WEB - SCRAPPING

Scraping for Key Skills Analysis

Project, DIA - 1

Submitted by:-

Manisha RAWLA
Guillaume GALLET
Audrey DREYS
Shubham SAINI

INTRODUCTION

This project focuses on utilizing web scraping techniques to extract valuable data from job-hunting websites, specifically Indeed and LinkedIn. The primary objective is to analyze and identify the skills in demand within various industries. The project not only serves pedagogical purposes by applying classroom concepts but also aligns with professional and personal development, encompassing data analysis, processing, web application development, and data visualization. Additionally, we emphasize the aspect of Responsible AI (RSE) by ensuring compliance with website data policies.

In the fast-evolving job market, understanding the skills in demand is crucial for both job seekers and employers. This project aims to provide insights into the current industry requirements by scraping job postings from popular platforms like Indeed and LinkedIn.

OBJECTIVES:-

2.1 Pedagogical:

Apply and practice web scraping techniques learned in class:

The primary pedagogical objective is to put into practice the theoretical knowledge gained during class sessions focused on web scraping. This involves understanding the intricacies of web pages, identifying the structure of HTML elements, and employing appropriate libraries (such as BeautifulSoup in Python) to extract relevant data. Practical implementation will reinforce the understanding of concepts learned in a classroom setting.

Learn new practices and techniques for handling the complexity of various websites:

Web scraping often presents challenges due to the diversity in website structures. Our aim is to encounter and overcome these challenges by learning new practices and techniques. This may involve adapting scraping strategies based on different HTML structures, handling dynamic content using tools like Selenium, and addressing potential roadblocks posed by various websites.

2.2 Professional:

Identify problems related to skill demand in various industries:

The professional objective is to identify and address real-world problems related to skill demand in different industries. By analyzing job postings, we aim to understand the specific skills that employers are seeking, facilitating informed decisions for job seekers and employers alike.

Formulate working hypotheses to address the identified problems:

We will formulate working hypotheses to address the identified problems upon identifying patterns and trends in the scraped data. These hypotheses will guide further analysis and serve as a foundation for developing strategies to meet the job market demands.

Develop a pedagogical approach to solving these problems:

In developing a pedagogical approach, we aim to draw on academic principles and theories to formulate systematic solutions. This involves applying classroom teachings to real-world scenarios and creating a bridge between theoretical knowledge and practical problem-solving.

2.3 Personal:

Utilize data analysis and processing skills to extract relevant information:

The personal objective centers around applying data analysis and processing skills to extract pertinent information from the scraped data. This involves cleaning and transforming raw data, utilizing statistical methods for analysis, and ensuring the accuracy and reliability of the information extracted.

Create data visualizations to communicate findings effectively:

Communication of findings is crucial, and creating effective data visualizations is key to achieving this. By employing visualization tools and techniques, we aim to present complex insights clearly and understandably, enhancing the overall impact of our analysis.

2.4 Responsible AI (RSE):

Adhere to website data policies to ensure ethical data scraping practices:

Responsible AI (RSE) is an integral aspect of the project. We commit to strictly adhering to the data policies of the websites being scraped. This includes obtaining proper permissions, respecting terms of use, and ensuring that our scraping practices align with ethical standards, fostering a responsible and respectful approach to data acquisition.

Emphasize responsible handling of collected data throughout the project:

Throughout the project lifecycle, we will prioritize the responsible handling of collected data. This involves implementing secure data storage practices, respecting user privacy, and anonymizing sensitive information. The goal is to ensure that the project aligns with responsible AI principles, fostering trust and ethical conduct in data-driven endeavors.

Methodology

3.1 Data Collection:

3.1.1 Web Scraping Techniques:

Utilizing Python-based libraries such as BeautifulSoup and Selenium, we implemented web scraping techniques to extract job postings from two major job-hunting websites, Indeed and LinkedIn. The scraping process involved navigating through the websites, parsing HTML structures, and collecting relevant information such as job titles, skills required, company names, and job descriptions.

3.1.2 Cross-Referencing Data:

To ensure the reliability and accuracy of the collected data, we implemented a cross-referencing mechanism. This involved comparing and validating job postings from Indeed against those from LinkedIn, minimizing potential discrepancies and enhancing the robustness of our dataset. Additionally, we considered the data policies of both websites to ensure ethical and responsible data scraping practices.

3.2 Data Analysis:

3.2.1 Identification of Key Skills:

After the successful extraction of job postings, we conducted a comprehensive analysis to identify the key skills in demand across different industries. Natural Language Processing (NLP) techniques were applied to analyze job descriptions and extract relevant keywords and skills. This process allowed us to compile a list of skills that are frequently mentioned in job postings within each industry.

