Project Report - Data Collection Lab 094290

Amit Zalle - 213126402 - amit.zalle@campus.technion.ac.il
Raz Biton - 315507780 - raz.biton@campus.technion.ac.il
Shalev Hermon - 208159400 - 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

For results, please see APPENDIX section.

Original Data Sets

The 'Job Advisor' leverages 'profiles' and 'companies' datasets to match users with suitable job positions, analyzing profile similarities.

Additional Data Collection

Additional data includes job postings, user data, and information about companies from 'LinkedIn', 'Indeed', and 'ZipRecruiter', top job search websites.

Tools and Methods:

- Bright Data Web Browser: Used with Playwright to connect to job search pages and retrieve data via the Bright Data proxy network.
- Playwright: Navigated job search and company pages, interacted with job listings, and managed page navigation.
- Beautiful Soup: Parsed HTML content to extract job details.
- Tkinter UI: Allowed manual input for user profile completeness.
- Additional Methods: Overcame challenges accessing LinkedIn data and infinite scrolling.

Overall, the project successfully gathered job postings, user data, and company information, overcoming challenges with network solutions and ensuring data completeness through manual UI input.

Additional Data Integration

Additional data integrates into two models: Model 1 identifies similar profiles, while Model 2 recommends job positions using open listings and company data.

Enrichment

- Open job positions items: consist of job listings gathered from multiple platforms. The initial size: a(Open Positions)=0.1, actual enrichment size: 7182 for models evaluation. For each new user the 'Job Advisor' can collect new updated job listings.
- Companies: Added companies not in the original dataset, a(Companies)=0.2, actual enrichment: 2530.
- User data: Includes all available user information, with further collection deemed unnecessary. a(people) = 2, actual enrichment is 1. Further collection is unnecessary to achieve our objectives.

Data Analysis

BM25 - Profile to Profiles Similarity

Data Assumptions and Filtering: We assumed that profile data comes from the same distribution, allowing us to learn about a new profile from others. To optimize model results, we filtered profiles without past jobs or education. we applied the same pre-processing on profiles and the user to be able to compare them.
Analysis Techniques: We employed the BM25 information retrieval model for profile-profile similarity. By treating the user as query and existing profiles as documents, we assessed their similarity based on term frequencies. Preprocessing involved converting profiles into word lists using PySpark UDF functions, followed by representation as TF-IDF vectors.
Feature Selection: We applied three separate information retrieval to capture aspects of profile qualification and general description: past and current companies for experience, detailed degree information for education, and general information like the "about" column for personal description. This holistic approach yielded success in both professional and general similarity assessment.

Word2Vec - Profile to Job Similarity

Data Assumptions: We assume access to top profiles similar to the user, obtained from the first model.
Analysis Techniques and Feature Selection: This model recommends relevant jobs using Word2Vec. We extract job descriptions from LinkedIn profiles of users similar to the target user, focusing on relevant skills and experience. Additional preprocessing involves removing stop words. Our feature selection combines domain knowledge and statistical analysis:

- Domain Knowledge: Emphasizing past job descriptions aligns with the assumption that similar users' jobs are relevant recommendations for the target user. Skills and experience in job descriptions inform potential matches.

- Statistical Analysis (Word2Vec): By transforming job descriptions and open position titles into numerical vectors, we calculate similarity scores, reflecting semantic alignment between user experience and job requirements. This approach moves beyond simple keyword matching.

Combining domain knowledge with Word2Vec allows us to generate targeted job recommendations aligned with the user's background.

AI Methodologies

BM25 - Profile to Profiles Similarity

We employed the BM25 information retrieval model to rank profiles based on their similarity to a new profile. Using TF-IDF values calculated earlier, we computed BM25 scores for each document for every term in the query.The formula used is:

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

where N is the number of documents, q_i is term i in q, and $k_1 = 2, b = 0.75$ are hyper parameters.

After calculating scores for each segment, we normalized and aggregate them with the user's preferences weights, then sort the profiles based on the results.

Word2Vec - Profile to Job Similarity

We utilized Word2Vec for semantic analysis, leveraging word embeddings to capture the meaning of words. By analyzing past job descriptions from similar professionals' LinkedIn profiles and scraped open position titles, we move beyond simple keyword matching. Word2Vec's ability to represent words as numerical vectors allows us to calculate similarity scores based on semantic relationships. This enables us to identify relevant job recommendations for the user, even when descriptions use different terminology but convey similar skills and requirements.

Evaluation and Results

In the evaluation of our model, we came across a problem. since our model finds similarity and recommendations on untrained data, there is no numerical way to evaluate our model.

We have chosen to evaluate it in a human-evaluation, on a random example profile. for this user, our model shown much success. giving profile who are very similar to him in the BM25 model, and job offers who very much fit his past jobs and education. You can see those result in the Images, Graphs, Plots section below.

