FINDING SIMILAR JOBS VIA SUMMARIES USING LSH AND MINHASHING

DARYA SAVITSKAYA

Contents

1. Introduction

When dealing with large and heavy datasets, such as job descriptions or resumes, it is often a problem to identify items that are "similar" without having to compare every single pair of items. The basic approach of comparing all pairs is computationally prohibitive. For example, if we have 1 million job summaries, comparing every pair would result in half a trillion comparisons. As datasets grow larger, this approach becomes impossible due to the quadratic increase in computation time. (Rajaraman and Ullman, 2022) In this project, using a dataset that contains various characteristics of 1297332 job postings on Linked in, such as link, summary, company, level, type, etc., the objective is to implement a system that detects pairs or sets of similar job descriptions by analyzing the summary of the job posting on Linked In.

2. Data Preprocessing

The dataset, extracted via the Kaggle API, initially contained 1297332 job postings, i.e. two columns a link and a summary. No null rows were present. I started building a systems utilizing only 10% of the dataset, as an average summary contains 550 words, which is quite a lot and computationally heavy. Further, every summary was split into a list of words. Whether to split a summary into words or shingles was a big concern. An important advantage of using words is memory usage - shingles add combinations of letters that capture the order of words. However, an advantage turns into disadvantage when we understand that by not including these additional pieces of words, we are loosing important information about sentence structure and content. However, given similar context of the texts (all of them are job summaries on the same website) and that the information is already pretty heavy, the decision was made to use words. Therefore, every summary was transformed into a list of words. To exclude the words that give us zero to no information, I inititally adapted a list of common stop words in english, such as 'a', 'the', 'in', etc. However, to adapt the list of stop words for this particular dataset, I counted the number of mentions of each words and looked closely at the most popular. The first 30 worlds ended up being mostly common stop words, such as common preposition and pronouns. Additionally, in the context of the dataset. common words like 'work' or 'experience' were also in those thirty words that were removed from consideration. It was also decided to look at the least popular words - it ended up being typos, links, words meshes with symbols - they were also removed.

3. Methodology: Algorithms and Implementation

3.1. **Algorithms.** For the task of scaling the similarity search across a large dataset of job descriptions, I have chosen a combination of MinHashing and Locality-sensitive Hashing as the core techniques. MinHashing is used to efficiently approximate the Jaccard similarity between job descriptions, which are treated as sets of words. By creating compact Minhash signatures for each job description, I can reduce the dimensionality of the data while still preserving the ability to compare sets for similarity. This technique allows for fast, memory-efficient comparisons without needing to store or process the entire dataset in full detail.

Once theMinhash signatures are generated, I use LSH to group similar descriptions into the same buckets with high probability. This technique avoids the computationally extensive process of comparing every pair of items by concentrating only on csandidate pairs that end up in the same bucket. As discussed in the chapter 3, LSH hashes the Minhsh signatures using several hush functions to reduce risk of false positives, it follows that dissimillar items are inlikely to be compared (Rajaraman and Ullman, 2022).

The implementation starts by generating Minhash signatures for each job description through the mh function, which filters out unwanted words, such as stop words and rare words, before encoding the remaining words into the Minhash object. These Minhash signatures serve as compacted representations of the job descriptions. Next, an LSH instance is configured with a similarity threshold of 0.5 and 64 permutations, ensuring that only job descriptions with potential similarity are compared. In essence, it means that there are 64 distinct hash functions to generate a signature for each set of words in a job description. The resulting signature is a 64-dimensional vector, where each value is derived from one of the hash functions. This signature acts as a compressed representation of the original set. The Minhash signatures are then inserted into the LSH structure. There the signatures are hashed into buckets - each is divided into smaller sections or bands, which are then hashed to determine the bucket where the item will be stored. Accordingly, items with similar signatures are likely to end up in the same or nearby buckets and only items in the same bucket are considered potential candidates for similarity checks.

3.2. **Handling Large Datasets.** The solution employs Locality-Sensitive Hashing combined with Minhashing to efficiently scale to larger datasets by reducing the number of comparisons between job descriptions. Instead of a quadratic time complexity, where all pairs would need to be compared, LSH narrows the focus to potential similar pairs, drastically cutting down unnecessary computations. By using words instead of shingles and excluding stop words as well as rare words, the system becomes more efficient in identifying meaningful patterns. The mh function, which applies 64 permutations, further optimizes the process by generating compact representations of each job description.

To handle the large size of the dataset effectively, I propose analyzing it in pieces, processing 10% at a time. While this chunking strategy helps in managing memory usage, an additional proposed step is to store Minhash signatures across batches. This ensures that potential similar pairs from different chunks are compared, maintaining the accuracy of the similarity detection while avoiding missed pairings between batches. These optimizations allow the solution to process larger datasets efficiently, even when handling just 10% of the original dataset at present

4. Results

As an example output, let us look at the jobs that were paired with job 0 and job 1. In the tables below can be found the jobs and their shortened descriptions. Job 0 can ba classified as

Job ID	**Job Title**	**Summary**
job_0	Restaurant Manager	Rock N Roll Sushi is hiring a Restaurant Manager! As our Restaurant Manager, you'll never be bored. You'll be responsible for making sure our restaurant runs smoothly.
job_7690	Assistant Team Leader	Assists the Team Leader in all aspects of daily operations including profitability, expense control, buying, merchandising, labor, regulatory compliance, and special projects.
job_39366	Facilities Manager	Facilities Operations: Oversee day-to-day facility operations, including maintenance, repairs, and compliance with safety regulations at Holsworthy Barracks.
job_86682	Lead Java Programmer	The Lead Java Programmer designs, estimates, develops, and implements well-tested solutions to satisfy application development requests within allocated time and budget.
job_52319	Sous Chef	We are looking for a passionate and experienced Sous Chef to inspire, lead, and mentor our kitchen team with a commitment to excellence.
job_14616	Quality Analyst	Looking for a Quality Analyst with strong knowledge in medical devices or pharma with 2 years of experience, handling reports and working in a regulated environment.

