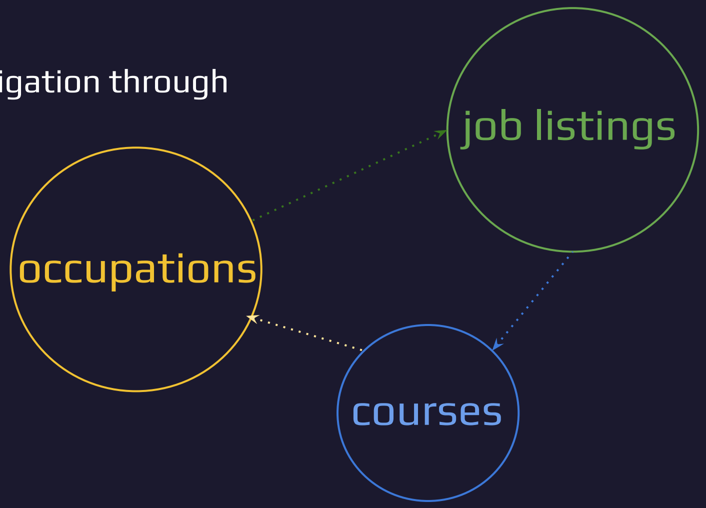
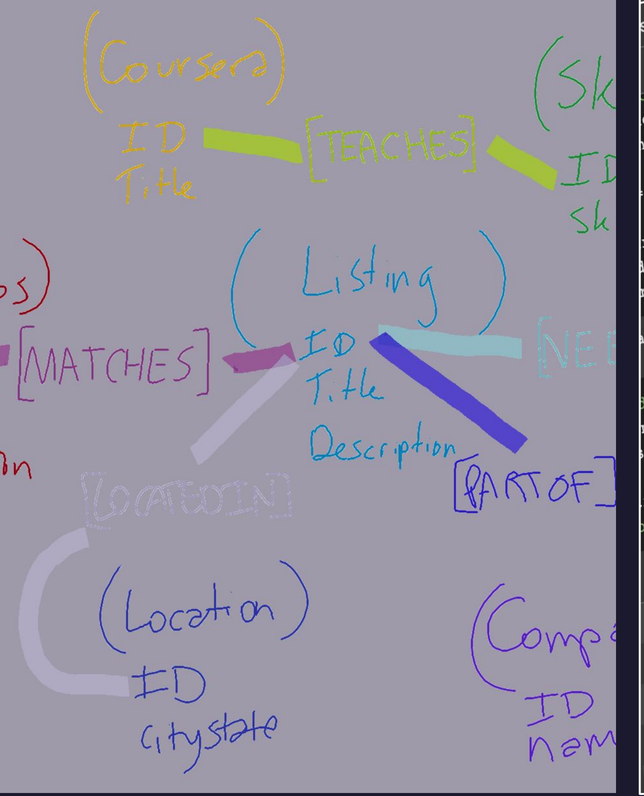
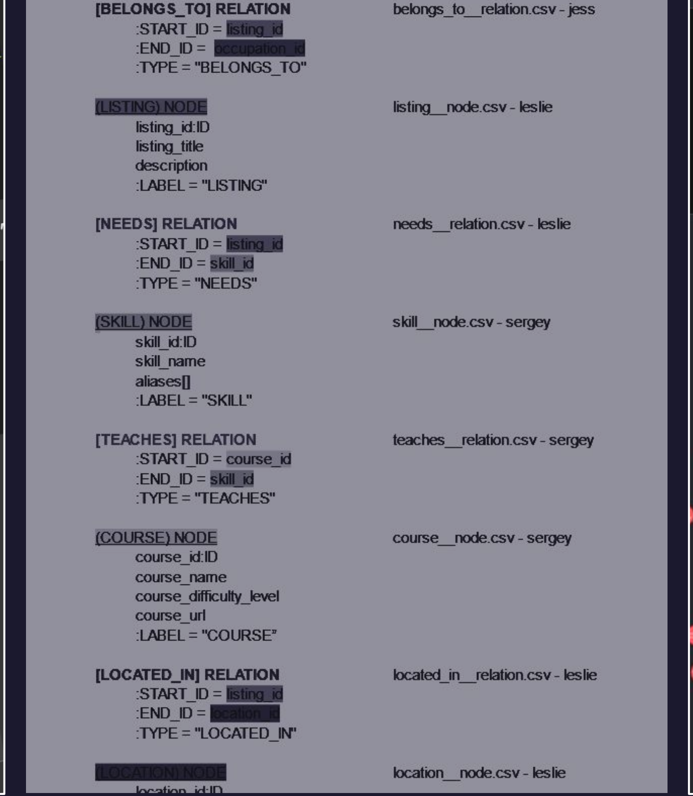
The Jobissimo Project

