Personalized Job Recommendation

APAN 5430 - Group 4

Minkyung Kim | Boni Rosen | Kibaek Kim | Kas Kiatsukasem | Philip Suchakrey

Table of Contents

1 Business Background & Problem

2 Data source & Procurement

3 Selected technologies & Rationale

4 System Development Process

5 Results & Evaluation

6 Limitations & Opportunities for Improvement

7 UI Demo

Business Background & Problem

In an **online job application portal and ATS** today, most recruitments are **relying on keyword-based search** and macthing where recruiters enter skill keywords or technical skill, and candidates are ranked only by those texts overlap.

This approch overlooks:

1. Candidate contextual relevance
2. Full story of candidate's background

System overlooks lead to:

1. Missed the right candidate opportunity
2. Poor overall match quality

Project Goal and Use Case

The goal of this project is to **develop a personalized job recommendation system:**

- Calculates a **match score** between a candidate's resume & job descriptions
- **Rank job postings by compatibility**, showing the most relevant and well-aligned positions first, based on **both explicit qualifications and implicit signals**
- Provides **resume improvement recommendation** using LLM

Use case:

- **Candidate side**: Upload their resumes to receive a ranked list of job openings that best fit
- **Recruiter side:** Identify the most relevant candidates based on overall profile similarity.

Data source

The goal is to combine data from two sources:

- **LinkedIn Job Board:** Most up-to-date and realistic job postings
- **Hugging Face job dataset:** to ensure sufficient data volume and diversity of jobs listed

Web Scraping from Linkedin Job Board

- Scraped ~30K LinkedIn job posts using linkedin_jobs_scraper
- 100 predefined titles (e.g., Digital Marketing Specialist)
- Extracted fields: Job Title, Company, Date, Link, Insights, Description
- Saved up to 10K records per JSON file (30K Total)

Job Description Data Set

The raw data source contained 22 columns with details about job opportunities. We adjusted the dataframe and filtered unnecessary columns. Finally, we have ~1GB data set filtered and includes only 10 important columns for the project

Source: https://huggingface.co/datasets/lof223/job_description

Combined Dataset Size: 1.2GB

Linkedin

1. Job Title - The title of the job position
2. Company - The company name
3. Company Link - URL link to the company
4. Date - The date the job was posted
5. Date Text - Text representation of the date
6. Job Link - URL link to the job posting
7. Insights - Additional insights about the job
8. Description Length - The character count of the job description
9. Description - The full text of the job description

Hugging Face

```
df_all_jobs.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1615940 entries, 0 to 1615939
Data columns (total 11 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Job Id            1615940 non-null  int64
 1   Experience        1615940 non-null  object
 2   Salary Range      1615940 non-null  object
 3   location          1615940 non-null  object
 4   Country           1615940 non-null  object
 5   Job Posting Date  1615940 non-null  object
 6   Job Title         1615940 non-null  object
 7   Company           1615940 non-null  object
 8   Job Description   1615940 non-null  object
 9   Benefits          1615940 non-null  object
 10  skills            1615940 non-null  object
dtypes: int64(1), object(10)
memory usage: 135.6+ MB
```

**Final columns in each data sources*

Selected technologies & Rationale

01 Named Entity Recognition

Purpose:
To extract explicit, **factual entities** from resumes and job descriptions.

- Detects skills, company names, job titles, certifications, degrees, and other key entities.
- **Captures what the candidate knows or has done.**

NER: Skill Score

02 Word/Sentence Embedding

Purpose:
To measure **contextual similarity** between Resume and JD text.

- Computes similarity between the two embeddings to evaluate semantic closeness.
- **Captures what the candidate meant in the resume, even when the wording is different.**

Word Embedding: Semantic Score

03 Topic Modeling

Purpose:
To identify which **domain or industry topics** each document belongs to (e.g., Data Science, Finance).

- Trains an LSA model on JD text to learn latent topic distributions.
- **Captures whether both texts belong to the same industry or domain.**