Continued on next page

Table 1 – continued from previous page

Job ID	Job Title	Summary
job_66356	Audit Accounts Manager	This is a management role with a varied split of audit and accounts, focusing on overseeing the workforce and WIP for a significant client portfolio.
job_19516	Associate, Alternatives Investments	The Associate, Alternatives Investments plays a key role in helping drive sales and building relationships with Financial Professionals selling Franklin Templeton Alternative Investment products.
job_67791	Patient Specialist Representative	We are seeking a motivated Patient Specialist Representative with customer services skills to optimize client financial interaction and patient care in a medical office setting.
job_20866	Radiology Technologist Pool	Performs diagnostic radiographic and fluoroscopic procedures, supports operating room and emergency department imaging services, and handles imaging patient care.
job_94894	Assistant Salon Manager	Great Clips is seeking an Assistant Salon Manager with strong technical skills, communication skills, and leadership ability.
job_22071	Safety Manager	This position provides overall leadership and direction for creating a culture of safety throughout the manufacturing facilities and ensuring compliance with regulations.
job_5651	Registered Nurse	Provides direct patient care to patients using the nursing process in accordance with applicable standards, and assists patients with daily living tasks in a psych ward.
job_7793	Pediatric Psychologist	Seeking a pediatric psychologist to expand the Integrated Behavioral Health program and provide behavioral health services to children in General Pediatrics clinics.
job_122523	Learning Support Teacher	Develop and implement an educational program for students assigned to the Learning Support program, collaborating with clinical support teams.

The job postings share several key similarities, primarily revolving around leadership, management, and customer service responsibilities. Even though the industries vary from health care to hospitality, retail and facilities management, all

these roles require managing teams, maintaining standards, and ensuring smooth operations.

Job ID	Job Title	Summary
Job 1	Registered Nurse (RN)	Manages patient care by collaborating with physicians, providing support, and performing assessments and interventions.
Job 100505	Family Dollar Customer Service Rep	Assists customers with merchandise, transactions, and store upkeep, ensuring a safe and positive shopping environment.
Job 107645	City Manager of Altoona, PA	Oversees municipal operations, manages a city budget, and implements policies established by the City Council.
Job 29675	Sr. Insurance & Contracts Specialist	Manages insurance matters, reviews contracts, and coordinates legal claims/issues within the company.
Job 84946	Accounting Manager	Establishes and monitors financial systems, drafts budgets, and provides data for informed financial decisions.
Job 46071	Senior Recruitment Consultant	Manages client relationships, assesses client needs, and provides recruitment and business solutions.
Job 99769	Construction Inspector	Inspects construction projects, performs field testing, and communicates findings to managers and contractors.
Job 126070	Physician - Otolaryngology	Provides care for patients with ear, nose, and throat conditions, earning $350,000-$450,000/year.
Job 105236	Counselor	Oversees patient treatment in addiction recovery, coordinating care from admission to discharge.
Job 89429	Radiologic Technologist	Performs diagnostic radiographic procedures, assisting medical staff with patient care in a clinical setting.
Job 58221	Senior Manager, Security Operations	Manages security operations and systems, ensuring service reliability and maintaining security standards.
Job 117402	Marketing Coordinator	Supports marketing activities like proposal development, client development, and content writing for transportation clients.

Continued on next page

Table 2 – continued from previous page

Job ID	Job Title	Summary
Job 3498	Logistic Manager	Manages logistics operations, ensuring proper shipment and handling of products, maintaining warehouse efficiency.
Job 99059	Lead Sales Associate	Provides superior customer service, manages store operations in the absence of store management, and maintains a clean store.
Job 4914	Substance Use Disorder Counselor	Guides patients in addiction recovery through counseling and develops personalized treatment plans.

Here the variation in industries is even more noticeable. Healthcare is prominent, with roles like registered nurses and other medical professionals, retail is also here through customer service and store management positions. Additionally, the public sector appears, as well as construction and engineering field and corporate and administrative sectors. Even though some false positives are present, we can highlight a high concentration of health-care jobs, that have similar responsilities such as patient assessment, care planning, and coordination with multidisciplinary teams. The jobs are all connected by high management factor and high responsibility.

5. Conclusion

This project serves as the foundation for a scalable system via which it is possible to efficiently detecting similar job descriptions in large datasets. The combination of Minhashing and LSH offers a practical solution that balances accuracy and performance. A system like that could be used in job recommendation engines, where similar job descriptions can be suggested to users based on their current search or preferences or already applied positions, enhancing the job-seeking experience and making the platform more user-friendly. Further refinements and improvements to the underlying techniques, such as experimenting with different hashing techniques or increasing the number of hash functions, could make this system even better in identifying relevant job opportunities for job-seekers.

6. Appendix

A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*, 3rd ed., Cambridge University Press, 2022. Chapter 3.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work, and including any code produced using generative AI systems. I understand that

plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.