10
2.6	Example of a Zero-Shot prompt and answer [19].	11
2.7	High-level overview of the full Zero-Shot process extracted from <i>Clavié and Soulié (2023)</i> [16].	13
2.8	Schematic representation of Negative-Sampling processes, extracted from <i>Decorte et al. (2022)</i> [23].	14
2.9	Pipeline for Fine-grained Danish Skill Classification in Kompetencer, extracted from <i>Zhang et al. (2022)</i> [22].	14
3.1	Diagram representing the system's pipeline, with different possible flows.	15
3.2	Example of an UA's DPUC web page for the course of Programming Fundamentals [28].	17
3.3	Gantt diagram containing the tasks that were and will be developed during this dissertation.	19

List of Tables

List of Code Excerpts

Glossary

ESCO	European Skills, Competences, Qualifications and Occupations	LLMs	Large Language Models
API	Application Programming Interface	NLP	Natural Language Processing
UA	University of Aveiro	O*NET	Occupational Information Network
ISCO	International Standard Classification of Occupations	DPUC	Pedagogical Dossier

Introduction

Nowadays, technological advances and continuous knowledge have established a permanent need for lifelong learning, delivered through specialization courses and, more recently, micro-credentials. This importance is based on the ability of such courses to keep professionals up to date with technological and knowledge developments, guaranteeing the updating of existing skills (upskilling) or the acquisition of new ones (re-skilling) [1], [2]. However, the business sector is currently facing a challenge in trying to identify the specific training courses that can fill the gaps in the skills of its employees [3].

Currently, there are actions to co-create training offers between Higher Education Institutions (HEIs) and companies in the same business community with the aim of creating training offers designed to keep up with the constant changes in the world of work.

However, the difficulty in linking a training offer (by its title, description or objectives) to the skills acquired at the end of its successful completion, still represents a barrier in the process of people and companies choosing a training offer.

To this end, the European Commission recently made available a database containing the multilingual taxonomy of European Skills, Competences, Qualifications and Occupations (ESCO) [1]. ESCO acts as a dictionary that describes, identifies and classifies professional occupations and skills relevant to the EU labour market and the education and training sectors, providing descriptions of 3008 occupations and 13 890 skills associated with these occupations, translated into 28 languages (those of the EU as well as Icelandic, Norwegian, Ukrainian and Arabic).

Even though great progress has been made in skills matching with the emergence of ESCO, the current limitations of the taxonomy's Application Programming Interface (API) [4] become apparent when it fails, on its own, to accurately return a list of ESCO skills for texts related to training or educational offers, often suggesting unrelated skills.

Thus, the idea for this dissertation was born: to develop a computational system capable of determining ESCO competences associated with training offers.

Therefore, in order to address the above-mentioned gaps, several approaches will be

discussed in this dissertation, especially the ones using Large Language Models (LLMs), due to their powerful text processing capacities, role-playing ability and human language comprehension. Those will serve as inspiration for the development of the proposed system's pipeline.

1.1 OBJECTIVES

The main goal of this dissertation is to develop a computational system capable of processing training offers' information, and map that data to ESCO skills, also taking advantage of Large Language Models (LLMs) and the ESCO API for that purpose. Our case study will be based on the Pedagogical Dossiers (DPUCs) and course description of the micro-credentials implemented at the University of Aveiro (UA).

A DPUC is an analogous of an identity card for UA's courses. It contains many fields that describe the educational offer, such as the contents, objectives, learning outcomes, requirements, recommended bibliography, and evaluation methodology.

Regarding the training offer in the micromodule format, herein designated as micro-credentials are, as per definition of the European Commission, "the record of the learning outcomes that a learner has acquired following a small volume of learning. These learning outcomes will have been assessed against transparent and clearly defined criteria" [5].

Therefore, the specific objectives of this dissertation can be summarized as follows:

- Implementation and testing of a system to manage the skills of UA's educational offer and to match them to ESCO skills.
- Development of a pipeline that connects the ESCO framework (API), UA's DPUCs and an LLM framework.
- Survey the State of the Art regarding occupations and skills taxonomies across the world and in national contexts.
- Study and further integrate the ESCO API in the system's pipeline.
- To test several LLM frameworks and NLP libraries to assess their applicability in the task of mapping UA's educational offer to ESCO skills.
- To deploy the system's pipeline in order to automate the skills matching process.
- Evaluate the system's performance, trusting on Course Directors' manual verification of the skills' matching.

Finally, the ultimate goal of this dissertation is to provide the UA academic community with a trustable framework that maps the institution's educational offer to standardized and specific skills of the European occupations and skills taxonomy (ESCO). This will help current and future students to have a better understanding of the University's educational offer and companies' Human Resources representatives to acknowledge and recognize UA's former students (alumnis) skills upon hiring them.

1.2 DOCUMENT STRUCTURE

This dissertation is composed by two more chapters. The next chapter, Chapter 2, offers an overview of the existing taxonomies in the field of occupations and skills and a discussion

about features of Large Language Models that can be used to simplify the skills matching process. This chapter also provides the analysis of current approaches using these models for skill extraction.

In Chapter 3, the methodology of the work developed throughout this pre-dissertation, as well as a projection of future work for the next semester is presented. This Chapter also includes a Gantt diagram containing the tasks that were and will be developed during this dissertation work.

State of the Art

2.1 TAXONOMIES FOR CLASSIFICATION OF OCCUPATIONS, SKILLS AND COMPETENCES

The collection of data referring to occupations, skills and competences from a variety of sources resulted in the creation of taxonomies, which are structured databases for occupational classifications [6].

The classification through various taxonomies plays a crucial role in the modern labour market and educational landscape. They serve as structured frameworks to categorize and describe occupations and the skills required for them in a standardized way, facilitating communication, analysis, and planning across different sectors and geographical boundaries.

Nowadays, their importance is undeniable, so we highlight the following advantages [2], [7]:

- Standardization and clarity - these taxonomies provide a common language for describing jobs and skills which is particularly important in a globalized economy where workers, employers, educators and lifelong learners operate across different regions and countries. For example, if an individual completed a Bachelor's degree in Portugal and a Master's degree in France and, for some reason, wants to work in Germany, these taxonomies would help German employers to understand which skills the individual possesses and recognize their higher education degrees.
- Labour market analysis - taxonomies enable government agencies, economists, and researchers to analyze labour market trends, track occupational shifts, and predict future workforce needs. This information is vital for policy making, especially in areas like employment, education, and immigration.
- Career development and guidance - For individuals, these taxonomies offer valuable insights into the skills and qualifications required for various occupations, aiding in career planning and development. They are particularly useful for lifelong learners who are looking to update their skills or change career paths.
- Recruitment and human resource management - Enterprises utilize these taxonomies to define job roles, assess skill requirements, and align their workforce with organizational

needs. This alignment is crucial for effective human resource management and strategic planning.

- Educational planning and curriculum development - Education institutions use these taxonomies to align their curricula with the current and future needs of the labour market. This ensures that the skills and knowledge they impart are relevant and meet industry standards.

In this section, some examples of taxonomies will be presented with special focus on ESCO since it will be the matter of study of this dissertation.

