Kariyer.net Knowledge Base

Sabancı Üniversitesi FACULTY OF ENGINEERING AND NATURAL SCIENCES

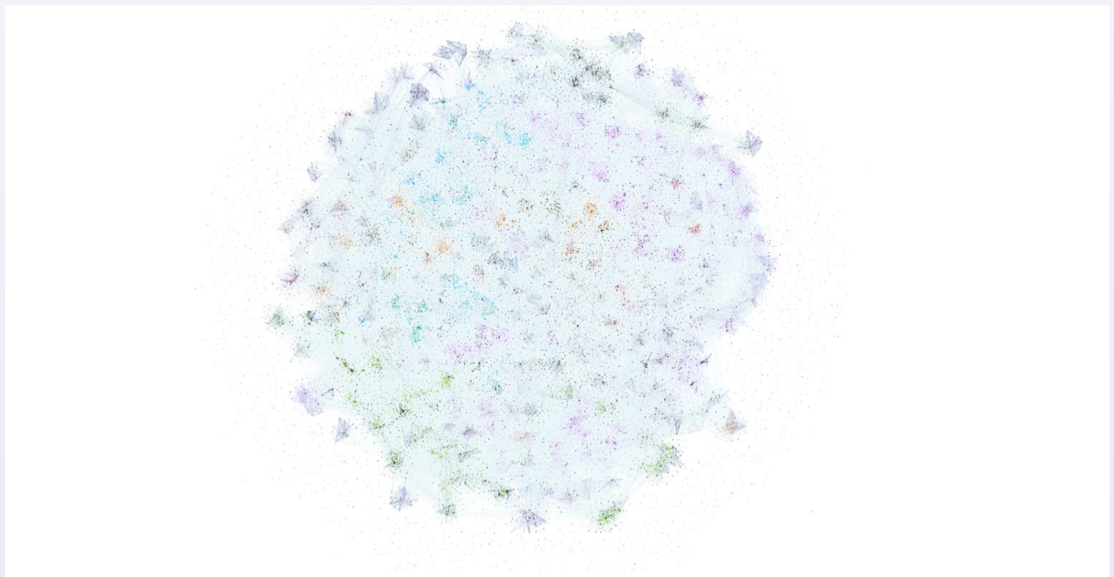
iLab

Student(s) Sima Adleyba

Faculty Member(s) Onur Varol

Company Advisor(s) Ömür Bali

ABSTRACT



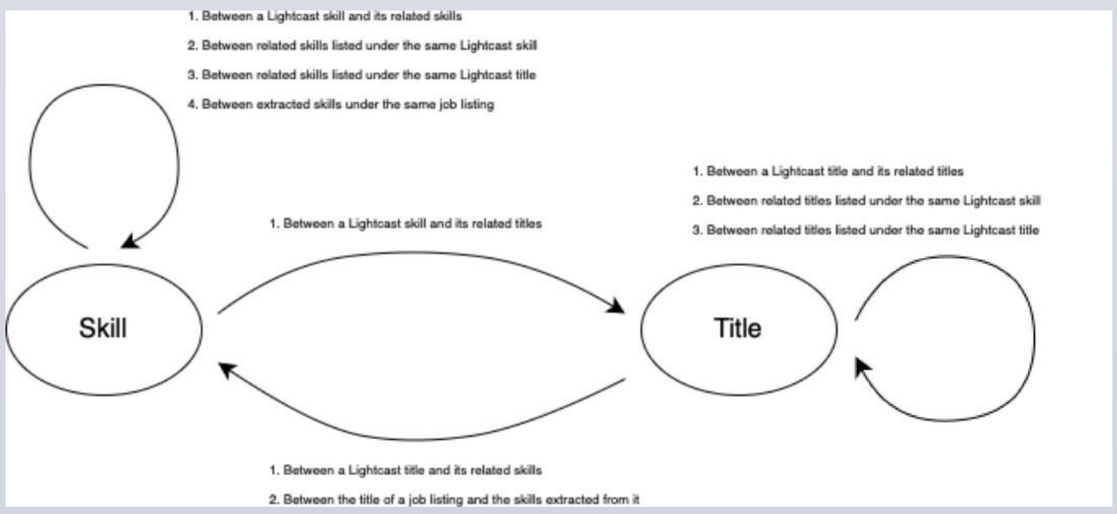
Effective job matching on job platforms is challenging, particularly in understanding the relationships between skills, job requirements and candidates. Kariyer.net, Türkiye's largest online job platform operating under iLab Holding, aims to address this challenge with the development of Kariyer.net Knowledge Base, which leverages data and state-of-the-art models to construct a graph where skills, job titles, job listings, and their relationships are systematically represented. The development process involved several major steps. The first step involved web scraping to gather the necessary skills and job titles data. Then, exploratory data analysis was conducted in order to understand the underlying characteristics of the gathered data. After data was examined, a skill extraction pipeline powered by a state-of-the-art large-language model was implemented. Extracted skills were embedded using a vector embedding model and clustering methods were performed to normalize the skills. In addition to skills, job titles were also normalized for consistency. Mentioned skills and job titles represent the nodes of the skills graph. The representation of their relationships through edges is being finalized by utilizing hyperparameter tuning for coefficients assigned for different edge sources. Once finalized, the constructed graph is expected to provide insights about the relationship between represented entities and these insights will be utilized in enhancing job recommendations, providing automatic skill suggestions for HR, and improving Kariyer.net's understanding of skill-career paths.