Topic Modeling: Topic Score

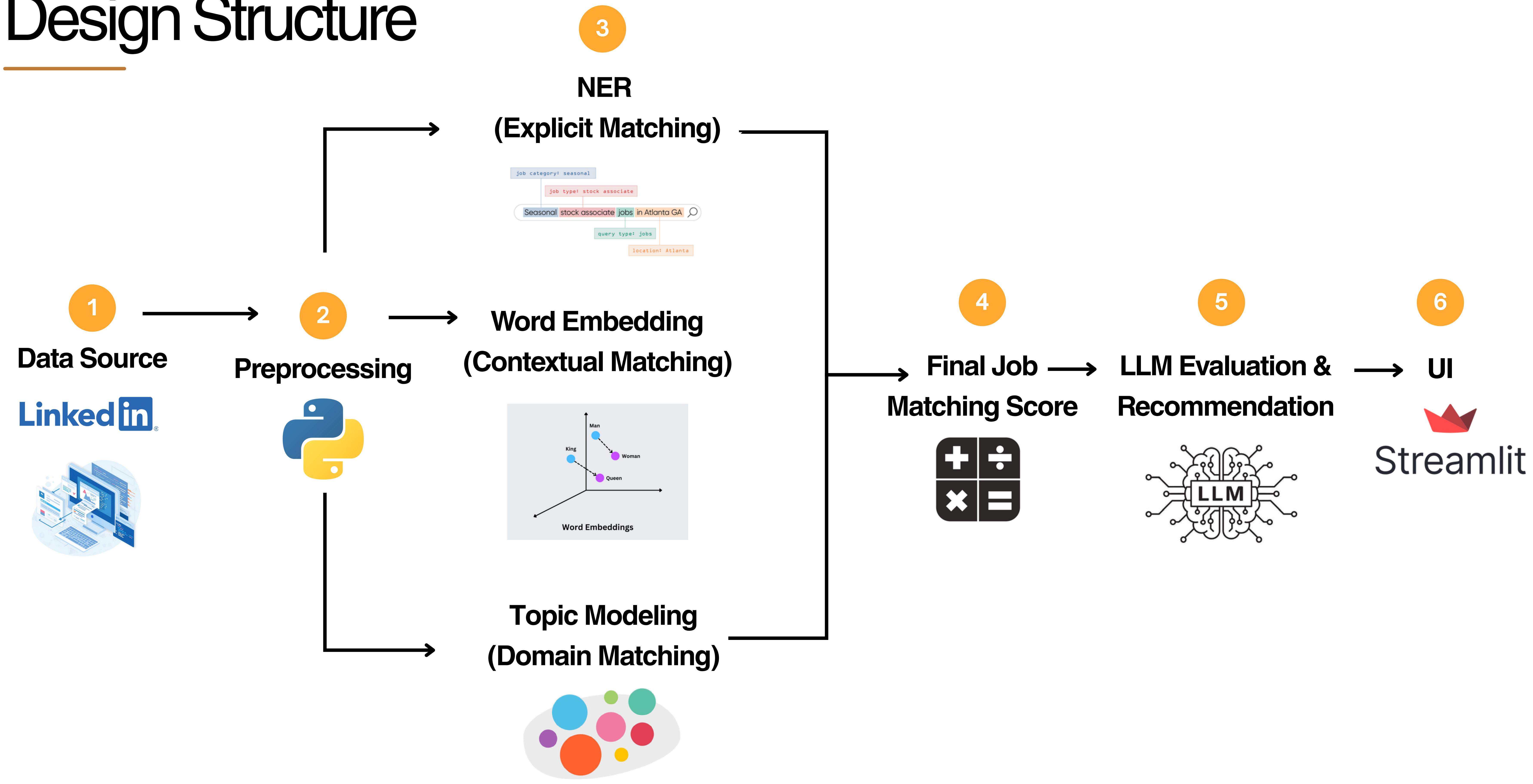
Integrated Scoring Engine - Job Recommendation

Candidate Matching Score Formula: 0.45×SkillScore+0.35×SemanticScore+0.20×TopicScore

avg_topic_semantic = (topic_score + semantic_score) / 2

Selected Final Job Matching Score Formula = avg_topic_semantic + (1 - avg_topic_semantic) * skill_score

Design Structure



Development Process

Data Cleaning

Resume

- **Text extraction** using pymudpf package
- **Text Cleaning & Normalization**
  - Remove Invisible Characters
  - Remove Personal Identifiable Information sections (Name, Emalil, Address)
  - Strip HTML and Special Entities
  - Whitespace & Layout Cleaning
  - Reformat Bullet Points & Fix Typography
  - Keep Only Meaningful Resume Sections
- **Save final cleaned resume dataset**

Job Post

- **Unify required text fields** into a single column
- **Text Cleaning & Normalization**
  - Unicode normalize + remove zero-width characters
  - Remove emails, phone numbers, URLs
  - Remove names like: linkedin jobs, glassdoor jobs, etc.
  - Normalize bullet points
  - Whitespace & Layout Cleaning
- **Identify and remove duplicate** job descriptions
- **Save final cleaned job description dataset**