2.1.1 ESCO

The definition provided by the European Commission for ESCO states “European Skills, Competences, Qualifications and Occupations (ESCO) is the European multilingual classification of Skills, Competences and Occupations. ESCO works as a dictionary, describing, identifying and classifying professional occupations and skills relevant for the EU labour market, education and training” [1]. This taxonomy allows computational systems to use the three pillars of ESCO (occupations, skills and qualifications) [8] and the relationships between them to address different use cases, such as matching job seekers to jobs on the basis of their skills, suggesting training programs to lifelong learners, understanding which qualifications are demanded or often requested by employers for working in a specific occupation, among others.

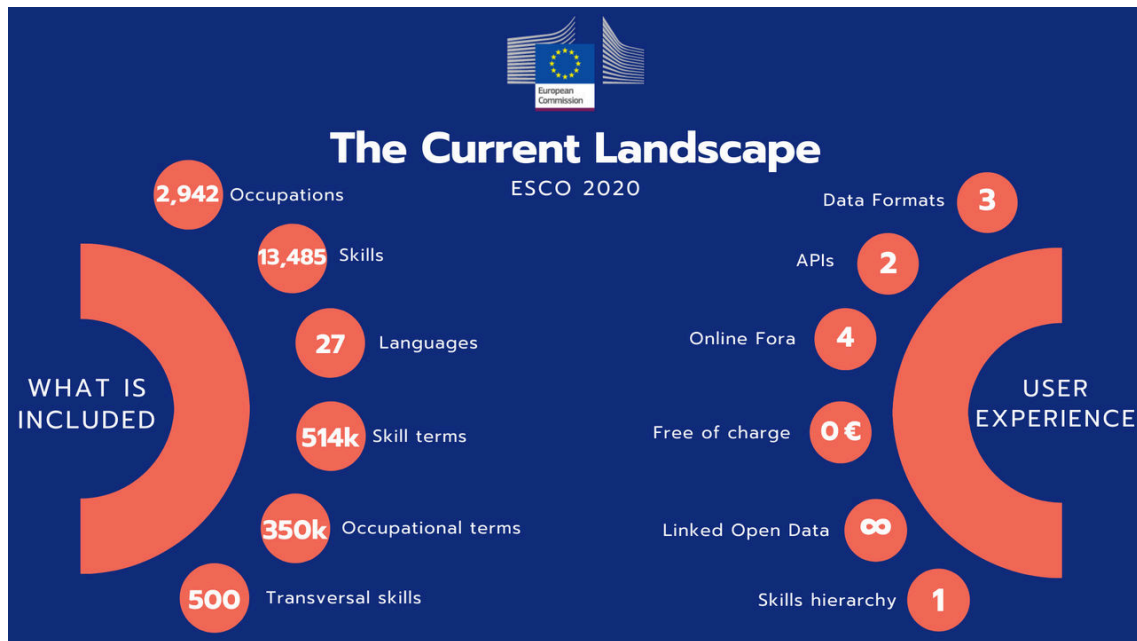


Figure 2.1: ESCO’s Landscape in September 2022 - The taxonomy currently contains information about 3,008 occupations and 13,890 skills [1].

People and enterprises change jobs and employers more frequently than in the past, thus new skills are regularly needed while geographic and job mobility increases. Both employers and job seekers are turning to digital methods for posting and applying for jobs, as well as for seeking and providing training opportunities. Companies and educational institutions

need clear and updated information on skills and qualifications to better address skills gaps in education [3].

Therefore, the aim of ESCO is to support job mobility across Europe and, consequently, achieve a more integrated and efficient labour market, by offering a “common language” on occupations and skills that can be used by different stakeholders on employment, education and training topics [1].

ESCO’s classification can be accessed through two types of APIs: a web-service API, available online and a Local API which, as the name suggests, needs to be installed locally [4].

The ESCO Web Services API is a web-based service that offers access to various versions of the ESCO classification, encompassing functionalities that address most ESCO business use cases. Featuring an easy-to-use web interface for linked data (disposed hierarchically), each concept is identified as an URI, making it a perfect setup for non-technical users. This service allows text search for Occupations, Skills & Competences and Qualifications.

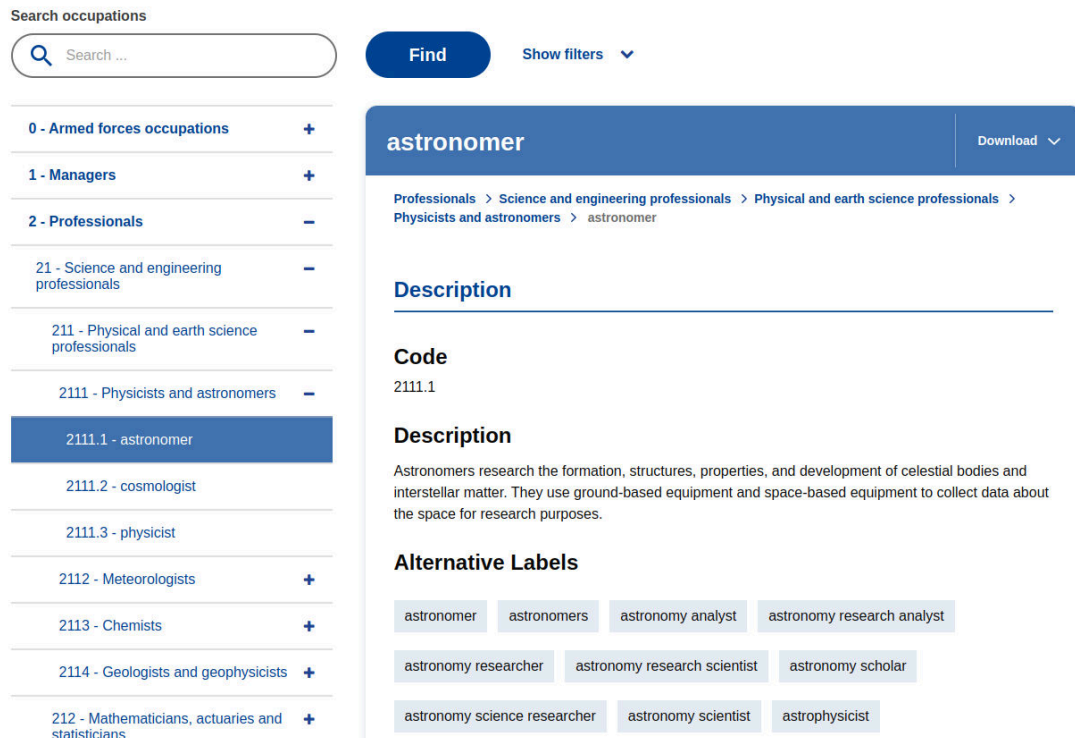


Figure 2.2: Example of an ESCO’s Web Service page [9].

The ESCO Local API is a downloadable version of the ESCO API, which can be installed on a computer or server, providing local access to ESCO’s information. Compared to the Web Services API, this local version assures increased performance and the independence from the availability of the service provided by the European Commission.

2.1.2 ISCO

The International Standard Classification of Occupations (ISCO) is a globally recognized framework developed by the International Labour Organization (ILO) to categorize and

organize occupations [10]. Similar to ESCO, the ISCO serve as a vital tool for labour market analysis, providing a standardized system for classifying jobs based on the skills, education, and duties involved. ISCO is structured hierarchically, with major groups divided into sub-major groups, minor groups, and unit groups, each with different levels of detail and specificity. The two latest versions of ISCO are ISCO-88 (dating from 1988) and ISCO-08 (dating from 2008) [11]. ESCO was actually influenced and inspired by ISCO, particularly in its approach to structuring occupations. ESCO maps each occupation to an ISCO-08 code and uses ISCO-08 as a hierarchical structure for its occupations pillar, indicating a direct link and a level of compatibility and interoperability between the two systems [12].

