Final project - Data Lab 1
GNNious Solution to an Old Problem

Diana Morgan (336472261), Sergiy Horef (341332336), Veronika Sorochenkova - (342588589)

Datasets and Project Videos, Auxilary Files on GitHub

Contents

1 Project Introduction

Finding the right job is a complex challenge, often relying on inefficient manual screening. Our project addresses this by using Graph Neural Networks (GNNs) to improve job matching.

Each user profile is represented as a graph with two types of nodes: the user and their previous job positions. By leveraging GNNs, we generate feature vectors that align user profiles with suitable job positions based on cosine similarity.

This approach enhances job recommendations by capturing deeper relationships between users and jobs, improving efficiency, and reducing search time.

2 Data Collection and Integration

To train our model, we used a combination of original user profile data and external data about job postings. The original dataset contained user profiles, including their past job positions.

As the original data provided to us by Bright Data did not include job postings, we performed web scraping using Bright Data's Scraping Browser proxy to collect job postings from Indeed.com . Additionally, to enhance our dataset, we incorporated open-source job datasets (that we have found by searching on the internet) and merged it with our scraped data[1]. This combined dataset provided a richer and more diverse job pool, improving the model's ability to match candidates with relevant positions.

By integrating multiple data sources, our solution ensures a comprehensive job matching process, leveraging both structured user information and a broad set of job postings to improve recommendation accuracy.

In this context, we consider each individual job posting as a single item (row), such that its properties (company name, position title, job location, salary, job type and job description) are the columns.[2]

In total, our merged dataset contains 133834 such (non-None) rows, out of which we have scraped 3780.

3 Data Analysis

3.1 Analysis Techniques

Mostly, our data analysis consisted solely of comparing the distributions of None and non-None elements in different columns for feature selection, so we talk about it more in the next subsection.

3.2 Feature Selection

Our feature selection process had two major steps.

Because we are doing matching of profiles with job-postings, we needed to select the columns (features) to represent the encoding of a profile, and the same for the job-postings.

3.2.1 Features for Job-Postings

As we were planning to use previous experiences of a profile as nodes affecting the final feature-vector used to find new job positions, we wanted the job postings to have the same features as the encoding of an experience for a profile.

Taking the intersection of the experience features and the columns of job postings that we have scraped, we ended up with the following: company name, position title, job location and job description.

[1] Please read Appendix 1 to see the full links to the datasets.

[2] Picture of an example of a few rows from the combined dataset can be found in Appendix 2.

3.2.2 Features for Profiles

While choosing the features to represent the profiles our decisions were heavily based upon the column analysis we have done in the previous homeworks in the course, as well as general domain knowledge.

We have included a picture of all columns in the profiles table in Appendix 2, however out of all columns we have only considered the following (others are either irrelevant, or have more than 50% missing): about, city, education, posts and experience, and also id and name for reconstruction.
Out of those posts had very small number of non-empty rows, so we removed that too.
In total we are left with about, city, education and experience which are a good representation of human features, in our opinion.

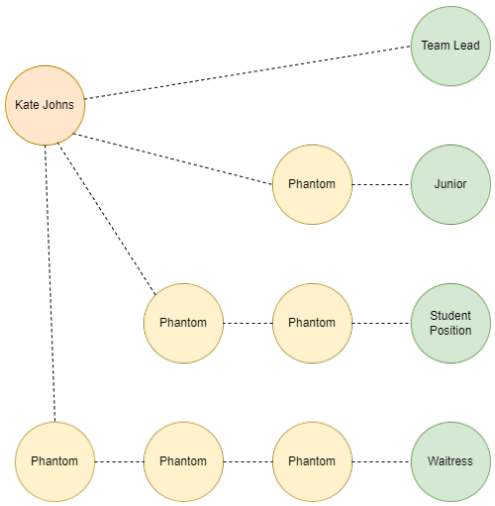
4 AI Methodologies

Our approach utilizes Graph Neural Networks (GNNs) to represent user profiles based on the properties of the profile as well as person's previous job positions. We employed the following methodologies:

- **Initial Embeddings:** Each user, as well as each job posting or user experience was encoded into a single column - by a simple concatenation of all the textual fields. Later, this single textual representation was used to create an embedding from a Bert model (Kenton and Toutanova (2019)) (from spark-nlp) with a dimension of 1024, and only considered the first 512 words. This is not a problem in our case, as on average, the description (being the longest text) had 19 words, and other properties (like title or company name) had around 3 to 4 words. For users a very similar situation holds, with about being the longest and containing few words on average.
- **Graph Representation:** Each user profile is modeled as a graph, where 3 major node types are present.