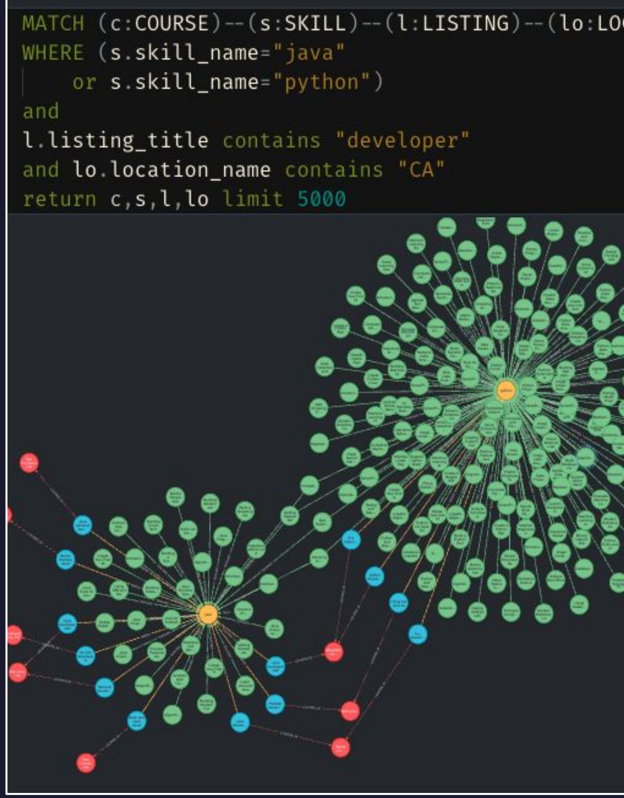
navigation through



Jessica Allan
Leslie Joe
Sergey Gurvich







Problem Statement

The average person spends 900,000 hours at work over the course of their life-time, so it stands to reason that changing career paths should be taken seriously. The skill set a person has is integral to advancing through the working world, so we set out to connect current occupations to job listings to the Coursera courses that will give you the skills to land your dream job.

What is our knowledge graph supposed to accomplish?

- Correctly connect job listings to occupations
  - Find more information about the job on the listing
    - Avg Salary
    - Career Outlook
- Correctly establish a relationship between what skills are needed for a job listing and what skills are taught through courses on Coursera
  - Which courses would best teach the majority of needed skills for a job posting
- Holistic view of jobs and skills

Dataset One: Job Listings Dataset

Dataset was of 22,000 job listings in the US from dice.com as provided by PromptCloud on Kaggle.

We kept the following data

- Job **Titles**
- Job **Descriptions**
- Job's **Company**
- Job's **Location**
- Job's list of needed **skills**

Dataset Two: Occupations

Dataset was obtained from ONET, an online database that contains all occupations in the United States.

We kept the following data

- Occupation **Title**
- Occupation **Synonyms**
- Occupation **Description**
- Expected **Career Outlook**
- Occupation **Salary**