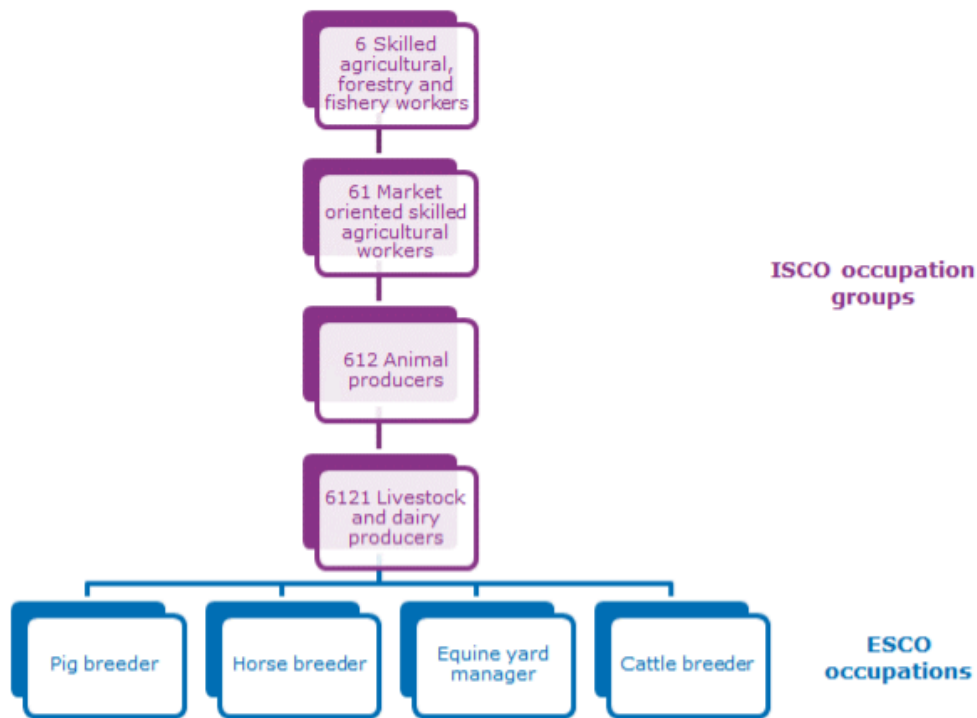


Figure 3: The structure of the occupations pillar

Figure 2.3: Role of ISCO-08 in the hierarchical structure of the ESCO occupations pillar [12].

The alignment with ISCO-08 enables ESCO to:

- Maintain consistency with a globally recognized occupational classification system, ensuring harmonization with international standards.
- Enhance its utility for the European labour market - facilitating comprehensive market analysis and mobility across borders.

2.1.3 O*NET

Occupational Information Network (O*NET) is the North American equivalent of ESCO, developed under the sponsorship of the U.S. Department of Labor/Employment and Training Administration (USDOL/ETA) [13], comprising occupations from the Standard Occupational Classification (SOC) system and their corresponding skills, knowledge, and abilities [8]. Similar

to ESCO and ISCO taxonomies, O*NET aims to provide a comprehensive database of job characterisation and work skills that standardizes and categorizes the labour market, in this case, of the United States scenario.

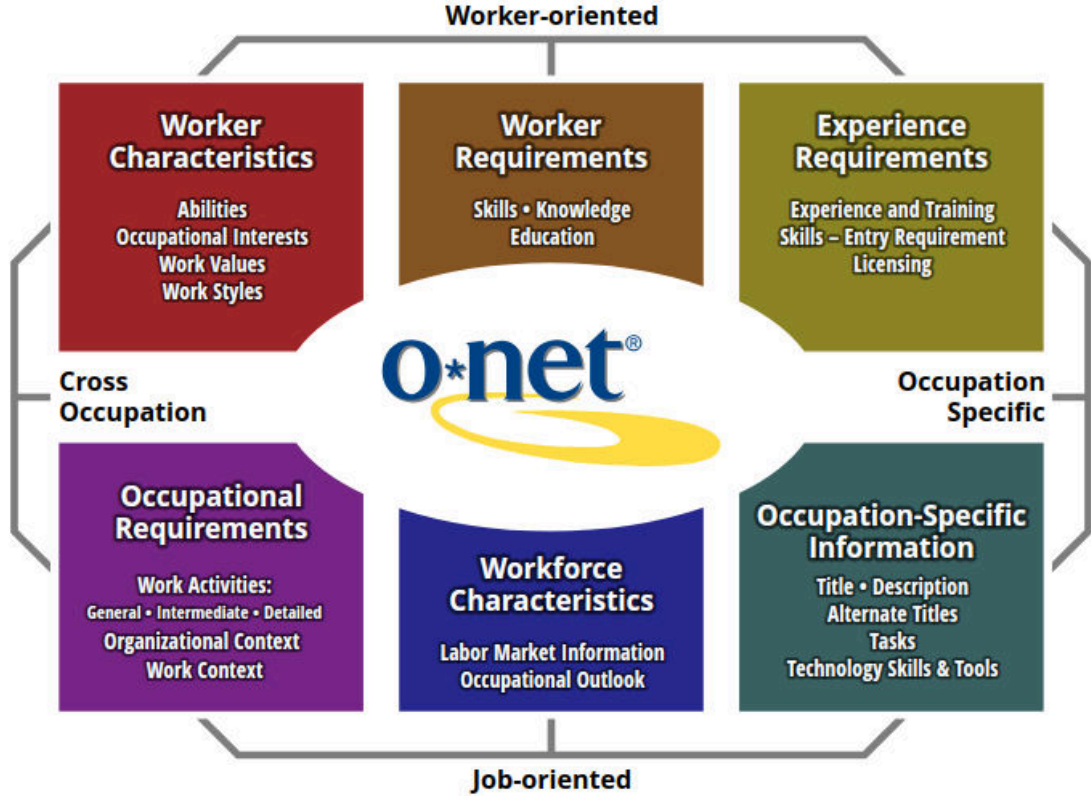


Figure 2.4: The O*NET® Content Model schematic representation [14].

The O*NET Content Model [14] serves as the structural foundation for the O*NET database, outlining critical occupational information by integrating job-oriented and worker-oriented descriptors. Developed through job analysis research, it includes six domains that detail attributes and characteristics relevant both across occupations (cross-occupational descriptors) and within specific jobs (occupation-specific descriptors), facilitating a comprehensive understanding of various roles and worker qualities.

According to Fareri *et al.*, when compared to ESCO, the O*NET taxonomy lacks in the level of detail, as the latter possesses six times more skills and three times as many job profiles [8]. Furthermore, there was a cooperation between the ESCO secretariat and the United States (U.S) Department of Labor to develop a crosswalk between the two frameworks [13] with the aim of “creating a bridge” that supports interoperability between two labour market standards using machine learning models and human validation.