  - One node is used to represent the profile and its embedding.
  - Another type of node is used to represent the previous job positions of this profile. (Their embedding corresponds to the embeddings of the corresponding jobs.)
  - And one last node type is called a "phantom" - which has a zero-embedding and its sole purpose is to slow done the propagation of information from the job position to the profile.

  Edges in the graph are directional and pass from the job position nodes to the profile node through the phantoms, such that the temporal order of job positions in the experience list of the person determines the number of phantoms on this path.
  If the position is at index i in experience list (starting from 0), its path will contain i phantoms - and so $i + 1$ steps will be needed to propagate the information from the job to the profile.
  Note: in case a profile doesn't have previous positions, the graph contains only the profile node and a singe phantom connected to it.



- **Graph Embeddings:** We trained a GIN (Graph Isomorphism Network) (Xu et al. (2018)) which is (theoretically) the most expressive type of a Graph Neural Network, to generate an embedding for the profile node such that it will be close in cosine-similarity to a "perfect" job position for this person.

  A bit on GNNs: at each iteration of a general graph neural network for each node it takes the embeddings of its neighbors (w.r.t the directed edges that enter it), applies some function to each embedding (defined as a "message"), and later applies some other function to the aggregation of all the messages. This final result is used to set the new embedding of the node.
  Now, in GIN both of those functions are learnable simple Feed Forward Neural networks, and therefore it tends to capture complex relations very quickly.

- **Model Training:** We have trained our model in a supervised setting.

  – We took a random sample of close to a 100 profiles (this small number is due to the computational limitations of our cluster), which we chose to be informative - so only profiles which had a non-empty about and at least one previous position with a non-empty description could be chosen.

  – These profiles were divided into a training and a test sets in a ratio of 80% to 20%. In order to prevent data-leakage (when some data is present at the training stage which will not be present at the test) we randomly chose half of the training set and hid the previous positions from the model.
  In this way, half of the points could teach the model to use the previous experience information. And another half would make sure that our model can work in cases when a person is searching for her first job (arguably the more interesting cases).

  – At each epoch in training, the model was given a graph built for each training profile. Given the model's prediction for a given person, the loss is calculated as follows:

$$\frac{1}{\#pos}\sum_{i}^{\#pos}(pred-(\frac{1}{2}+\frac{1}{2^i}))^2$$

  That is, the loss is a mean squared error between the cosine similarity of the embedding predicted by the model and each of the previous positions of a person, and a number which in our opinion should represent that earlier positions have less importance.
  In this way the lowest similarity (in our opinion) can be $\frac{1}{2}$. And the similarity with the current (or at least most recent) position should be 1. The importance decays exponentially with the position index.

  – Now, if we had only shown our model "positive" cases, it is possible that it would simply learn to maximize the cosine similarity (as we compare it with values greater than 0.5). Therefore for each person we also find the similarity with $\frac{\#pos}{2}$ negative cases, which are

embeddings from the job-listings dataset chosen to be the most dissimilar in cosine similarity to the average of the positions of the person.

- **Cosine Similarity Matching:** After obtaining the embedding for the profile, we uses cosine similarity (of the profile and job embeddings) to rank job positions based on their relevance to a given user profile.

5 Evaluation and Results

We have picked at random one person who had previous jobs and one who didn't, and for each we have run our model to predict the next possible positions.
The full results can be found in Appendix 3. However, for the person who had previous positions, where each one was different from any other (and arguably not jobs at all) Khuong Phan, our model returned the best next position to be "Change Management Coordinator" (which is not bad at all!) (All Predictions).