Data Cleaning - Result

Resume

```
===== CLEANED TEXT PREVIEW =====

SKILLS
 - Programming & Analytics: Python, R - Version Control: Git,
 - Databases & Data Management: PostgreSQL, MongoDB - Environment Contr
 (NoSQL), Navicat, SAP ERP - Microsoft: Excel, PowerPoint, Word, Outloo

 PROFESSIONAL EXPERIENCE
CEF SOLUTIONS, INC. CONSULTING Seoul, South Korea
Business Enablement Associate Sep 2023 - Mar 2024
 - Conducted market research and workflow analysis for a leading Asian
 synthesizing insights that informed strategic recommendations adopted
 - Designed and delivered proposal decks and presentations for stakehol
 stakeholder decision-making.
 - Optimized and standardized process workflows for new initiatives res
 - Managed project timelines and deliverables, leading structured meeti
 continuous progress.

SAMSUNG ELECTRONICS AMERICA Ridgefield Park, USA
Data Analyst Jul 2022 - Jul 2023
 - Performed data validation and exploratory analysis for extensive ven
 reducing transaction errors and ensuring data integrity.
 - Conducted P&L analysis leveraging financial datasets to optimize pri
 long-term operational sustainability.
 - Analyzed large volumes of customer chargeback cases, investigating d
 and strengthen vendor compliance.
 - Leveraged SAP ERP extractions and Power Query to create recurring re
 efficiency and ensuring consistent accuracy.
 - Monitored dashboard data accuracy, identified issues, and delivered
```

Job Post

```
---- Last cleaned job text (wrapped, width=100) ----
Job Title: Business Development Consultant (Travel Partnerships)
CloudofGoods.com Role: Business Development Consultant (Part-Tim
Miami, Port Canaveral), Los Angeles, Anaheim, Las Vegas, Atlanta
hours/week About Cloud of Goods Cloud of Goods is the largest eq
U.S., delivering mobility scooters, wheelchairs, strollers, crib
U.S. cities and select international destinations. We make trave
elderly travelers, families with children, and those who need eq
destinations. Role Overview We are seeking part-time Business De
markets to establish partnerships with travel industry organizat
attractions, cruise lines, conference centers, travel agencies,
focused on outbound business development (cold calling, cold ema
the goal of signing partnership agreements or MOUs. Consultants
local travel industry partners into our affiliate program, where
customers who rent through CloudofGoods.com. Key Responsibilitie
and approach potential partners: hotels, attractions, cruises, c
tour operators. Pitch Cloud of Goods' rental marketplace and exp
guests/customers. Negotiate and secure agreements, MOUs, or affi
Conduct cold calls, cold emails, and in-person site visits to ge
and demos about Cloud of Goods services. Build strong local rela
hospitality ecosystem. Affiliate Program Growth Sign up travel a
Cloud of Goods' affiliate program. Promote affiliate benefits, i
referring customers. Support affiliates with onboarding and mark
Representation Represent Cloud of Goods in local travel industry
and trade shows. Provide feedback on market trends, competitive
```

Named Entity Recognition - Skill Extraction

Modeling Pipeline

1 Custom Skill Dictionary

- Master Skill Dictionary combines 530 core technical and business skills with extended multi-industry

2 Identify technical skills and keywords

- Convert unstructured text into structured skill entities
    - Resumes
    - Job descriptions

3 Utilize pretrained NLP model - spaCy

- en_core_web_lg → (higher accruacy)
- Create a Skill Matcher using PhraseMatcher
    - Skill in skill list → mark them as "SKILL."
- Tokenization: Splitting text into small pieces
- Name Entity Recognition: Labeling important entities in text.

4 Final Similarity Score

- Jaccard Similarity Score (Resume vs Job)

Extracted Skills - Resume

SKILLS

- Programming & Analytics: Python SKILL , R SKILL - Version Control: Git SKILL ,
- Databases & Data Management: PostgreSQL SKILL , MongoDB SKILL - Environment Control: Docker SKILL
(NoSQL SKILL), Navicat, SAP SKILL SAP ERP SKILL - Microsoft: Excel SKILL , PowerPoint, Word, Outlook

PROFESSIONAL EXPERIENCE

CEF SOLUTIONS, INC. CONSULTING Seoul, South Korea

Business Enablement Associate Sep 2023 - Mar 2024

- Conducted market research SKILL and workflow analysis for a leading Asian financial institution's auto-loan initiative, synthesizing insights that informed strategic recommendations adopted by senior management.
- Designed and delivered proposal decks and presentations for stakeholders, enhancing project clarity and accelerating stakeholder decision-making.
- Optimized and standardized process workflows for new initiatives resulting in notable increases in operational efficienc SKILL .
- Managed project timelines and deliverables, leading structured meetings and documentation to ensure alignment and

Extracted Skills - Job Post

Event Manager

Specialize in conference and convention planning. Coordinate speaker sessions, exhibitors, and attendee experiences.

Oversee event registration and marketing.

