Project Report - Data Collection Lab 094290

Amit Zalle - email
Raz Biton - raz.biton@campus.technion.ac.il
Shalev Hermon - shalevhermon@campus.technion.ac.il

Project Introduction

This project tackles the challenge of limited job discovery using AI and comprehensive web scraping to recommend the most relevant open positions for a user. The necessity arises from the vast amount of job postings available online, making it difficult for individuals to identify the most suitable opportunities. Current job search platforms often rely on keyword matching, which can miss out on positions with a slightly different wording but a perfect fit for the user's skills and experience. This project addresses this gap by employing a multi-faceted data-driven approach:

- Web Scraping for Open Positions and Company Details: We leverage web scraping techniques to gather data on a vast pool of open positions from various online sources. This scraped data includes details like job titles, types, seniority levels, locations, descriptions, links, and crucially, information about the companies offering those positions. This additional layer of data allows for a more informed job search experience.

- Identifying similar professionals: We utilize the BM25 similarity algorithm to analyze the user's LinkedIn profile and find individuals with the most comparable career paths. .

- Matching skills to open positions: By training a word2vec model on the past job descriptions of these similar professionals and the descriptions of scraped open positions, we uncover hidden relationships and semantic similarities. This allows us to recommend open positions with titles that might not be an exact keyword match but are highly relevant to the user's background.

The outcome is a personalized list of recommended open positions that goes beyond keyword matching. By incorporating comprehensive web scraping, we expand the search scope and unlock a wider range of relevant positions with valuable company details. This combined approach offers a more comprehensive and relevant job search experience for the user.

Data Collection and Integration

Data from original data-sets

The 'Job Advisor' utilizes 'profiles' and 'companies' datasets to match users with suitable job positions, analyzing profile similarities and enriching recommendations with company data to optimize the job-seeking process.

Additional Data Collected

Additional data collected included job postings, user data, and information about companies mentioned in the postings. Job listings were scraped from 'LinkedIn', 'Indeed', and 'ZipRecruiter' all considered top 3 job search websites according to GeeksforGeeks.

Tools and Methods:

- Bright Data Web Browser: Integrated with Playwright, connects to job search pages, performs search queries, and retrieves job listings, companies, and user data via the Bright Data proxy network.
- Playwright: Utilized for navigating job search and company pages, interacting with job listings, extracting job details, and managing page navigation and timeouts.
- Beautiful Soup: Used to parse HTML content from pages retrieved by Playwright, extracting specific elements like job titles, company names, and locations.
- Tkinter UI: manual input, ensuring user profile data completeness.
- Additional Methods: Overcoming challenges in accessing LinkedIn data involved employing alternative methods, including the use of guest URLs and persistent attempts. Additionally, navigating LinkedIn's infinite scrolling of jobs search page necessitated the utilization of alternative methods, such as extracting URLs from Chrome DevTools without scripts and CSS.

Overall, the project successfully gathered job postings, user data, and company information from LinkedIn, Indeed, and ZipRecruiter. Utilizing tools like Bright Data Web Browser, Playwright, and Beautiful Soup, we overcame challenges with workaround network solutions and ensured data completeness through manual UI input.

Additional Data Integration

The additional data enhances our solution by integrating it into two models. Model 1 identifies the most similar profiles to the user using collected profile data, while Model 2 utilizes open job listings and company data to recommend suitable job positions for the identified profiles.

Enrichment

- Open job positions items: consist of job listings gathered from multiple platforms, each including details such as title, company name, company URL, job location, seniority level, employment type, job function, industries, about section, job URL, and posting date. The initial collection size is denoted as a(Open Positions)=0.1, with an actual enrichment size of 7182 for models evaluation. For each new user the 'Job Advisor' can collect new updated job listings.

- Companies items: consist of companies not present in the original dataset, featuring the same attributes. The initial collection size is denoted as a(Companies)=0.2, resulting in an actual enrichment size of 2530. The collection process depends on identifying missing data. For each new user collected job listings, the 'Job Advisor' can collect the additional companies data if needed.