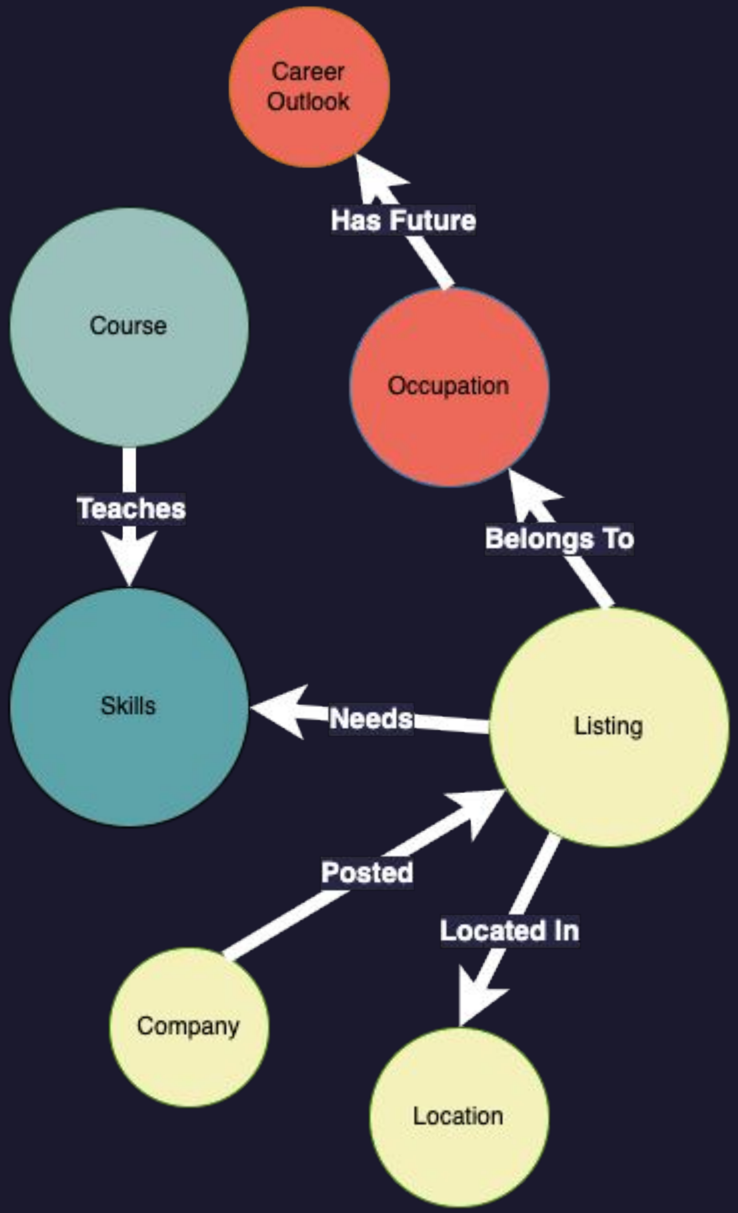
Dataset Three: Coursera Courses

Dataset was of 3,500 online classes offered by coursera.org as provided by Khushee Kapoor on Kaggle.

We kept the following data

- Course **Names**
- Course **Description**
- Course **Difficulty Levels**
- Course **URLs**
- Course’s list of taught **skills**

Our Plan: Graph Model



Cleaning: Job Listing Dataset

SKILLS:

- Removed special characters except for + and # for C++ and C##
- Kept stopwords and lower/uppercase for Spacy and Stanza
- Some "skills" were filled out as things like "read job description" and "see below" (not case sensitive)
    - Removed those skills
- Some "skills" were duplicates of the job description or full sentences (like "this job requires Python and Java knowledge")
    - Extracted most ORG and PRODUCT entities with Spacy and Stanza

DESCRIPTION

- Removed special characters
- Kept stopwords and lower/uppercase for Spacy and Stanza
- For jobs with empty skill lists, we extracted most ORG and PRODUCT entities from the description

	company_name	description	location_name	listing_title	listing_skill_name
0	Digital Intelligence Systems, LLC	Looking for Selenium engineers. must have soli...	Atlanta, GA	AUTOMATION TEST ENGINEER	[ecommerceretail qa, lan, peoplesoft, bourne s...
1	University of Chicago/IT Services	The University of Chicago has a rapidly growin...	Chicago, IL	Information Security Engineer	[systems administration, network monitoring, i...
2	Galaxy Systems, Inc.	GalaxE.SolutionsEvery day, our solutions affec...	Schaumburg, IL	Business Solutions Architect	[business inteligence, enterprise solutions ar...

Cleaning: Coursera Dataset

- Needed to clean any special and foreign language characters from Course Name, Course Description and original Skills columns

- The original Skills columns was in the format: "Skill_1   Skill_2   Skill_3". Converted this column to be as type 'list' to be able to append the skills extracted from other columns (next step).

	course_name	course_description	course_difficulty_level	course_url	course_skills
0	Write A Feature Length Screenplay For Film Or ...	Write a Full Length Feature Film Script In th...	Beginner	https://www.coursera.org/learn/write-a-feature...	[Drama, Comedy, peering, screenwriting, film, ...
1	Business Strategy Business Model Canvas Analys...	By the end of this guided project, you will be...	Beginner	https://www.coursera.org/learn/canvas-analysis...	[Finance, business plan, persona user experien...
2	Silicon Thin Film Solar Cells	This course consists of a general presentation...	Advanced	https://www.coursera.org/learn/silicon-thin-fi...	[chemistry, physics, Solar Energy, film, lambd...