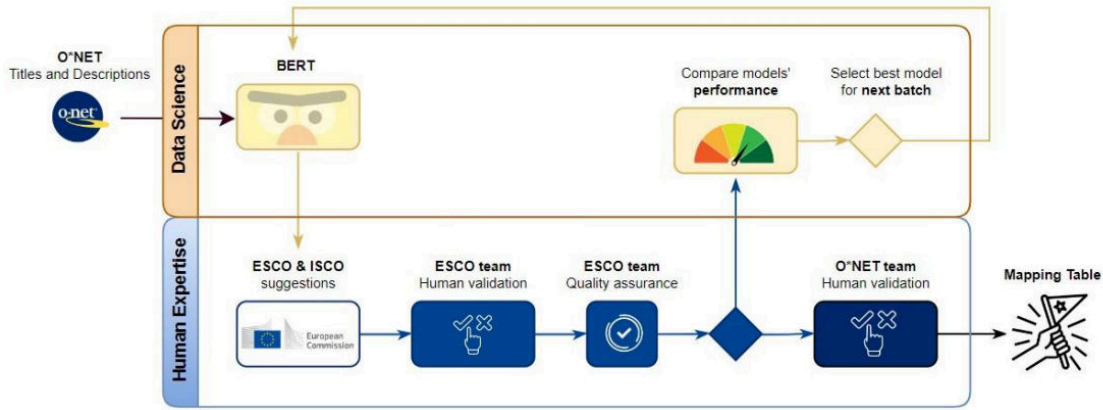


Figure 2.5: ESCO-O*NET crosswalk methodology [13].

The European Commission developed AI models (fine-tuned BERT base-uncased) to match O*NET occupations to ESCO occupations based on textual similarity [13]. These models were trained using expert feedback, national taxonomies, qualifications, and job vacancies. The best model suggested ten ESCO occupations for each O*NET occupation, representing the highest semantic similarity. Then, the ESCO team validated these suggestions, distributing the tasks among members and following specific rules to determine the relationship between O*NET-ESCO pairs. The validation process included quality checks and, if necessary, third-party involvement for disagreements.

This project involved two parallel efforts: iterative improvements by the data science team based on validation feedback and collaboration with the U.S. Department of Labor for further validation and refinement. The final crosswalk table, agreed upon by both teams, is now available on the ESCO Portal [1], [13].

2.2 LARGE LANGUAGE MODELS (LLMs)

Large Language Models (LLMs) are Neural Networks with billions of parameters and trained on vast quantities of unlabelled data (to understand language patterns, grammar and semantics) and labelled data (to guide the model towards more specific tasks) [15]. LLMs have significantly impacted the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI), representing a major transition in how machines understand and generate human language and how they return the information. The history of LLMs is rooted in the development of Machine Learning (ML) and NLP, with their emergence causing an extreme improvement in algorithm enhancing, computational power, data availability and data retrieval.

As previously mentioned, NLP has been fundamental to the development and operation of LLMs. Its approach usually involves the execution of a software pipeline with the aim of extracting information from text [6]. These approaches have been widely used in different fields, for example, in the analysis of customer reviews, social media usage, patents and, more recently, in the analysis of skill demand and job profiles [6], [8].

The field’s progression from syntactic analysis to understanding semantics and pragmatics of language has directly influenced the capabilities of these models. NLP techniques are crucial in pre-processing data for LLMs, handling tasks like tokenization, part-of-speech tagging, and entity recognition, which are vital for training these models effectively. Furthermore, advances in NLP have continuously pushed the boundaries of what is possible with LLMs, leading to more sophisticated and context-aware models.

Nowadays, LLMs are an integral part of a wide range of applications. From chatbots and virtual assistants to advanced text generation and translation, they have shown their ability to improve performance on text retrieval tasks through the generation of synthetic training data based on real examples [16], transforming the way machines interact with human language.

With the increasing use of LLMs, the concern in developing and optimizing prompts to achieve the best possible answers gave rise to a new discipline, Prompt Engineering [17]. Prompt engineering gathers a certain set of skills and techniques that enhance the interaction and development with LLMs, allowing for their augmentation with domain knowledge and external tools.

In the context of this work, we will discuss two important techniques of Prompt Engineering, **Zero-Shot Prompting** and **Role-Play Prompting**.

In NLP and LLMs, Zero-Shot Prompting is understood as providing a prompt to the model that is not part of the training data, but for which the model can provide a desirable result or answer [18]. This technique relies mainly on the model’s reasoning.

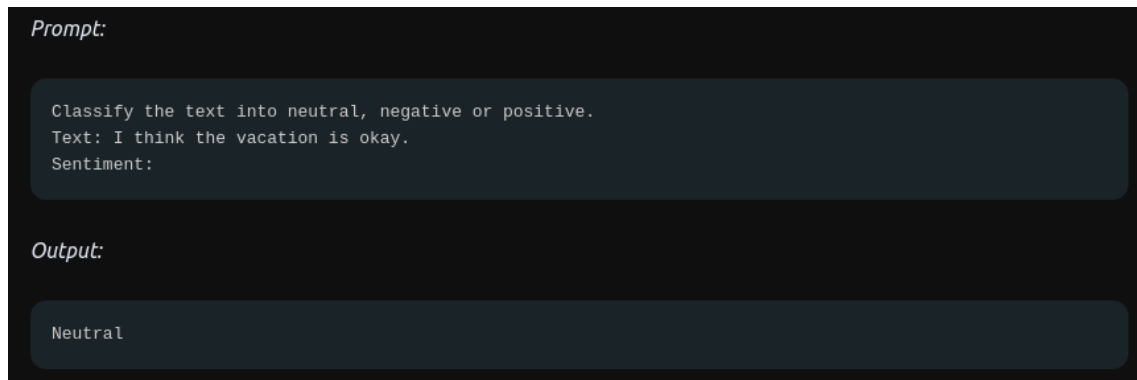


Figure 2.6: Example of a Zero-Shot prompt and answer [19].

In Figure 2.6 a simple application of **Zero-Shot Prompting** is represented, where no context or further information is given, just a simple task that tests the model’s reasoning. In the next section, an application of this technique, focused on skills matching, is explored.

Modern LLMs, such as *ChatGPT* and *Google Bard* show an extraordinary capacity for role-playing, being able to not only embody human characters, but also non-human entities, such as the Linux terminal [20]. That capacity can be referred to as **Role-Play Prompting**. The **Role-Play Prompting** allows the models to simulate complex human-like interactions and perform certain tasks differently, regarding the context prompted and the entity they are embodying. This represents an important feature to the development of the present work and will be further discussed in Chapter 3.

2.3 APPLYING NLP AND LLMs TO OCCUPATION AND SKILL TAXONOMIES

As previously discussed, NLP and LLMs offer sophisticated tools that can be used to interpret and process text from job descriptions and educational/training offers, helping in mapping their content to ESCO competences, for example.

In this section, various aspects and approaches related to the use of NLP and LLMs will be covered, in the specific context of skill extraction and subsequent alignment with occupational taxonomies, specifically focusing on the matching of these skills to the ESCO framework.