3.2.2 Trend and Pattern Analysis:

We delved into trend and pattern analysis to uncover insights within the job postings. This involved examining the frequency of specific skills, identifying emerging trends, and understanding the evolving demands of the job market. Statistical tools and visualization aids were employed to present our findings in a clear and interpretable manner.

3.3 Data Visualization:

3.3.1 Visual Representation of Insights:

To effectively communicate our findings, we employed various data visualization techniques. Graphs, charts, and interactive visualizations were created to represent the distribution of skills, trends in demand, and other relevant patterns. Tools like Matplotlib, Seaborn, and D3.js were utilized to enhance the visual appeal and interpretability of the presented data.

3.3.2 User-Interactive Visualizations:

We prioritized user interaction by incorporating features such as hover-over tooltips, filtering options, and dynamic charts. These elements empower users to explore the data, gain deeper insights, and customize their viewing experience based on their specific interests.

PRACTICAL APPLICATIONS & IMPLICATIONS:

4.1.1 Heading: Aligning Education with Industry Demand

In today's rapidly evolving job market, the significance of aligning education with industry demand cannot be overstated. Our project's findings offer a practical solution for both educational institutions and private training providers to bridge the gap between academic curricula and the skills demanded by employers.

4.1.2 Content:

Our analysis of skills in demand across various industries serves as a valuable resource for educational institutions. By leveraging this data, schools and universities can tailor their course offerings to ensure that students are equipped with the skills most sought after in the job market. This alignment not only enhances the employability of graduates but also strengthens the connection between academia and industry.

4.1.3 Implications:

Curriculum Enhancement: Educational institutions can use our findings to enhance their curricula, incorporating relevant skills and technologies that are currently in high demand.

Adaptive Learning Paths: By regularly updating their course content based on our insights, schools can provide students with adaptive learning paths, preparing them for the ever-changing landscape of the job market.

4.2 Empowering Professional Development:

4.2.1 Heading: Tailoring Professional Development Programs

Our project extends its impact to the professional development sector, offering insights that can guide the design of short courses and training programs provided by private institutions.

4.2.2 Content:

Private training providers can leverage our findings to design targeted short courses and workshops aimed at addressing specific skill gaps identified in our analysis. This approach ensures that professionals can continually upskill or reskill themselves in alignment with industry requirements, enhancing their competitiveness in the job market.

4.2.3 Implications:

Customized Training Modules: Private institutions can create customized training modules based on the identified skills, providing professionals with focused and relevant learning experiences.
Market-Relevant Certification: Offering certifications aligned with the skills in demand enhances the market relevance and value of professional development programs.

4.3 Enhancing Career Guidance:

4.3.1 Heading: Informed Career Decision-Making

Our project's insights have implications for career guidance services, enabling individuals to make more informed decisions about their educational and career paths.

4.3.2 Content:

Career guidance counselors in schools and universities can use our findings to provide students with up-to-date information on the skills required for their desired career paths. This empowers students to make informed decisions about their education and career choices, aligning their aspirations with market realities.

4.3.3 Implications:

Tailored Career Path Recommendations: Career counselors can tailor their advice based on real-time data, ensuring that students are well-prepared for the skills landscape they will encounter upon entering the workforce.

Job Market Awareness: Students gain a heightened awareness of the dynamic nature of the job market, allowing them to proactively adapt their skill sets for future success.

GIT-HUB, code commit link

shubhamsaini20/WebScrapping-DataProcessing (github.com)

RESULT (Summarize in presentation)

"To provide a concise overview of our findings and insights, we have prepared a presentation summarizing the key results. To make it easily accessible, we have uploaded the presentation to the following link: https://www.canva.com/design/DAF5bNGWGr8/L1hl9y-d8zuzMvzebQP7hA/edit?utm_content=DAF5bNGWGr8&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton
Please feel free to explore the slides for a visual representation of our analysis and the skills in demand across various industries."

Link for presentation (Result summarize)-
https://www.canva.com/design/DAF5bNGWGr8/L1hl9y-d8zuzMvzebQP7hA/edit?utm_content=DAF5bNGWGr8&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

References:

A. Data Policies:

Indeed Privacy Policy: https://www.indeed.com/legal/privacyfaq

LinkedIn Privacy Policy: https://www.linkedin.com/legal/privacy-policy

B. Relevant Literature:

- *"Web Scraping with Python: Building Your Own Tools to Extract Data"* by **Ryan Mitchell**: https://www.oreilly.com/library/view/web-scraping-with/9781491985564/
- "*Text Mining in Practice: A Case Study Approach"* by **Bodnar and Gupta**: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119282105
- *"The Future of Jobs Report 2020"* by **World Economic Forum**: https://www.weforum.org/publications/the-future-of-jobs-report-2020/

C. Additional Resources:

Beautiful Soup Documentation: https://readthedocs.org/projects/beautiful-soup-4/

Selenium Documentation: https://www.selenium.dev/documentation/

D. School Portal

Devinci Learning