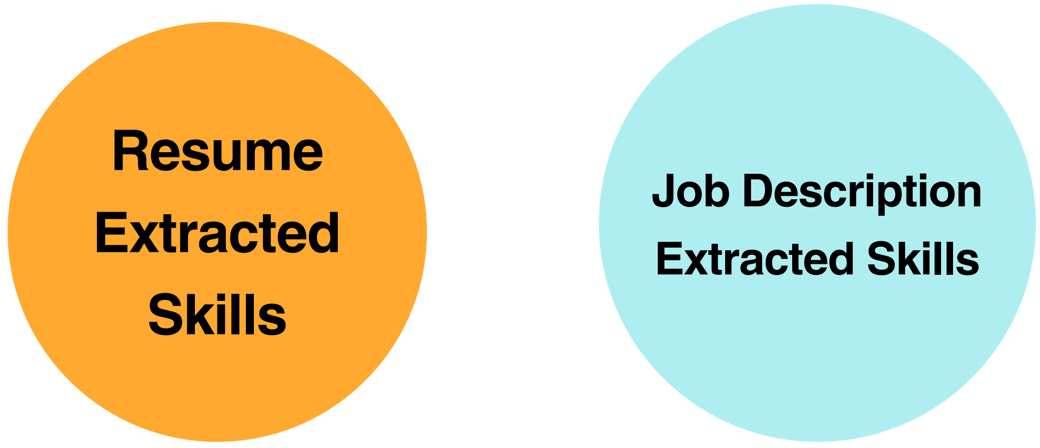
A Conference Manager coordinates and manages conferences, meetings, and events. They plan logistics SKILL , handle budgeting SKILL , liaise with vendors, and ensure the smooth execution of events, catering to the needs and expectation of attendees.

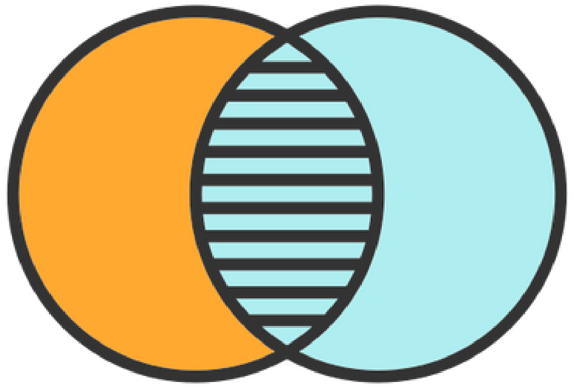
Event planning SKILL Conference logistics SKILL Budget management Vendor coordination Marketing and promotion

Client relations

Named Entity Recognition - Score and Result

Jaccard Similarity Score Calculation





What both sets share (overlap)
What they have in total (union)

Jaccard = $| A \cap B | / | A \cup B |$

Example

[A] Resume Skills
{python, sql, tableau, aws}

[B] Job Description Skills
{sql, python, excel, aws, communication}

Jaccard Similarity Score
Overlap = {python, sql, aws} = 3
Union = {python, sql, tableau, aws, excel, communication} = 6

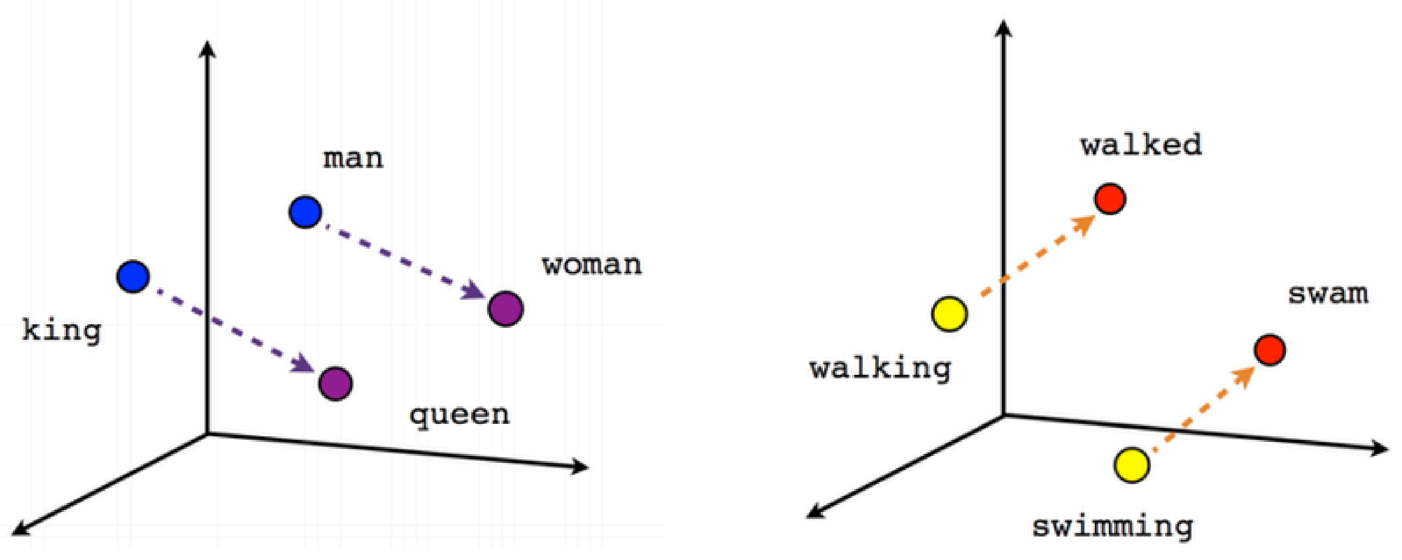
Jaccard = 3 / 6 = 0.50 → **50% matched**