2.3.1 Skill-Span Extraction

The paper *SkillSpan: Hard and Soft Skill Extraction from English Job Postings* [21] presents an innovative approach to extract skills from job postings. In summary, it consists of a dataset of job postings annotated at the span-level¹ for hard and soft skills. Initially, the authors focused on manual annotations by domain experts, to accurately identify specific spans of text that correspond to skills in job postings. This manual annotation is crucial because it will be used to train and fine-tune various BERT models for automatic skill extraction and classification of the manually annotated skill spans. The effectiveness of these models is also explored, highlighting the importance of fine-grained, span-level annotations for a more accurate representation of skills in job postings.

While the final goal of this study is to provide a well-built approach on automated skill extraction, the initial step of dataset creation and annotation relies heavily on human expertise, which is a very lengthy and time-consuming process, especially if we take into account that ESCO contains 13,890 skills. In the following section, a method to fully automate this process is proposed.

2.3.2 Zero-Shot Matching

In *Large Language Models as Batteries-Included Zero-Shot ESCO Skills Matchers*, the authors explore the application of LLMs in skill matching tasks, particularly mapping job descriptions to ESCO competences. In this approach, a well-thought synthetic data generation is carried out [16]: for each of the 13,890 skills contained in ESCO, they prompted GPT-3.5 to generate 40 example sentences that could be used in a job posting in order to refer to the skill. The prompt used for this purpose is provided by the authors in the appendix of the paper.

Furthermore, Clavié and Soulié [16] propose a two-step process for their skills matching pipeline, composed of initially generating a list of potential matches for a given job description, followed by re-ranking these matches. To generate the list of potential matches, the authors use two distinct approaches. The first one relies on simple logistic regression classifiers and the second one is based on cosine similarity between the job description text and the sentences previously generated by GPT-3.5.

¹A span is a specific, continuous segment of text

In the second phase of the skills matching, LLMs are employed as zero-shot re-rankers (explained in section 2.2 [17]) i.e., without further context, the LLM is prompted to rank the list of potential matches according to their relevance or appropriateness for the job post.

This approach, requiring no human annotations, significantly outperforms previous methods, highlighting the transformative impact of LLMs in automating and refining the process of mapping complex job descriptions to specific skill sets within a recognized framework, such as ESCO.

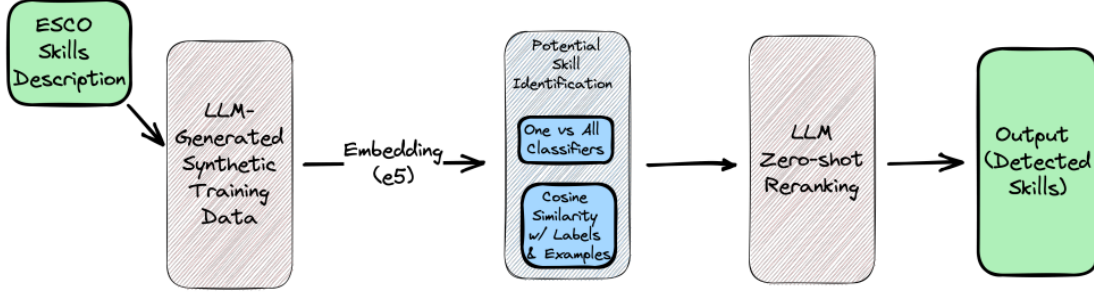


Figure 2.7: High-level overview of the full Zero-Shot process extracted from *Clavié and Soulié (2023)* [16].

2.3.3 Distant Supervision

In *Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning* [22] and *Design of Negative Sampling Strategies for Distantly Supervised Skill Extraction* [23], a distinct approach is taken. In these two papers, the authors rely on distant supervision for skill extraction, although with some differences between them. “Distant supervision is a technique of labelling data for relation extraction utilizing an existing knowledge database” [24], in this specific context, the database is ESCO.

In *Decorte et al. (2022)* [23], the authors use distant supervision to create a training set for skill extraction (Fig. 2.8 - 2). They collect sentences from job vacancies (Fig. 2.8 - 1) that explicitly mention skills of the ESCO taxonomy, a highly precise method that can, however, lead to false negatives. Then, negative sampling strategies are introduced to improve learning (Fig. 2.8 - 3). This step consists in “combining ‘positive sentences’ for a given skill (i.e., sentences labelled with that skill during the distant supervision step) with sentences not containing that skill (referred to as ‘negative sentences’)”. Finally, a pre-trained RoBERTa model is used to train binary classifiers for each skill (Fig. 2.8 - 4).

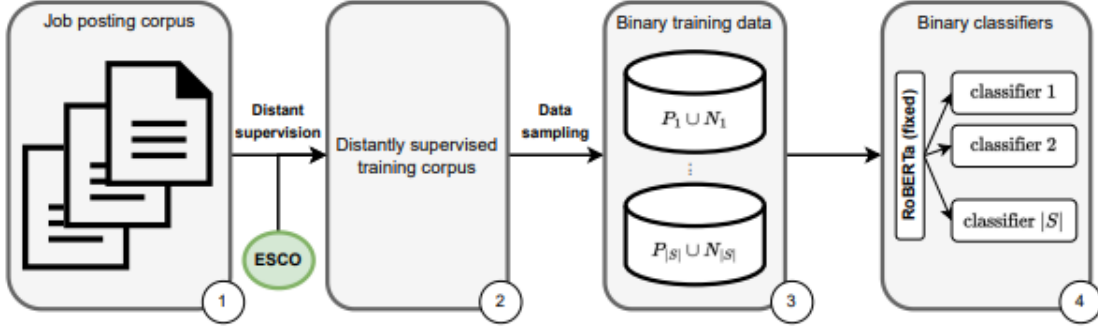


Figure 2.8: Schematic representation of Negative-Sampling processes, extracted from *Decorte et al. (2022)* [23].

In *Zhang et al. (2022)* [22], the authors also use distant supervision, in this case, with the ESCO API as reference. The authors annotate a Danish job posting dataset at a span-level, i.e., they identify and label spans within the dataset that correspond to particular skills (following the same logic as explained in the section 2.3.1), while using the ESCO API for distant supervision to obtain fine-grained labels. After that, various BERT models are fine-tuned on the distantly supervised labels and used to classify the spans. In addition, authors also presented experiments based on Zero-Shot and Few-Shot Prompting to address the problem of skill extraction.

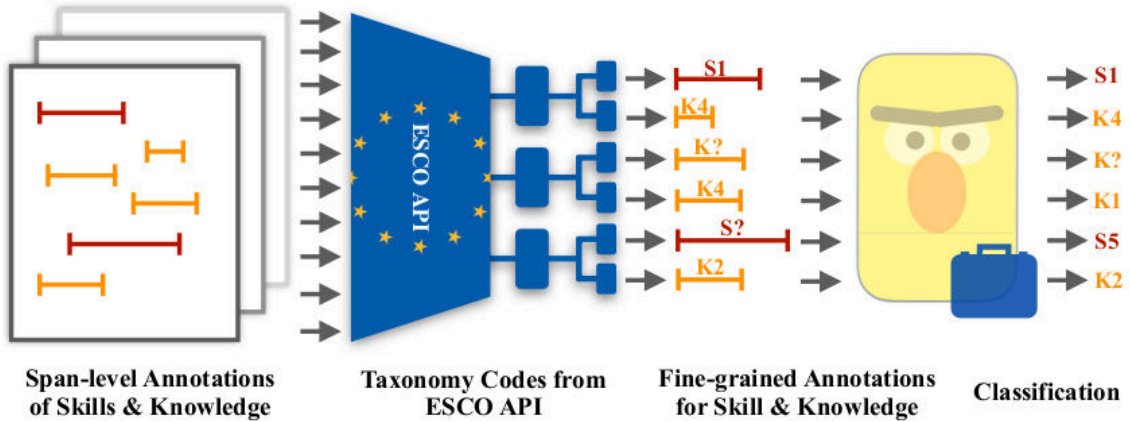


Figure 2.9: Pipeline for Fine-grained Danish Skill Classification in Kompetencer, extracted from *Zhang et al. (2022)* [22].