And for a person who did not have previous jobs (Kyle Scheich), but based on her about was experienced with coding, SQL, etc. our model had predicted the first job to be "Product Manager - Data Science", which is highly relevant (All Predictions).

In our opinion these are not bad results, and we can certainly see why these jobs are proposed. (Note: it is important to keep in mind that we had to take a sample of 1000 points from the set of all jobs, and some of those are nulls (for some reason, as we do filter nulls out), so it is possible that these are indeed the best jobs present.)

The main problem of our domain is that there is no objective metric by which predictions of our model could be evaluated. How can we decide if some job is better than another, if they are approximately similar in their domain and other properties?
Natural language and intrinsic human complexity add an extra layer of harness to the evaluation problem.

We have considered using a Large Language Model to evaluate the results - that is, to provide it with the information about a person and the jobs predicted by our model, and ask it to give a "grade" to the predictions.
This seemed like a natural choice given the natural language of the results. However, this approach has two major flaws: firstly - we couldn't get a reasonable enough model to work on our cluster; and secondly - LLMs tend to hallucinate, and this problem becomes worse as the simplicity of the model increases.
In total, given our computational limitations, we couldn't rely on LLMs.

We have resorted, therefore, to use our loss as an evaluation measure (on the test set). It seems relatively reasonable to us that the encoding should be relatively similar to the person's previous positions, and our results clearly support our claims.
In order to measure how much information our model actually learns, we calculated the loss of a Null model, that is - of a model that always returns a 0 vector (and so has cosine similarity of 0), and have compared between the two.

Here is a graph which shows the test loss of the model and Null model as a function of the epoch.

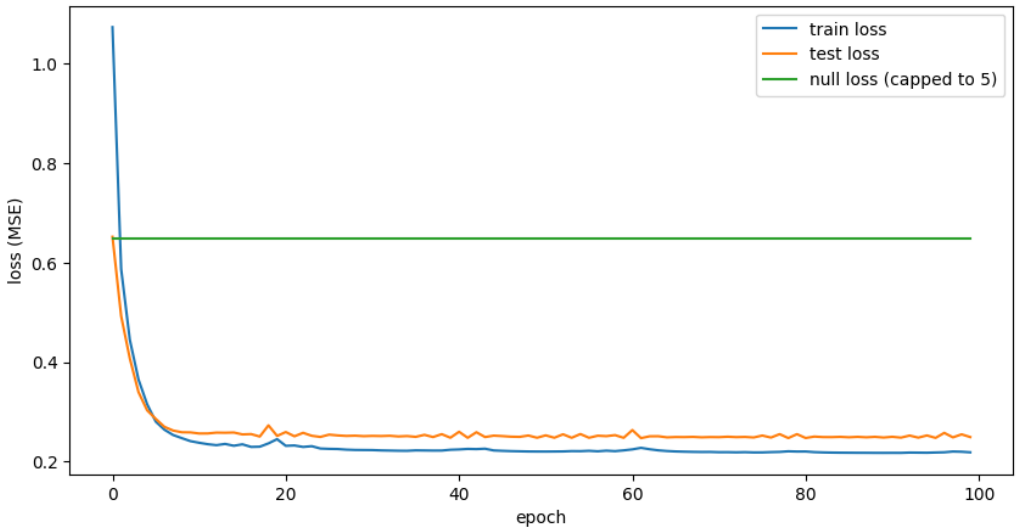


Figure 1: MSE loss comparison for train, test and baseline

It can be clearly seen that our model learns useful information which helps it to distinguish between different jobs.

6 Limitations and Reflection

During the project, we faced several constraints that influenced our approach and outcomes:

- **Resource Availability:** Our main limitation was the lack of computational resources. The virtual machine we used did not have enough capacity to incorporate larger datasets or experiment with more sophisticated models.
  It could be said that this is mostly a self-imposed constraint, as we could have gone to work locally and therefore have access to more resources. However, we made a point of working solely on DataBricks and spark such that if more resorces are available this could be a scalable project.