Word Embedding

- **How word embedding works?**
  - Core idea → **“meaning”** captured via geometric distance in vector space
  - Numerical vector representations of text where similar words/documents → similar vectors

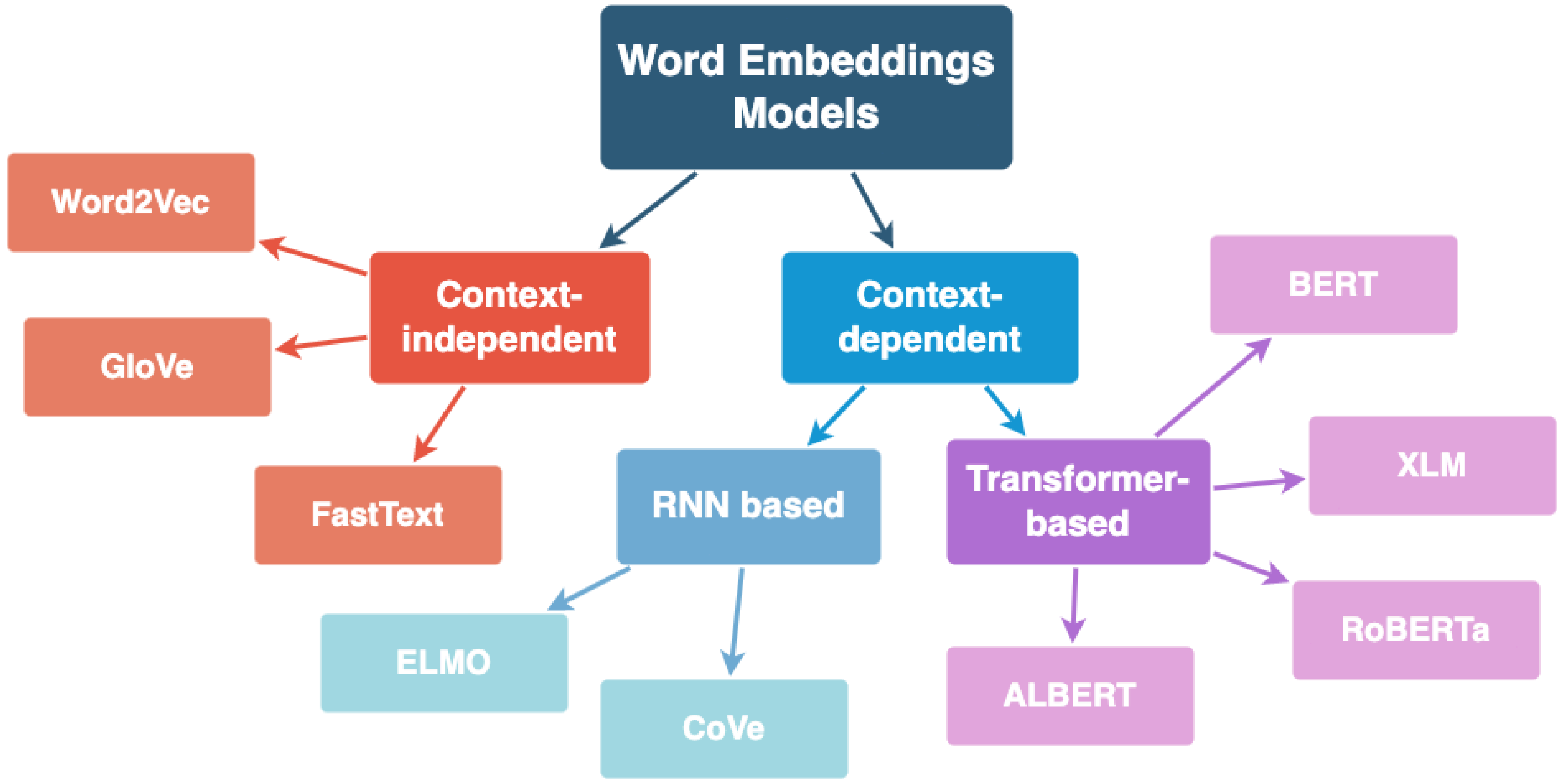
- **Why we need word embeddings in job post–resume matching ?**
  - Enable semantic understanding beyond keyword matching
  - Word embeddings capture:
    - Role similarity
    - Skill relationships
    - Contextual meaning