There are ways new ways to examine those kind of models using LM, but those are still new and will not necessarily work, and with our limited time we were unable to use them.

Limitations and Reflection

Despite our success in scraping job listings, challenges persisted. Scraping websites lacking APIs led to data collection delays. While we addressed login requirements on platforms like LinkedIn, Bright Data's limitations hindered access to necessary user data. This forced manual data collection via a UI.

Filtering profiles for relevance from the static table led to data loss, potentially excluding valuable profiles. This filtering could be eliminated with more data, faster GPU processing, or additional information.

In the W2V model, inconsistent job descriptions posed challenges. Abbreviations and unclear wording hindered word2vec's accuracy in capturing job requirements. Additionally, limited information in job titles may cause word2vec to overlook crucial skills or experience mentioned elsewhere.

Better hyperparameter selection is possible with more time and computational resources for additional model runs.

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.

APPENDIX

Data collection pipeline: in Figures 3-5 you can see the general approach of the 'Job Advisor' data collection. First through the UI it gather some user complementary data and basic preferences for first job filtering, then it scraped the user linkedin profile and Job listings.

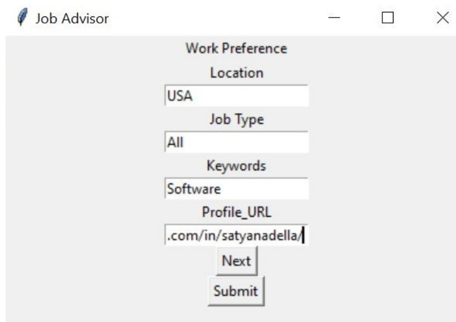


Figure 1: After completing his profile data manually, the user enters his basic job preferences.

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 OpportunityChewy 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?tr...	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.

	Title	Similarity	Company_Name	Company_URL	Job_Location	Seniority_level	Employment_type
0	Foreman/Project Manager	0.499091	Certified Apartment Staffing	https://www.linkedin.com/company/prolific-staf...	Arlington, TX	Mid-Senior level	Full-time
3	Admin/Compliance Analyst-Trainee (Korean)	0.493735	ecocareers	https://uk.linkedin.com/company/ecocareers?trk...	New York, NY	Internship	Full-time
4	SENIOR ACCOUNTANT/ACCOUNTING MANAGER	0.493459	Milestone Property Management	https://www.linkedin.com/company/milestone-pro...	Portland, OR	Mid-Senior level	Full-time
6	SENIOR ACCOUNTANT/ACCOUNTING MANAGER	0.493459	Source 1 Solutions	https://www.linkedin.com/company/source-1-solu...	Clearwater, FL	Mid-Senior level	Contract
7	HR Coordinator/Recruiter	0.491202	American Pool	https://www.linkedin.com/company/american-pool...	Miami, FL	Entry level	Full-time
8	Inside/OSP Manager	0.486677	TekSynap	https://www.linkedin.com/company/teksynap?trk=...	Arlington, VA	Mid-Senior level	Full-time

Figure 4: Most Recommended Open Jobs For User

Images, Graphs, Plots

We will show the process of our model for a user from the profiles static table with the ID: "denise-rathburn-9138a961"

BM25 model

We start by getting the BM-25 score for each segment:

Top Scores for Experience

position	current_company	experience	id	bm25_score
Credit - Accounts Receivable	{"company_id":"robert-half-international"...,"name":"Robert Half","title":"Credit - Accounts Receivable"}	{"company":null,"company_id":null,"description":"Our ... placement of accounting and financial professionals like myself... financial reporting, ..., tax research ...}, {"company":null, ..."title":"Accounts Receivable/Collections...},{"company":null,... Credit Card accounts and ... Accounts Payable ...Corporate Credit Card accounts ...	denise-rathburn-9138a961	46.38732977
Assistant Credit Manager at OMEGA Federal Credit Union	{"company_id":null...,"title":"Assistant Credit Manager at OMEGA Federal Credit Union"}	[{"company":"OMEGA Federal Credit Union",...description":"Assistant credit Manager overseeing ... VISA Credit Card ... the Loan Dept which ... VISA Credit Card Program....	laura-hillard-3b295260	21.31781984
Accounts Payable at Top of the World Headwear	{"company_id":"top-of-the-world-brand",...., "title":"Accounts Payable ...	[{"company":"Top of the World Headwear"... {"description":"Accounts Payable",... , Accounts Payable",...	debi-smith-a2aa7367	13.76351008

Top Scores for Education

educations_details	education	id
Walsh College of Accountancy and Business Administration	[{"degree":"Bachelor of Business Administration (BBA)",...,"title":"Walsh College of Accountancy and Business Administration",...	denise-rathburn-9138a961
Walsh College of Accountancy and Business Administration	[{"degree":"Bachelor of Business Administration (BBA)",...,"field":"Computer Information Systems",...,"title":"Walsh College of Accountancy and Business Administration",...	billmckenziemi
Walsh College of Accountancy and Business Administration	[{"degree":"Bachelor of Business Administration (BBA)",...,"title":"Walsh College of Accountancy and Business Administration"...	michael-thibault-a4134816

we see in the two graphs the similarity of the information of the top scores for the corresponding segment. This shows how successful the BM25 model is in finding the similarity between the users.