- **Technological Constraints:** Due to hardware limitations, we had to simplify certain aspects of our approach, which may have impacted the model's performance and accuracy.
  For example, we have started by planning to create larger graphs where companies could server as additional connectors between different profiles, but had to backtrack to one-profile graphs due to an impossibility of running such a system. The same goes for the number of points we could use to train our model on.

- **Time Restrictions:** We were also constrained by time, as issues with the virtual machines delayed our progress and limited our ability to experiment with additional methodologies.

Despite these challenges, our project demonstrates the potential of GNNs for job matching. Future work could explore better resource allocation and alternative computational approaches to overcome these limitations.

7 Conclusions

As we have seen throughout this project, even with such a small number of training samples our model manages to learn non-trivial information and provide reasonable results.
This clearly shows the value of approaches using graph neural networks, and gives much promise for the achievements such models could provide if made scalable.

8 Appendix 1 - Links to public Datasets

Dataset of 124k job postings on Kaggle - Linkedin 124k.
Dataset of 82k tech job postings on GitHub - Tech Linkedin 82k.
Dataset of 4k job postings on GitHub - Glassdoor 4k.

9 Appendix 2 - Images

	company_name	title	location	salary	type	description
1	CyberCoders	Senior Data Engineer	Alexandria, VA	null	Full-time	Job Title: Senior Data Engineer Location: Alexandria, VA Salary Range: $120k - $150k Re
2	Johnson & Johnson	Principal Full Stack Software Engineer.	Santa Clara, CA	null	Full-time	Ethicon, part of Johnson & Johnson Medical Devices Companies, is recruiting for a Princi
3	Microsoft	Senior Software Engineer	Washington, DC	$120,000.00/yr - $189,000.00/yr	Full-time	Microsoft's WCB health team is looking for a Senior Software Engineer who loves to learn,
4	Microsoft	Senior Software Engineer	Reston, VA	$120,000.00/yr - $189,000.00/yr	Full-time	Microsoft's WCB health team is looking for a Senior Software Engineer who loves to learn,
5	Microsoft	Senior Software Engineer	Irving, TX	$108,000.00/yr - $175,000.00/yr	Full-time	Microsoft's WCB health team is looking for a Senior Software Engineer who loves to learn,
6	Microsoft	Senior Software Engineer	Atlanta, GA	$106,000.00/yr - $172,000.00/yr	Full-time	Microsoft's WCB health team is looking for a Senior Software Engineer who loves to learn,
7	Microsoft	Senior Software Engineer	Charlotte, NC	$105,000.00/yr - $170,000.00/yr	Full-time	Microsoft's WCB health team is looking for a Senior Software Engineer who loves to learn,
8	Cleerly	Staff - Data Scientist	Denver, CO	$130,000.00/yr - $200,000.00/yr	Full-time	Position Title: Staff Data Scientist Position: Full Time, Exempt Supervises Staff: No Salary
9	Microsoft	Senior Software Engineer	Reston, VA	$120,000.00/yr - $189,000.00/yr	Full-time	Are you passionate about architecting and developing mission critical cloud solutions that
10	Microsoft	Senior Software Engineer	Charlotte, NC	$105,000.00/yr - $170,000.00/yr	Full-time	Are you passionate about architecting and developing mission critical cloud solutions that

Figure 2: Example of rows in the job-listings dataset

1	about	string
2	avatar	string
3	certifications	array<struct<meta:string,subtitle:string,title:string>>
4	city	string
5	country_code	string
6	current_company	struct<company_id:string,industry:string,link:string,name:string,title:string>
7	current_company:company_id	string
8	current_company:name	string
9	education	array<struct<degree:string,end_year:string,field:string,meta:string,start_year:string,title:string,url:string>>
10	educations_details	string
11	experience	array<struct<company:string,company_id:string,description:string,duration:string,duration_short:string,end_date:string,locat...
12	followers	bigint
13	following	bigint
14	groups	array<string>
15	id	string
16	languages	array<struct<subtitle:string,title:string>>
17	name	string
18	people_also_viewed	array<struct<profile_link:string>>
19	position	string
20	posts	array<struct<attribution:string,created_at:string,img:string,link:string,title:string>>
21	recommendations	array<string>
22	recommendations_count	bigint
23	timestamp	string
24	url	string
25	volunteer_experience	array<struct<cause:string,duration:string,duration_short:string,end_date:string,info:string,start_date:string,subtitle:string,titl...
26	courses	array<struct<subtitle:string,title:string>>