- User data item: encompasses all available user information. The initial estimation of the dataset size is indicated as a(people) = 2, yet the actual dataset size is 1. It has been determined that further collection is unnecessary to achieve our objectives.

Data Analysis

BM25 - profile to profiles similarity

Data Assumptions and Filtering- we firstly assumed that the data about the profiles comes from the same distribution, and therefore we can learn about a new profile from the others. Additionally to lower the running time and to enhance the model result we filtered profiles that didn't had a past job or any education. Additionally we entered the new profile to the profiles table so we can preprocess both at once, and also use the new profile for the prediction.

Analysis Techniques- In our first model for profile-profiles similarity we used the information retrieval model BM25. This model is an improved normalized model of the tf-idf representation model. To use those model, we consider the profile we want to find similarity for as query, and the profiles in the data as documents, and sort the "documents" according how fit they are to our query using normalized counting of the query terms in the documents. To do so, we need to turn the profiles into list of words they include. And we did so in the preprocessing using UDF function on the pyspark data frame. After that we represent the different profiles using tf-idf vectors that fit each document to vector of the terms it contains and the number of appearances they have.

Feature Selection- we apply 3 different information retrieval for each aspect of the profile qualification and general description. The first segment is education, which we build from past and current companies' names and positions. The second segment is education, which we extract from the detailed about the

person past degrees, and the extra courses he's taken. And the last segment is the general information about the individual, which we extracted from column's like the "about" that is the person own description about himself. The combination of those three shown much success and yield it own aspect of similarity between people both in regard to professional information and general one.

Word2Vec - Profile to Job Similarity

Data Assumptions- we assume that in this part we get the top profiles that is similar to the user, that we get from the first model.

Analysis Techniques and Feature selection- Leveraging word2vec, we analyze job descriptions from similar user profiles (experience on LinkedIn). These descriptions naturally capture relevant skills. We clean the data (remove stop words) to improve analysis. Word2vec then transforms these descriptions and scraped job titles into numerical vectors for comparison. This allows us to identify relevant jobs beyond keyword matching, based on semantic similarities Our feature selection process leverages a combination of domain knowledge and the statistical analysis capabilities of word2vec:

• Domain Knowledge: We primarily focus on the past job descriptions within the experience sections of the similar users' profiles. This approach aligns with the assumption that jobs held by users with similar profiles are likely to be relevant recommendations for the target user as well. Skills and experience are often reflected in job descriptions, making them a rich source of information for identifying potential matches.

• Statistical Analysis (Word2vec): We convert both cleaned job descriptions and scraped job titles into numerical vectors using word2vec. Comparing these vectors allows us to calculate a similarity score, reflecting the semantic alignment between the user's past experience and job requirements. This goes beyond keyword matching.

By combining domain knowledge with word2vec's capabilities, we identify relevant features representing the user's skills and experience. This enables generating targeted job recommendations that align with the user's background.

AI Methodologies

BM25 - profile to profiles similarity

To sort the profiles based on their similarity to the new profile, we assumed a information retrieval world, where we need to sort the profiles as a documents similarity to the profile's query. we used the tf-idf we calculated before to create

BM25 score for each segment words using the formula:

$$BM25(D,q) = \sum_{i=1}^{n} \left[\log\left(\frac{N - df(q_i) + 0.5}{df(q_i) + 0.5} \right) \cdot \frac{tf(D,q_i) \cdot (k_1 + 1)}{tf(D,q_i) + k_1 \cdot \left(1 - b + b \cdot \frac{length(D)}{\frac{1}{N} \sum_{j=1}^{N} length(D_j)} \right)} \right]$$

When N-number of documents, $q_i - termiinq, k_1 = 2, b = 0.75 \quad hyperparameters.$

For each segment calculate the scores for the new profile, and then normalized them and sum them up in weighted average according to the preferred weights of the user. We sort based on the result.

Word2Vec - profile to job similarity

We use Word2vec for Semantic Analysis by Harnessing Word Embeddings. Our primary analysis technique hinges on word2vec, NLP tool that falls under the umbrella of machine learning algorithms. Word2vec excels at creating word embeddings, numerical representations of words that capture their semantic meaning. Words with similar meanings occupy similar positions in this vector space. This allows us to analyze the textual descriptions from two key data sources:

• Past Job Descriptions of Similar Professionals: We extract job descriptions from the "experience" section of LinkedIn profiles identified as similar to the user's profile using the BM25 similarity algorithm. These descriptions hold valuable insights into the user's skills and experience.