For this dissertation, the **Zero-Shot** approach and the **Distant Supervision** approach will be combined to obtain maximum advantage of both LLM and ESCO API. The data generation method used by *Clavié and Soulié (2023)* [16] could also be adapted for this work in order to process all of the ESCO skills, but generating educational offers instead of job postings. However, the other presented approaches will also be considered during the implementation of this work, as they provide interesting insights and solutions for different problems.

Methodology

In this chapter, aspects related to the work methodology will be discussed, such as initial approaches, aspects to consider when developing the system, and solutions to the problems raised on the literature review. Therefore, for a better understanding of the work outcome, we start by analysing a small diagram (Figure 3.1), that summarizes the interactions and relationships between the different parts involved in the system.

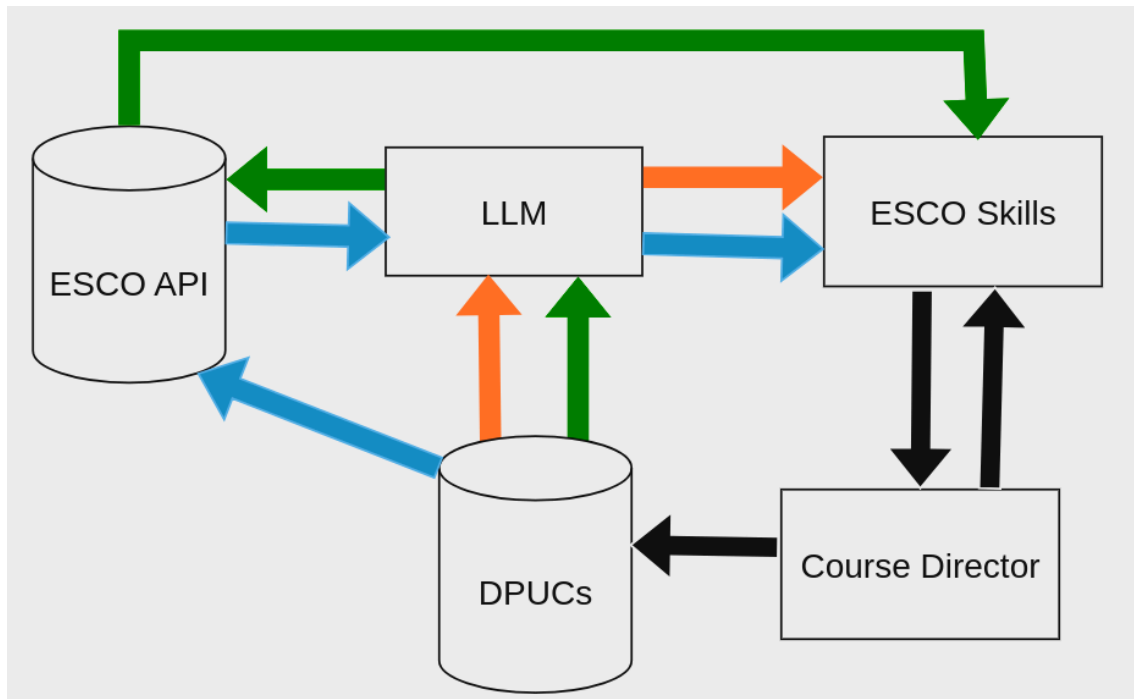


Figure 3.1: Diagram representing the system's pipeline, with different possible flows.

The five entities in the diagram consist of:

- ESCO API - Explained in section 2.1.1, it is crucial for the system to work properly, as it is the main source of information for skills matching.

- DPUC - Document containing the detailed information of a specific UA course or micro-credential, the structure of which was explained in Section 1.1. UA has a database that contains all DPUC from disciplines currently ministered.
- LLM - Will be responsible for assisting in the pre- and post-processing of data from the DPUCs and the ESCO API.
- ESCO skills - These are the final product of the system, which are wanted to be as closely related to the course or micro-credential in question as possible.
- Course Director - Professor responsible for managing the DPUC of the course for which they are responsible (human expert in the matter) and for evaluating the skills determined (i.e. approving or discarding the determined skills based on their similarity to the learning outcomes of the course).

Furthermore, there are three possible paths on Figure 3.1, each representing a different approach to the system flow:

The **orange** path is more related to a test rather than a real possibility for the system's flow. In this approach, the information coming from courses' DPUCs is provided directly to a LLM (a few tests were conducted using both *ChatGPT 4* [25] and *Google Bard* [26]) and it is then prompted to return a list containing the ESCO skills that accurately match the course's description. Since no LLM in the market was trained with ESCO data, the tendency is for these models to hallucinate and give general skills that rarely match the specific ESCO ones.

In the **blue** path, the information coming from the DPUCs is provided to the ESCO API in an HTTP query (the process of which will be further explained). After that, ESCO API returns a list of skills that should match the input of the query. This list is then provided to the LLM alongside with the already mentioned DPUC information for post-processing. The skills that do not appropriately match the DPUC are discarded, trusting in the model's ability for this serialization task. This process is repeated for each course and micro-credential.

In the **green** path, DPUC's information is firstly provided to the LLM for pre-processing to obtain a list of keywords that appropriately represent the description. This process is conducted so that only crucial information is provided to the ESCO API before the query (acting as a filter). Finally, the query is done to the API to obtain the list of skills.

A combination of the blue and green path can take place with a double passing through the LLM framework for pre- and post-processing, as it could significantly improve the results.