OBJECTIVES

The primary objective is to uncover meaningful patterns between skills and job titles in order to develop additional features for the platform, and enhance Kariyer.net's insight into evolving career trajectories. To do so, these were to main steps:

- Construction of a structure in the shape of a skills taxonomy or skills graph that reflects the job landscape in Turkey. To construct the structure, available Kariyer.net data and new data from online sources were utilized.
- Development of a skill extraction component that extracts the relevant information from the job listings effectively.
- Development of a pipeline that maps the extracted information from job listings to the skills structure accurately

PROJECT DETAILS



As the first step, a literature review was conducted to understand the methodologies employed by other job-searching platforms. After reviewing existing methods, Kariyer.net's unique context—shaped by differences between the Turkish and global job markets—along with recent advancements in large language models not utilized in previous work, was taken into account. Following this, the objectives stated above were determined.

PROJECT DETAILS (CONTINUED)

Next, the source of online skills data was determined: Lightcast. Then, a web-scraping script was implemented and categorized skills, job titles, and occupations data were collected. After data collection, data cleaning and exploratory data analysis were performed. With these initial steps done, a graph-based representation was deemed more suitable.

In this phase, several graphs—differing in how their edges are defined—were constructed using collected Lightcast skills. After examination, it was agreed to build new graphs using Kariyer.net data for comparison, which made it necessary to develop the skill extraction pipeline.

The skill extraction pipeline collects job listings from Kariyer.net's database, utilizes LangChain and ChatGPT 4o-mini to extract skills in a structured format, and adds them back to the database. To address inconsistencies—such as duplicate entries caused by collecting data from both Kariyer.net and Lightcast, or variations in how the same skill or title is written due to HR preferences—a normalization component was developed. Several experiments were conducted, and HDBSCAN clustering with vector embeddings was selected. After embedding and normalization, the component adds the relevant information back to the database.

Since all data required to construct the graph's nodes and edges had been collected, the next phase focused on defining the edges. Both Kariyer.net job listings and additional metadata associated with Lightcast-extracted skills were used to establish these connections.

Once edge definitions were finalized, a graph was constructed where edge weights reflected their frequency. The resulting graph had over 114,000 nodes—44,378 skills and 70,074 titles—and approximately 61 million edges. Analysis revealed the clustering step had not reached the desired level of accuracy, so the algorithm will be rerun with a different configuration, and hyperparameter tuning will be applied to weight edges from different sources before integrating the final graph into the automated pipeline.

CONCLUSIONS

The project had three main objectives: (1) to construct a skills taxonomy that reflects the job landscape in Turkey, (2) to develop a skill extraction component that identifies relevant information from job listings, and (3) to build a model that accurately maps the extracted information to the skills taxonomy. As the project progressed, first objective evolved into the construction of a heterogeneous graph, which is more suited to representing complex relationships between skills and job titles. The second objective was realized completely. The developed pipeline uses LangChain and ChatGPT 4o-mini to extract skills from job listings, and HDBSCAN clustering with embedding to normalize them. The third objective is partially realized, with graph construction underway using re-clustered entities and weighted edge definitions refined through hyperparameter tuning.

The project has a meaningful contribution to the previous state-of-the-art. Unlike earlier works that rely on structured data, the Kariyer.net Knowledge Base project handles unstructured and noisy job listings, which significantly increases the complexity of the task. This makes it more applicable for most of the real world applications. Additionally, utilizing a fully graph-based approach rather than a taxonomy makes Kariyer.net Knowledge Base a more flexible alternative for representing skills and job titles. The use of large language models such as ChatGPT 4o-mini and related frameworks such as LangChain further distinguished the project from prior approaches. Overall, the combination of the reasons mentioned above makes Kariyer.net Knowledge base a meaningful contribution.

REFERENCES

Lin, S., Yuan, Y., Jin, C., & Pan, Y. (2023). Skill graph construction from semantic understanding. In Companion Proceedings of the ACM Web Conference 2023 (pp. 978–982). Association for Computing Machinery. https://doi.org/10.1145/3543873.3587667

Macskassy, S. (2022, November 30). Building LinkedIn's Skills Graph to Power a Skills-First World. LinkedIn Engineering Blog. https://www.linkedin.com/blog/engineering/skills-graph/building-linkedin-s-skills-graph-to-power-a-skills-first-world

Macskassy, S., Jin, C., Lin, S., Wei, X., & O'Neill, M. (2023, March 21). Building and maintaining the skills taxonomy that powers LinkedIn's Skills Graph. LinkedIn Engineering Blog. https://www.linkedin.com/blog/engineering/data/building-maintaining-the-skills-taxonomy-that-powers-linkedins-skills-graph

Yan, J., Macskassy, S., Sun, L., Zhou, D., Kou, R., & Li, Z. (2023, December 13). Extracting skills from content to fuel the LinkedIn Skills Graph. LinkedIn Engineering Blog. https://www.linkedin.com/blog/engineering/skills-graph/extracting-skills-from-content