Data Science Job Market Data Mining: Identifying the Most In-Demand Skills for Data Scientist

Jui-Ching Yu

juiching@usc.edu

5507402044

Problem Statement

The job market, particularly in the tech industry, is evolving at an unprecedented pace as new technologies and AI models continue to advance. As a current Data Science student and aspiring job seeker, I often worry about whether I possess the most up-to-date skills to stay competitive and the possibility of being replaced by AI tools in an entry-level position. This project aims to address these concerns by answering the question: "What skills are most in demand for data-science-related roles in 2025, and how do they differ across positions and regions?" By automating the process of collecting and analyzing live job postings, this project seeks to generate actionable insights that can help students identify which tools and technologies to learn, enable educators to align curricula with industry needs, and support professionals in planning effective upskilling strategies.

Data Sources

Data will be collected from LinkedIn, Glassdoor, Handshake, and Indeed, focusing on roles such as Data Scientist, Machine Learning Engineer, Data Engineer, and Data Analyst. Using Python's requests and BeautifulSoup (or SerpAPI for safer API access), around 1,000 job postings will be gathered. Key fields include job title, company, location, posting date, salary (if listed), and job description. All data will be publicly available and collected in accordance with fair-use guidelines, then stored as CSV files for analysis.

Analysis and Visualizations

The project will analyze the collected job postings to uncover the most in-demand skills and relationships between them. A frequency and TF-IDF analysis will identify commonly mentioned and distinctive technical skills. Clustering methods will group jobs by dominant skill sets. Optionally, the project will explore correlations between salary and the number or type of required skills.

For visualization, the results will be presented through bar charts (top skills), a word cloud (key terms overview), network graphs (skill co-occurrence), and scatter plots (salary vs. experience or skill count). Tools such as matplotlib, plotly, wordcloud, networkx, and scikit-learn will be used for analysis and visualization.