- **World embeddings models can be divided based context dependency.**



Male-Female

Verb tense



Word Embedding Models Tested

01 Word2Vec (Skip-gram)

Description:
Neural embedding model using token co-occurrence to learn word vectors

- Captures basic semantic relations (e.g., “analyst” ~ “consultant”)
- However, no contextual understanding due to most word vectors loses sentence structure

02 SBERT (Sentence Bert)

Description:
Transformer-based model producing sentence- and document-level embeddings

- Strong contextual understanding
- Handles long job/resume text accurately
- Robust to synonyms & variation in writing style

03 CareerBERT (Pre-trained model)

Description:
Pre-trained model for career/job similarity.

- Tailored for occupation matching
- Domain-specific vocabulary understanding
- Less stable similarity output in our testing

Final Word Embedding Model: SBERT (Sentence Bert)

SBERT consistently produced more accurate and stable rankings compared to Word2Vec and CareerBERT, especially for resumes with varied experience levels.

- **Model version used: all-MiniLM-L6-v2**
- **Dimension tested : 300 , 384, 768, and 1536.**

Word Embedding

Modeling Pipeline

1. **Cleaned Job/Resume Text**
   - Apply light cleaning (\n, unicode, tabs → space)
   - Keep punctuation & wording to preserve context
2. **Full-Text SBERT Embeddings**
   - Convert all cleaned texts to lists
   - Encode using SBERT latest model
   - Output: dimensional dense vectors
3. **Cosine Similarity**
   - For each resume, compute similarity with all jobs
   - semantic_score = cos(resume_emb, job_emb)
   - Higher score → more semantically aligned roles
4. **Ranked Results**
   - Sort jobs by semantic score (descending)
   - Stored the score for the final matching

Results

Top Jobs	Score
Senior FP&A Analyst	0.682
Senior Data Management	0.677
Data Product Owner	0.677
Python Developer	0.661
Frontend Engineer	0.654

Topic Modeling

Topic modeling compares meaning, not just keywords by mapping each resume into shared job themes

Map Themes
Convert jobs into shared topics

Topic Modeling
Match resumes to jobs by shared meaning

Compare Meaning
Match semantic themes beyond keywords

Align Roles
Connect related titles like analyst and BI

Technique Selection

LSA

- Works directly on TF-IDF vectors
- Uses SVD to reduce dimensions

✓ Fast and Efficient for retrieval

✓ Handles short/medium texts well

✓ Reduces noises

*We also developed the model using “LDA” as alternative technique for result comparison

Topic Modeling

Modeling Pipeline

1. **Job/Resume Cleaned Text**
   Cleaned input documents with job descriptions
2. **TF-IDF Vectorization**
   Convert text to weighted term vectors (1–3 grams)
3. **TruncatedSVD**
   Project into **120**-dimensional topic space
4. **Cosine Similarity**
   Compute pairwise match scores
5. **Ranked Results**
   Top-N recommendations per resume

Results

Top Jobs	Score
Delivery & Practice Head with verification	0.997
Database Operations Analyst	0.995
Application Engineer	0.994
Director Metals & Mining Consulting - Big4	0.993
Research Analyst - Energy & Resources	0.993

Results & Evaluation

Final Matching Score Calculation

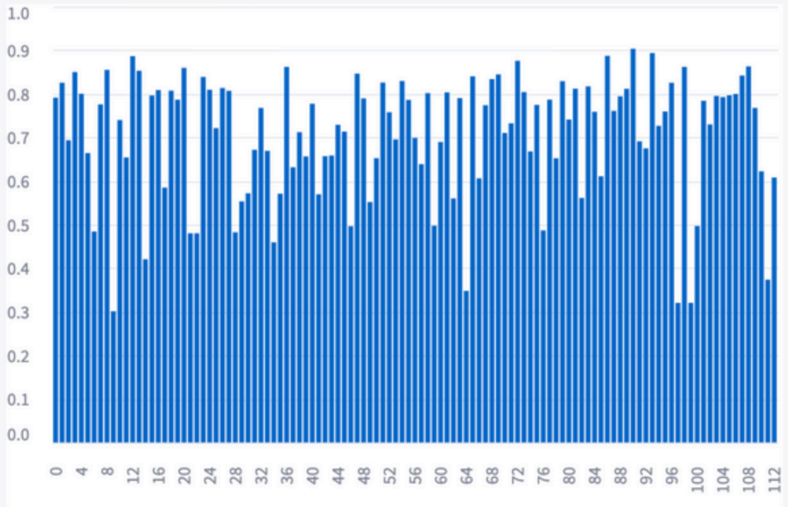
Job Matching Score Formula = avg_topic_semantic + (1 - avg_topic_semantic) * skill_score