• Scraped Open Position Titles: The web scraping component gathers titles of open positions from various online sources. These titles summarize the core requirements and responsibilities of the positions.

By leveraging word2vec, we can move beyond simple keyword matching between job descriptions and titles. Word2vec's ability to capture semantic relationships allows us to identify potential matches even if the descriptions use different terminology but convey similar skills and requirements. In essence, word2vec plays a critical role in our analysis by transforming textual descriptions into numerical vectors that capture the underlying meaning and relationships between words. This enables us to calculate similarity scores and identify relevant job recommendations for the user.

Evaluation and Results

text

Limitations and Reflection

Although we found ways to scrape job listings effectively from search result pages, we still faced challenges. Scraping websites without direct APIs caused delays in data collection. While we managed to address some issues, like dealing with login requirements on platforms like LinkedIn, Bright Data's browser limitations stopped us from accessing all the user data we needed. Specifically, Bright Data blocked the login option to LinkedIn, leading to incomplete user profiles for analysis. This gap between our plans and technical constraints necessitated manual collection of user data through a user interface (UI).

When using the profiles' static table, we filter the data to include only informative profiles. This loses data, and might be harmful if the new profile is also uneducated for example. This filtering can be removed given a larger data, faster gpu for running the algorithm, or more information.

In the W2V model, we came across inconsistent Job Descriptions- Scraped job titles might contain inconsistencies like abbreviations or unclear wording. This inconsistency can confuse word2vec, hindering its ability to accurately capture the skills and requirements of the position. additionaly we had limited Information in Job Titles- Job titles often provide a brief summary of the position. If the information is limited, word2vec might miss out on key skills or experience mentioned elsewhere in the job description.

Furthermore, we can decide on a better hyperparameters for the models, if we had more time and power to run more models.
text

Conclusions

Conclusions

• Identifying Similar Users: We employed the BM25 similarity algorithm to identify LinkedIn profiles with experiences most similar to the target user. This initial filtering step ensured we focused on job descriptions with a high likelihood of reflecting relevant skills.

• Identifying Relevant Skills: We developed a methodology that utilizes word2vec to analyze past job descriptions from these similar users. This allowed us to create a profile of the user's skills and experience based on the semantic meaning conveyed within the descriptions.

• Matching Skills to Opportunities: By comparing the user's skill profile, generated through word2vec analysis, with the semantic information extracted from scraped open position titles, we were able to calculate similarity scores. These scores reflect the degree of alignment between the user's background and the requirements of the positions.

• Data-Driven Recommendations: Our approach leverages the power of word2vec to generate more nuanced and relevant job recommendations for the user. This can empower individuals in their career exploration by highlighting opportunities that align with their skillsets.

In conclusion, this project has taken a significant step towards utilizing word2vec technology, in conjunction with BM25 for user identification, to bridge the gap between skills and opportunities. The ability to identify relevant skills from past experiences and match them to suitable job openings can be a valuable tool for both job seekers and employers.
text

APPENDIX

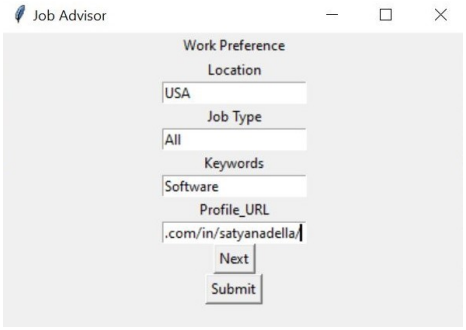

Figure 1: After completing his profile data manually, the user enters his basic job preferences.

current_company:name	education	educations_details	experience	followers	following	groups	id	languages	name
Microsoft	{'degree': 'Computer Science', 'end_year': '',...	The University of Chicago Booth School of Busi...	{'Company': 'Microsoft', 'Company ID': 'micros...	11M followers	500+ connections	NaN	NaN	{'Title': 'English', 'Subtitle': ''}	Satya Nadella

Figure 2: Some user data collected through scraping and manual input.