Figure 3: Full schema of columns in the job-listings dataset

10 Appendix 3 - Proposed Jobs

For a person for whom our model observes previous positions:

	id	name	pre_enc	experience_index
1	khuong-phan-15142523a	Khuong Phan	Radio Disc Jockey Fillmore, California, United States I thought I told you you can't scam me for work that I've done in life on ...	0
2	khuong-phan-15142523a	Khuong Phan	Spy Washington, District of Columbia, United States I think the grandpa is in the house with the newspaper but the kid and th...	1
3	khuong-phan-15142523a	Khuong Phan	Boss Ho Chi Minh City, Ho Chi Minh City, Vietnam If you had any idea how walmart cared about him then why do keep putting...	2
4	khuong-phan-15142523a	Khuong Phan	Lieutenant Tully, Hauts-de-France, France Date me you fool your boobis better I thoughts so good ol m the Donna's be bette...	3
5	khuong-phan-15142523a	Khuong Phan	Dancer Biên Hòa, Dong Nai, Vietnam I went dancing for a wedding but the date was tired she needed rest I date raped her an...	4
6	khuong-phan-15142523a	Khuong Phan	Writer Missouri, United States Hi Pam can you shower with me one more time I think the bonus could be bought everytime or ...	5
7	khuong-phan-15142523a	Khuong Phan	Senior Technical Writer Missouri, United States Hi Pam can you shower with me one more time I think the bonus could be bou...	6
8	khuong-phan-15142523a	Khuong Phan	Actor South San Jose Hills, California, United States San Jose State Iniversity can sick m u dick for some Crack u o u stupid a...	7
9	khuong-phan-15142523a	Khuong Phan	Entertainer San Bernardino County, California, United States The baby is hurs she was a little girl to be as a boy I know all kin...	8
10	khuong-phan-15142523a	Khuong Phan	Disk Jockey New York City Metropolitan Area I'll make love to you til your body stops let's go slow I ain't got nowhere to left t...	9
11	khuong-phan-15142523a	Khuong Phan	Singer San Jose, California, United States I drop the guy but he had it in the ducking movie tweet is what he called a guy for ...	10

Figure 4: Previous positions for a person for whom they are observed by the model

	company_name	title	location	description
1	Southwest Key Programs	Change Management Coordinator	Houston, TX	Job Summary:: The Change Management Coordinator will be responsible for developing and e
2	PrimeSource Building Products	Accounts Payable Clerk	Miami, FL	Who we are:Prime Matter Labs is a profitable and established personal care product manufac
3	At Home Group Inc.	Remote: Virtual Personal Assistant (CoPilot)	Unknown	Job Title: Virtual/Personal Assistant (Copilot)Schedule: Full time roles available Weekday: Mor
4	Equity LifeStyle Properties, Inc.	Remote: Virtual Personal Assistant (CoPilot)	Unknown	Job Title: Virtual/Personal Assistant (Copilot)Schedule: Full time roles available Weekday: Mor
5	Oracle	Hispanic Heritage Month Celebration - Consultant Opportuniti...	United States	Thank you for joining us for the Virtual Open House: Hispanic Heritage Month Celebration for
6	.	Interviewer	Unknown	About Optimism Optimism is a media company working to build a brighter web. We conceive,
7	Better	Interviewer	Unknown	About Optimism Optimism is a media company working to build a brighter web. We conceive,
8	Continental	Director of Operations	Warren, MI	Description The Director of Operations will lead the strategic development and expansion of c
9	Kaiser Permanente	Community Relations Consultant III	Aurora, CO	Salary Range: $37.93/hour - $49.07/hour Job Summary: Supports the coordination of commu
10	The Scotts Miracle-Gro Company	Assistant Marketing Manager	14111 Scottslawn Rd, Marysville, OH 43040	Here at Scotts Miracle-Gro there is no such thing as a typical day. Our culture is constantly en