📊 Evaluation Results

Total Resumes	Avg Final Score	Best Match	Worst Match	LLM Yes
113	0.708	0.903	0.301	51/113

Total Evaluations: 113

Score Distribution



Match Categories:

- 🟢 High Match (≥75%): 58
- 🟡 Medium Match (60-75%): 32
- 🟠 Low Match (<60%): 23

*Screenshot from our UI, showing the statistical result and visual of our similarity score

Detailed Results

index	nk	resume_file	job_rar	job_title	job_company	Final Scor	Skill Sc	Semantic	Topic S	resume_	job_skills_	matching	LLM Match
90	1	Minho_Kim_20251129_0	1	Data Engineer AWS - Hybric	TechTriad	90.32	38.46	74.42	94.12	19	17	10	❌ No
93	2	Nadia_Mahmoud_20251	1	Sr. Analyst, Digital Initiative	Toast	89.34	38.1	74.69	90.88	15	14	8	✅ Yes
86	3	Mia_Rodriguez_2025112	1	Senior Data Engineer	Envision Employn	88.7	36.67	71.17	93.15	20	21	11	✅ Yes
12	4	Ana_Martinez_20251129	1	Data Analyst	Wyndham Hotels	88.63	33.33	70.35	95.52	12	8	5	✅ Yes
72	5	Luca_Moretti_20251129_	1	Data Analyst	Helic & Co.	87.56	43.75	63.01	92.76	11	12	7	✅ Yes
108	6	Yin_Zhang_20251129_06	1	Data Analyst	Helic & Co.	86.27	31.82	68.37	91.37	17	12	7	✅ Yes
36	7	David_J_Frederickson.p	1	Procurement Executive	Rockhill Asia	86.17	26.32	77.69	84.77	13	11	5	✅ Yes
98	8	Resume Example 1.pdf	1	Procurement Executive	Rockhill Asia	86.17	26.32	77.69	84.77	13	11	5	✅ Yes
20	9	Asha_Patel_20251129_0	1	Data Engineer AWS - Hybric	TechTriad	85.94	24.24	73.73	89.14	24	17	8	✅ Yes
8	10	Amir_Boroumand.pdf	1	Sr. Full Stack Developer wit	Motion Recruitme	85.48	38.24	70.36	82.63	22	25	13	✅ Yes
13	11	Ari_Patel_20251129_063	1	Data Engineer AWS - Hybric	TechTriad	85.28	18.52	75.13	88.74	15	17	5	❌ No
3	12	Aisha_Thompson_20251	1	Data Scientist, Watchlist	RemoteHunter	84.98	33.33	59.66	95.3	17	11	7	✅ Yes
47	13	Jae_Kim_20251129_060	1	Data Analyst	Sports Research	84.62	15	74.38	89.43	13	10	3	✅ Yes
69	14	Lin_Wei_Chen_2025112	1	Data Scientist, Watchlist	RemoteHunter	84.44	28.57	64.51	91.93	16	11	6	✅ Yes
107	15	Steven_J_Vik_Incident_	1	Cyber Defense Engineer	Confidential	84.18	20	67.82	92.63	15	3	3	❌ No
65	16	Leena_Patel_20251129_	1	Sr Data Analyst with verific	Horizontal Talent	84.01	23.81	68.08	89.93	14	12	5	✅ Yes

Our Resume-Job matching model **provides the final similarity score** for each resume and **identifies the best-matching job**

LLM Evaluation & Recommandation

*GPT-OSS:20B (DECODED ONLY MODEL) USED

Prompt Engineering

system_prompt

You are an expert recruiter evaluating resume-job matches. Job descriptions often list ideal candidates with all possible skills - this is unrealistic. Focus on ESSENTIAL qualifications needed to perform core job functions. Consider transferable skills, related experience, and learning ability. A 'Yes' means the candidate can reasonably do the job, not that they have every skill listed.

prompt_template

Evaluate if this resume matches this job posting. Be practical and realistic.

JOB POSTING:
Title: {title}
Description: {job_text}

RESUME:
{resume_text}

EVALUATION APPROACH:
1. Identify 3-5 CORE skills/requirements essential for this role (ignore nice-to-haves)
2. Check if candidate has these core skills OR transferable/equivalent experience
3. Assess if candidate's experience level aligns with role expectations
4. Consider: Can they reasonably perform the core job functions? If yes → "Yes"

DECISION RULES:
1.Answer "Yes" if: Candidate has essential qualifications OR strong transferable skills that indicate they can learn/adapt
2.Answer "No" only if: Missing critical core requirements that would prevent basic job performance

OUTPUT FORMAT:
Line 1: "Yes" or "No"
Line 2: One-sentence explanation
If "No", then:
Line 3: "RECOMMENDATIONS:" followed by 3-5 specific, actionable recommendations
Line 4: "LINKEDIN_KEYWORDS:" followed by 5-10 relevant job search keywords

Evaluation & Resume Improvement Recommendation

***LLM Evaluates "Yes matching"**

LLM Evaluation:

- Match: ✅ Yes
- Reasoning: The candidate meets the essential qualifications—holding a bachelor's degree, possessing over 8 years of senior marketing experience, and demonstrating proficiency with HubSpot Marketing Hub, Salesforce, and marketing automation workflows, as well as analytics and reporting—which indicates they can effectively perform the core functions of the Marketing Operations Manager role.