	Title	company_Name	company_URL	Job_Location	Seniority_level	Employment_type	Job_Function	Industries	About	Job_URL	Posted	collected
0	Software Engineer, Platform	Tecton	https://www.linkedin.com/company/tectonai?trk=...	New York, NY	Not Applicable	Full-time	Engineering and Information Technology	Software Development	At Tecton, we solve the complex data problem i...	https://www.linkedin.com/jobs/view/software-en...	9 hours ago	2024-04-10 17:07:01.748296
1	Software Engineer	Phoenix Home Care and Hospice	https://www.linkedin.com/company/phoenixhomehc...	Longmont, CO	Entry level	Full-time	Engineering	Technology, Information and Internet	About SparkFunSince 2003, SparkFun has been he...	https://www.linkedin.com/jobs/view/software-en...	14 hours ago	2024-04-10 17:07:01.748296
2	Software Engineer (Java)	Bellwood	https://www.linkedin.com/company/bellwood-labs...	Atlanta, GA	Mid-Senior level	Full-time	Engineering and Information Technology	Internet Publishing	About The JobWe are seeking a Software Enginee...	https://www.linkedin.com/jobs/view/software-en...	2 weeks ago	2024-04-10 17:07:01.748296
3	Software Engineer I	Chewy	https://www.linkedin.com/company/chewy-com?trk=...	Bellevue, WA	Entry level	Full-time	Engineering and Information Technology	Retail	Our Opportunity:Chewy is hiring a Software Eng...	https://www.linkedin.com/jobs/view/software-en...	16 hours ago	2024-04-10 17:07:01.748296
4	Software Engineer - Builders	CoinList	https://www.linkedin.com/company/coinlist?trk=...	New York, NY	Entry level	Full-time	Engineering	Technology, Information and Internet	CoinList is where early adopters invest in and...	https://www.linkedin.com/jobs/view/software-en...	3 weeks ago	2024-04-10 17:07:01.748296
5	Software Test Engineer	Rishabh Software	https://in.linkedin.com/company/rishabh-softwa...	Santa Clara, CA	Not Applicable	Other	Engineering and Information Technology	IT Services and IT Consulting	Job Title: Software Test engineerLocation: San...	https://www.linkedin.com/jobs/view/software-te...	20 hours ago	2024-04-10 17:07:01.748296
6	Software Developer M1	Ncontracts	https://www.linkedin.com/company/ncontracts?fr...	Brentwood, TN	Mid-Senior level	Full-time	Engineering and Information Technology	Internet Publishing	Software Developer M1Remote	Product and Deve...	https://www.linkedin.com/jobs/view/software-de...	8 hours ago	2024-04-10 17:07:01.748296
7	Software Engineer, Platform	Tecton	https://www.linkedin.com/company/tectonai?trk=...	San Francisco, CA	Not Applicable	Full-time	Engineering and Information Technology	Software Development	At Tecton, we solve the complex data problem i...	https://www.linkedin.com/jobs/view/software-en...	9 hours ago	2024-04-10 17:07:01.748296
8	Engineer I - Software	General Dynamics Electric Boat	https://www.linkedin.com/company/electric-boat...	New London County, CT	Entry level	Full-time	Engineering and Information Technology	Defense and Space Manufacturing	The IT Software Engineering Group is seeking a...	https://www.linkedin.com/jobs/view/engineer-i-...	12 hours ago	2024-04-10 17:07:01.748296
9	Software Engineer Intern	Trioscope	https://www.linkedin.com/company/trioscope?trk=...	United States	Internship	Full-time	Engineering and Information Technology	Movies, Videos, and Sound	Trioscope is seeking a Software Engineer Inter...	https://www.linkedin.com/jobs/view/software-en...	7 hours ago	2024-04-10 17:07:01.748296

Figure 3: List of job listings from LinkedIn based on the user's basic preferences.

textssss

Images, Graphs, Plots

textsss

References