Figure 5: Proposed jobs for a person for whom previous positions are observed by the model

For a person for whom our model does not observe previous positions (they do not exist for that person).

	about
1	- More than 10 years of experience in JavaEE development, a solid foundation in JAVA, understanding of basic knowledge of IO, multithreading, reflection, security, etc., and a certain understanding of JVM principles and tuning; - Familiar with Spring, springMVC, Struts, Ibatis, hibernate, SOA service framework, etc., understand its principle and implementation mechanism, read the source code, and imitate its core; - Familiar with scripts jQuery, Extjs, web container tomcat, weblogic; - Familiar with SQL and database-based design and development, master database tuning such as Mysql and Oracle; - Familiar with mogonDB, redis construction and use; - Familiar with common commands of Linux operating system; - Familiar with SVN, GIT, Maven, ANT and other build tools - Understand Hadoop, zookeeper and other open source distributed systems, and MapReduce programming. - Familiar with various data structures and algorithm models - Be good at learning and communicating with others, be upright and honest, have a strong sense of professionalism, and have the ability to analyze and solve application problems; - Experience in designing and developing large-scale e-commerce websites and core banking systems

Figure 6: Previous positions for a person for whom they are observed by the model

	company_name	title	location	description
1	Verizon Media	Product Manager - Data Science	New Jersey	Position: Technical Process/Product Manager with Data Science Locati Verizon
2	The HON Company	ver 2	Beverly Hills, CA 90210	ver 2
3	City of Greensboro	EARLY CHILDHOOD EDUCATION TO KINDERGARTEN TRANSITION COORDINATOR	712 North Eugene St, Greensboro, NC 27401	Fair Labor Standards Act Classification: Non-Exempt Position Term: 12
4	La Joya Independent School District	Spanish Teacher	126 West State Street, Wellsville, NY 14895	The Wellsville Central School District has the following open position: S
5	College of Saint Benedict	FranU Student Worker Fed Study	5414 Brittany Drive, Baton Rouge, LA 70808	Student Worker Fed Study FranU Baton Rouge, La Job Title : Student W
6	La Joya Independent School District	Bilingual (Spanish) Program Assistant	Rolling Meadows, IL	JobID: 7833 Position Type: Program Assistants/Bilingual (Spanish) Dat
7	Aegis Living	Cook	1845 116th Avenue NE, Bellevue, WA 98004	Overview: Cook- Hiring Now! Are you interested in a career without late
8	Men's Wearhouse	3rd Shift Freight Sorter / ParT-time Warehouse Associate / Package Handler	7488 Brokerage Drive, Orlando, FL 32809	Location: 7488 Brokerage Drive, Orlando, FL 32809-5622 Category: W
9	US Army Reserve	United States Army	Fort Myers, FL	Full Job DescriptionLet's see what job in the US Army works best for yo
10	US Army	US Army	U.S. Government in Norwalk, CA 90650	Come sit down with US Army Recruiter: Sergeant Jones in our Norwalk

Figure 7: Proposed jobs for a person for whom previous positions are observed by the model

In both cases the negative jobs (sorted ascending by similarity) are:

	company_name	title	location	description	embedding
1	null	null	null	null	null
2	null	null	null	null	null
3	null	null	null	null	null
4	null	null	null	null	null
5	null	null	null	null	null
6	null	null	null	null	null
7	null	null	null	null	null
8	null	null	null	null	null
9	null	null	null	null	null
10	null	null	null	null	null

Figure 8: Caption

References

Kenton, J. D. M.-W. C. and L. K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Volume 1. Minneapolis, Minnesota.

Xu, K., W. Hu, J. Leskovec, and S. Jegelka (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.