***LLM Evaluates "Not matching"**

LLM Evaluation:

- Match: ❌ No
- Reasoning: The candidate lacks the required Java expertise (and JavaScript experience) essential for the core backend role, which would impede immediate contribution.

📝 **Resume Improvement Recommendations:**

• Develop proficiency in Java (e.g., build simple microservices with Spring Boot or Jakarta EE) through online courses or side projects.

• Gain working experience in JavaScript and Angular by contributing to front-end or full-stack tutorials, focusing on REST API integration.

• Highlight existing backend-like responsibilities (microservices, API design, CI/CD, Docker) and emphasize rapid learning of new languages.

• Seek certifications or short courses in Java and Angular to strengthen résumé credibility.

• Showcase transferable skills (Python, Docker, AWS, CI/CD, REST APIs, and ETL) that align with backend development tasks.

🔍 **LinkedIn Job Search Keywords:**

backend developer, data engineering, MLOps, Python, REST API, Docker, AWS, Azure, GCP, containerization, SQL, NoSQL, microservices, CI/CD, Java, JavaScript, Angular, Jakarta EE, FastAPI, Snowflake, Databricks.

```
backend developer, data engineering, MLOps, Python, REST API, Docker, AWS, Azure, GCP, containerization, SQL, NoSQL, mic
```

Evaluation Result

LLM as a Judge

- Collected 113 resumes (45 real and 68 synthetic) from peers representing a wide range of roles
- Each resume includes the user's actual current job title as ground truth for evaluation

Metrics

1. **LLM-Verified Top-1 Match Rate**
   → Percentage of resumes where the LLM answered "Yes" for the #1 recommended job.
2. **LLM-Verified Top-3 Match Rate**
   → Percentage of resumes where the LLM answered "Yes" for any job in the Top 3 recommendations.

Metric	Usage	Evaluated Match %	Logic
Top-1 Match %	Strict Accuracy	**38.34%**	Evaluates the model's ability to place the correct job at the very top of the ranking.
Top-3 Match %	Practical Usage	**41%**	Evaluates overall recommendation quality and relevant alternative quality beyond the top choice.

Findings & Conclusion

Findings

- The job-matching system that combines skill matching with semantic matching produced strong results
- From NER skill extraction reveal that job titles may vary, but the core responsibilities in the job description are usually consistent.
- LLM evaluation result always change every time we tested by nature
- SBERT's semantic score improved the relevance of Top-job matches. SBERT's also worked well to understand synonym and variation of writing style.
- Adding grams (1-3) and topics (120) helps improve the result in Topic modeling, and Using LSA runs much faster than LDA

Conclusion

- **Scalability**
  - Geographical Scale: job listings with multilingual support
  - Data Scale: Expand data collection across multiple job platforms and company sources
  - Enable real-time system integration for continuous job and resume processing.
- **Liscensing**
  - Partner with job-posting platforms and companies for more diverse job descriptions.
- **Business Value**
  - This technique can be implemented in existing ATS systems for companies, not just as a tool for job seekers.

Limitations & Opportunities for Improvement

What We Could Have Improved With More Time

- Expand and refine the skill dictionary for higher coverage
- Automate normalization and tuning for SkillScore, SemanticScore, and TopicScore
- Incorporate experience years to better reflect seniority and adjust match quality
- Expand the job post dataset with a wider range of industries to improve model generalizability.

Where the Demo/App Breaks

- Fails when job descriptions contain missing fields or extremely short/empty text
- UI may freeze for very long job descriptions or large batch uploads
- LLMs have token limitations during evaluation, which can lead to reduced performance in cases where resumes or job descriptions contain long or complex text.
- The app currently supports only PDF-format resumes, which limits flexibility for users with other file types.

UI Demo

Thank You

APAN 5430 - Group 4

Reference

Zielinski, D. (2023, December 21). Is your applicant tracking system hurting your recruiting efforts? SHRM. https://www.shrm.org/topics-tools/news/hr-magazine/applicant-tracking-system-hurting-recruiting-efforts

Cappelli, P. (2019, May 1). Your approach to hiring is all wrong. https://hbr.org/2019/05/your-approach-to-hiring-is-all-wrong