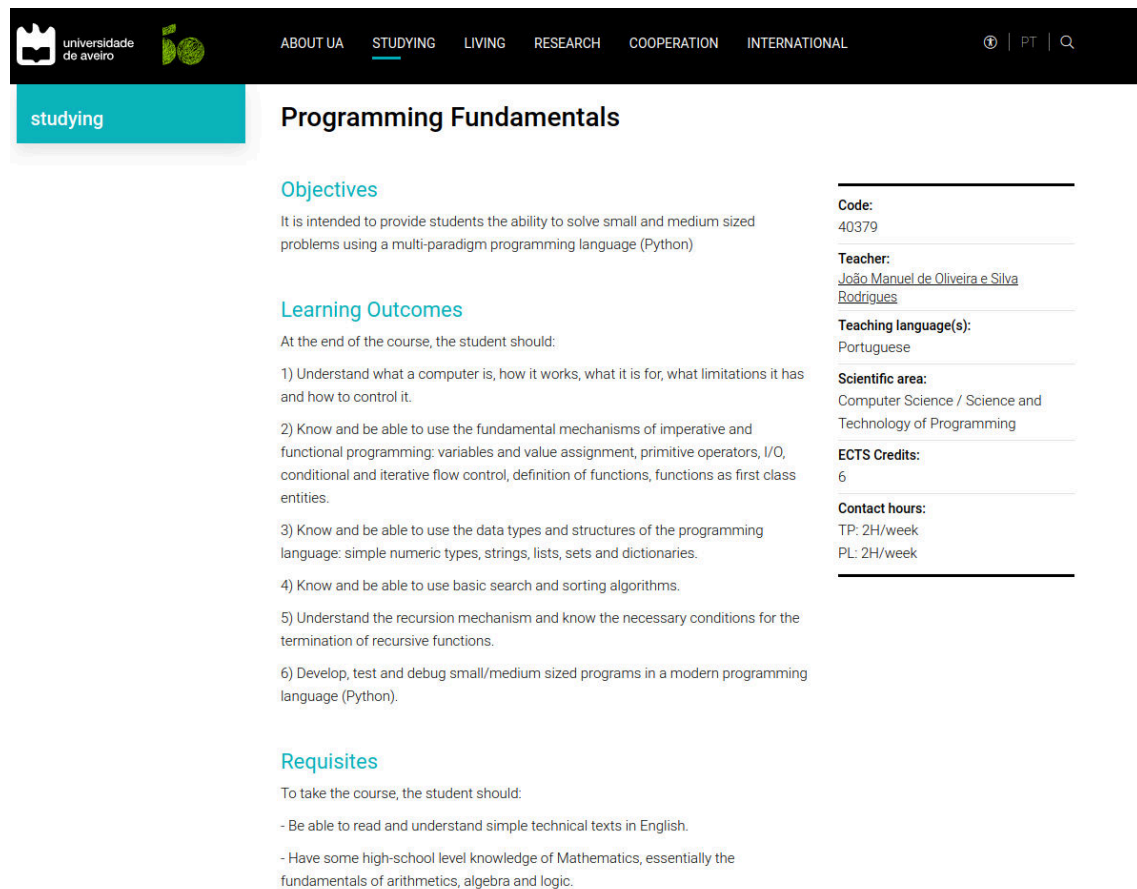
The final step will encompass the feedback of the Course Directors, after assessing the final list of skills for their lecturing course and discarding the skills that do not match the learning outcomes. The Course Directors are also responsible for managing the DPUCs of the courses they lecture. Both interactions are represented by the black arrows in Figure 3.1.

The possibility of using only the ESCO API to match an educational offer to ESCO skills has been set aside because after some tests, the API has shown incapable of returning an accurate set of skills for the provided information (i.e., some of the skills returned were non-sense and not related to the educational offer). An email was sent to the ESCO development team, delegated by the European Commission, in order to understand which algorithm or

logic was behind the fetching of these skills because it's not well detailed in the ESCO API documentation [27], but no further answer was given to this date.

As far as implementation is concerned, it is first necessary to get acquainted with the ESCO API. For that matter, a careful study and analysis of the API was carried out. Afterwards, a *Python* script was created to parse the information of a document that contains every UA DPUC. This document was filtered so that only the **course title**, **contents**, **objectives** and **learning outcomes** were extracted from all the DPUC fields (described in 1.1).

The University of Aveiro considers the microcredential courses as regular courses, therefore, the logic applied for extracting the desired fields from the microcredential description documents was the same.



universidade de aveiro ABOUT UA STUDYING LIVING RESEARCH COOPERATION INTERNATIONAL ⓘ | PT | Q

studying

Programming Fundamentals

Objectives

It is intended to provide students the ability to solve small and medium sized problems using a multi-paradigm programming language (Python)

Learning Outcomes

At the end of the course, the student should:

- 1) Understand what a computer is, how it works, what it is for, what limitations it has and how to control it.
- 2) Know and be able to use the fundamental mechanisms of imperative and functional programming: variables and value assignment, primitive operators, I/O, conditional and iterative flow control, definition of functions, functions as first class entities.
- 3) Know and be able to use the data types and structures of the programming language: simple numeric types, strings, lists, sets and dictionaries.
- 4) Know and be able to use basic search and sorting algorithms.
- 5) Understand the recursion mechanism and know the necessary conditions for the termination of recursive functions.
- 6) Develop, test and debug small/medium sized programs in a modern programming language (Python).

Requisites

To take the course, the student should:

- Be able to read and understand simple technical texts in English.
- Have some high-school level knowledge of Mathematics, essentially the fundamentals of arithmetics, algebra and logic.

Code:
40379

Teacher:
[João Manuel de Oliveira e Silva Rodrigues](#)

Teaching language(s):
Portuguese

Scientific area:
Computer Science / Science and Technology of Programming

ECTS Credits:
6

Contact hours:
TP: 2H/week
PL: 2H/week

Figure 3.2: Example of an UA's DPUC web page for the course of Programming Fundamentals [28].

After obtaining the DPUC data for input, it was necessary to obtain the list of skills. For that purpose, a previously developed script within our research group was used to query the ESCO API and obtain a list of skills for the provided text. The script uses the HTTP GET endpoint “/search” [26] and was adapted to fit the needs of this project.

Zhang *et al.* (2022) suggest the use of Levenshtein distance¹ to help finding the best match for a non-ESCO skill in the ESCO API [22]. This method relies on distant supervision (section 2.3.3) and can be successfully integrated in the system's flow to filter the API's results.

¹Distance given by the minimum number of operations needed to transform one string into the other

Using LLMs for this use case has demonstrated to be a consistent approach since the models can be applied for pre- and post-processing, in many ways (section 2.3). Our objective was to use a free and open-source LLM. After research and testing of open sourced LLMs, *Google Bard* [26] seemed like a perfect fit, not only because it was developed by Google, but also largely due to the fact that the open source community developed a *Python* package [29] that returns responses from *Google Bard* through cookie values, which allows for programmatically integration within the system proposed in this work.

Bard is a conversational generative artificial intelligence chatbot developed by Google, based initially on the LaMDA family of LLMs, later upgraded to PaLM and, more recently, to Gemini [29].

The capacity of *Bard* to perform text processing tasks and role-playing, embodying entities such as a Human Resources representative, a course director or a lifelong learner, allows a perfect fit in the system’s pipeline, as the prompt’s context can be modelled to achieve more filtered and objective answers.

Even though some advances and experiments have been conducted using *Bard* in this work, the possibility to explore other LLMs remains relevant, as they can greatly contribute to a more complete and accurate work.

Finally, when the list of skills is obtained for each UA course and micro-credential, they will be manually reviewed and assessed by the responsible course director to assure that every skill correctly matches the educational offer. This feedback will represent an extremely important step to evaluate and validate the system’s performance.

Figure 3.3 features a Gantt diagram that details the time estimates for each development and research task, as previously discussed. The development process of this dissertation will employ an iterative approach, consistent with the way its pipeline was described.