Final result of the model

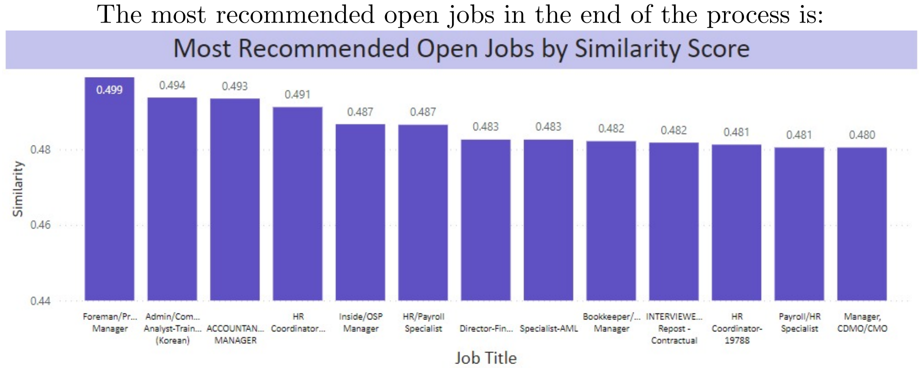
total_score	information_score	education_score	experience_score	id
217.5478732	36.32651082	42.05937306	139.1619893	denise-rathburn-9138a961
63.95345951	0	0	63.95345951	laura-hillard-3b295260
42.05937306	0	42.05937306	0	michael-thibault-a4134816
41.29053023	0	0	41.29053023	debi-smith-a2aa7367
40.05373341	0	0	40.05373341	brenttney-davis-b6428192
40.05373341	0	0	40.05373341	marcia-puntini-008a181b
40.05373341	0	0	40.05373341	debra-kahrmann-1496a16b
40.05373341	0	0	40.05373341	beth-breisblatt-bb9b63119
40.05373341	0	0	40.05373341	angela-evans-5ba613a2
40.05373341	0	0	40.05373341	john-craft-iii-4980b950
39.29764845	0	6.366909251	32.9307392	kelly-bulmahn-930654a2

The final result of the model will return the top profiles and will send them to the next model of Word2Vec.

Word2Vec model

We can see in the results below, that the most recommended open jobs for the user seems adjusted to his profile, based on his education, and experience. The main offers were either related to accounting - which similar to his past jobs, or related to managing, which fit his education of BBA (meaning offer new field of jobs that good for him).

The most recommended open jobs in the end of the process is:



In addition, we show details about the recommended open jobs:

Most Recommended Open Jobs by Similarity Score

Title	Company Name	Job Location	Similarity
Foreman/Project Manager	Certified Apartment Staffing	Arlington, TX	0.50
Admin/Compliance Analyst-Trainee (Korean)	ecocareers	New York, NY	0.49
ACCOUNTANT/ACCOUNTING MANAGER	Source 1 Solutions	Clearwater, FL	0.49
ACCOUNTANT/ACCOUNTING MANAGER	Milestone Property Management	Portland, OR	0.49
HR Coordinator/Recruiter	American Pool	Miami, FL	0.49
Inside/OSP Manager	TekSynap	Arlington, VA	0.49
HR/Payroll Specialist	Pressed Juicery	Culver City, CA	0.49
Director-Finance	NYU Grossman Long Island School of Medicine	New York, NY	0.48
Specialist-AML	KeyBank	United States	0.48
Bookkeeper/Office Manager	NorthPoint Search Group	Birmingham, AL	0.48
Bookkeeper/Office Manager	Staff Financial Group	Orlando, FL	0.48
INTERVIEWER/TRANSLATOR Repost - Contractual	State of Maryland	Maryland, United States	0.48
HR Coordinator- 19788	TalentZök	Orange, CA	0.48

We plot a word cloud of recommended open jobs based on the titles

The Similarity Spectrum: A Word Cloud Analysis

References

[1] Bright Data. How to scrape linkedin: 2024 guide.
[2] dev.to. Tutorial: Web scraping linkedin jobs with playwright.
[3] Scrapingdog. Web Scraping LinkedIn Jobs using Python (Building Job Scraper)