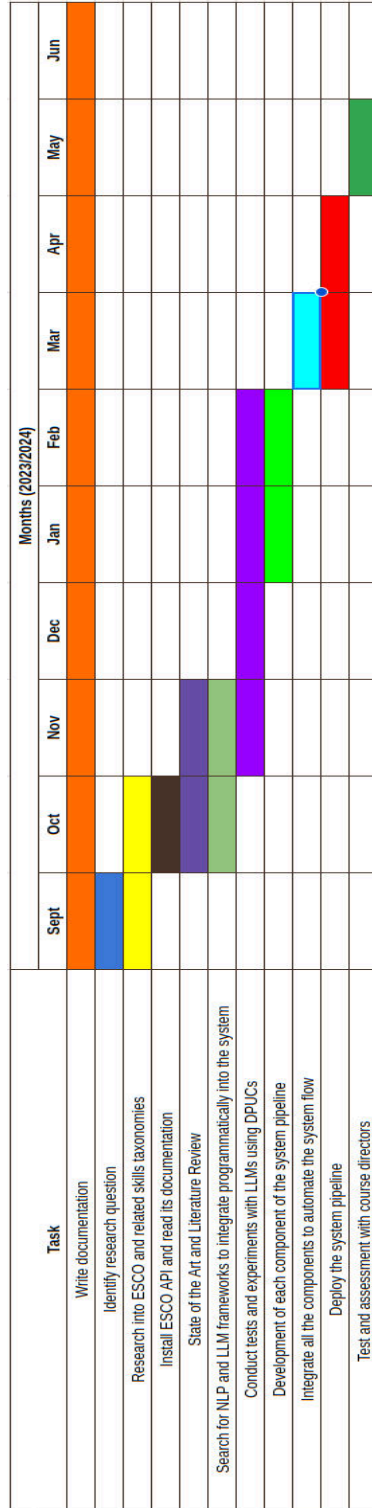


Figure 3.3: Gantt diagram containing the tasks that were and will be developed during this dissertation.

References

- [1] European Commission, *What is esco?*, Accessed: 15-11-2023. [Online]. Available: <https://esco.ec.europa.eu/en/about-esco/what-esco>.
- [2] International Labour Office - Hana Řihova, *Using labour market information*, Accessed: 27-11-2023. [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---ifp_skills/documents/publication/wcms_534314.pdf.
- [3] European Commission, *Why is esco needed?*, Accessed: 15-11-2023. [Online]. Available: <https://esco.ec.europa.eu/en/about-esco/what-esco/why-esco-needed>.
- [4] European Commission, *Esco api*, Accessed: 20-11-2023. [Online]. Available: <https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/esco-api>.
- [5] European Commission, *A european approach to micro-credentials*, Accessed: 06-01-2024. [Online]. Available: <https://education.ec.europa.eu/education-levels/higher-education/micro-credentials>.
- [6] F. Chiarello, G. Fantoni, T. Hogarth, V. Giordano, L. Baltina, and I. Spada, «Towards esco 4.0 – is the european classification of skills in line with industry 4.0? a text mining approach», *Technological Forecasting and Social Change*, vol. 173, p. 121 177, 2021, issn: 0040-1625. DOI: <https://doi.org/10.1016/j.techfore.2021.121177>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0040162521006107>.
- [7] World Economic Forum, *Building a common language for skills at work a global taxonomy*, Accessed: 10-12-2023. [Online]. Available: https://www3.weforum.org/docs/WEF_Skills_Taxonomy_2021.pdf.
- [8] S. Fareri, N. Melluso, F. Chiarello, and G. Fantoni, «Skillner: Mining and mapping soft skills from any text», *Expert Systems with Applications*, vol. 184, p. 115 544, Dec. 2021, issn: 0957-4174. DOI: 10.1016/j.eswa.2021.115544. [Online]. Available: <http://dx.doi.org/10.1016/j.eswa.2021.115544>.
- [9] European Commission, *Astronomer occupation page in esco*, Accessed: 08-01-2024. [Online]. Available: <http://data.europa.eu/esco/occupation/57a12047-4f1e-40ed-add5-9736923f231b>.
- [10] International Labour Office, *Isco - international standard classification of occupations*, Accessed: 02-12-2023. [Online]. Available: <https://www.ilo.org/public/english/bureau/stat/isco/>.
- [11] R-Project, *Esco - isco relationship*, Accessed: 05-12-2023. [Online]. Available: https://cran.r-project.org/web/packages/labourR/vignettes/occupations_retrieval.html.
- [12] European Commission, *International standard classification of occupations (isco)*, Accessed: 05-12-2023. [Online]. Available: <https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/international-standard-classification-occupations-isco>.
- [13] European Commission, *The crosswalk between esco and o*net*, Accessed: 23-11-2023. [Online]. Available: <https://esco.ec.europa.eu/system/files/2022-12/ONET%20ESCO%20Technical%20Report.pdf>.
- [14] U.S. Department of Labor, *The o*net® content model*, Accessed: 23-11-2023. [Online]. Available: <https://www.onetcenter.org/content.html>.
- [15] X. Lyu, S. Grafberger, S. Biegel, et al., *Improving retrieval-augmented large language models via data importance learning*, 2023. arXiv: 2307.03027 [cs.LG].
- [16] B. Clavié and G. Soulié, *Large language models as batteries-included zero-shot esco skills matchers*, 2023. arXiv: 2307.03539 [cs.CL].

- [17] DAIR.AI, *Prompt engineering guide*, Accessed: 21-12-2023. [Online]. Available: <https://www.promptingguide.ai/>.
- [18] Adrian Tam, *What are zero-shot prompting and few-shot prompting*, Accessed: 03-01-2024. [Online]. Available: <https://machinelearningmastery.com/what-are-zero-shot-prompting-and-few-shot-prompting/>.
- [19] DAIR.AI, *Zero-shot prompting*, Accessed: 03-01-2024. [Online]. Available: <https://www.promptingguide.ai/techniques/zeroshot>.
- [20] A. Kong, S. Zhao, H. Chen, *et al.*, *Better zero-shot reasoning with role-play prompting*, 2023. arXiv: 2308.07702 [cs.CL].
- [21] M. Zhang, K. N. Jensen, S. D. Sonniks, and B. Plank, *Skills span: Hard and soft skill extraction from english job postings*, 2022. arXiv: 2204.12811 [cs.CL].
- [22] M. Zhang, K. N. Jensen, and B. Plank, *Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning*, 2022. arXiv: 2205.01381 [cs.CL].
- [23] J.-J. Decorte, J. V. Haute, J. Deleu, C. Develder, and T. Demeester, *Design of negative sampling strategies for distantly supervised skill extraction*, 2022. arXiv: 2209.05987 [cs.CL].
- [24] P. Su, G. Li, C. Wu, and K. Vijay-Shanker, «Using distant supervision to augment manually annotated data for relation extraction», *PLOS ONE*, vol. 14, no. 7, pp. 1–17, Jul. 2019. DOI: 10.1371/journal.pone.0216913. [Online]. Available: <https://doi.org/10.1371/journal.pone.0216913>.
- [25] OpenAI, *Chatgpt*, Accessed: 08-01-2024. [Online]. Available: <https://chat.openai.com/>.
- [26] Google, *Google bard*, Accessed: 08-01-2024. [Online]. Available: <https://bard.google.com/>.
- [27] European Commission, *Esco api doc*, Accessed: 25-10-2023. [Online]. Available: https://ec.europa.eu/esco/api/doc/esco_api_doc.html.
- [28] João Manuel de Oliveira e Silva Rodrigues - Universidade de Aveiro, *Dpuc example - programming fundamentals*, Accessed: 06-01-2024. [Online]. Available: <https://www.ua.pt/en/uc/12286>.
- [29] Minwoo Park, *Google bard api*, Accessed: 08-01-2024. [Online]. Available: <https://github.com/dsdanielpark/Bard-API>.

APPENDIX A

Additional content