Cleaning: Occupation Dataset

- Had to request the data from the ONET API and parse the XML response
  - First API call was to get the list of all occupations
  - Second API call was to get expected career outlook for each occupation
  - Third API call was to get the expected salary for each occupation

- The occupation synonyms column was split into an array of synonyms rather than a long string by replacing the word 'and' with a comma, and then splitting the string on commas
  - ['Business Intelligence Analyst', 'Competitive Intelligence Analyst', 'Data Analyst', 'Intelligence Analyst', 'Market Intelligence Analyst', 'Market Intelligence Consultant', 'Strategic Business and Technology Intelligence Consultant', 'Strategist']

- Besides missing data, the returned data was pretty clean

	occupation_id:ID	job_code	occupation_title	occupation_synonyms	occupation_description	occupation_growth	occupation_salary
0	0	13-2011.00	Accountants and Auditors	[Accountant, Accounting Officer, Audit Partner...	Examine, analyze, and interpret accounting rec...	Bright	77250
1	1	27-2011.00	Actors	[Actor, Actress, Comedian, Comic, Community Th...	Play parts in stage, television, radio, video,...	Bright	
2	2	15-2011.00	Actuaries	[Actuarial Analyst, Actuarial Associate, Actua...	Analyze statistical data, such as mortality, a...	Bright	105900
3	3	29-1291.00	Acupuncturists	[Acupuncture Physician, Acupuncture Provider, ...	Diagnose, treat, and prevent disorders by stim...	Average	60570
4	4	29-1141.01	Acute Care Nurses	[Cardiac Interventional Care Nurse, Charge Nur...	Provide advanced nursing care for patients wit...	Bright	77600

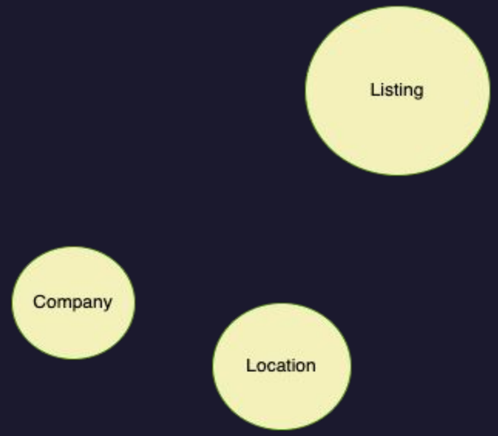
Creating Nodes

- Nodes were created out of the cleaned datasets
- Our Jupyter Notebooks were shared in a google drive using Colab
- The notebooks create an individual CSV for each node type
  - Stored into the "Output Dataset" folder
  - We followed a predetermined schema to keep names consistent

<u>(NAME OF NODE) NODE</u>
node_id:ID
node_title
attribute_1
atrribute_2
:LABEL = "NAME OF NODE"

Creating Nodes: Listing, Location & Company

- From the Listings dataset
    - Created the LISTING nodes out of the job's titles and descriptions as the only properties
    - Created the COMPANY nodes out of just the company names
    - Created the LOCATION node out of just the location name, as saved in the original Kaggle dataset as "City, State_Abbreviation"
    - The original Skills column was not included, as it would be used to create SKILLS node (further step)



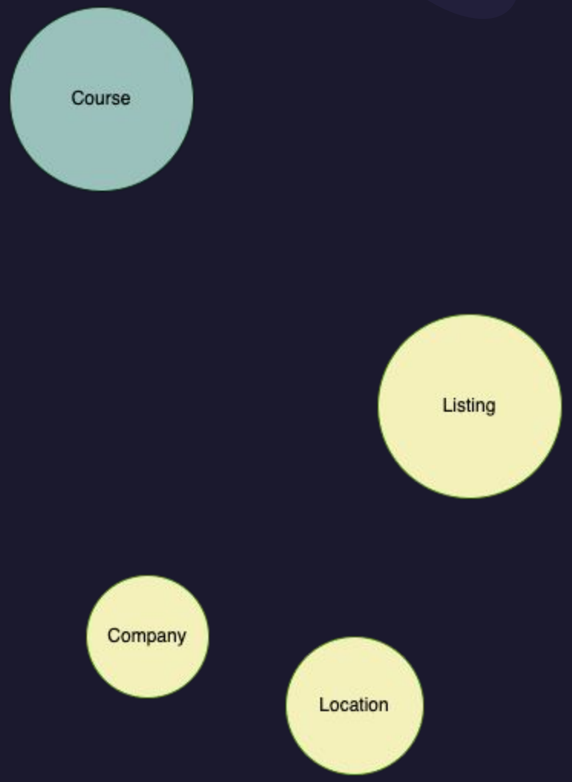
<u>(LISTING) NODE</u>
listing_id:ID
listing_title
description
:LABEL = "LISTING"

<u>(COMPANY) NODE</u>
company_id:ID
company_name
:LABEL = "COMPANY"

<u>(LOCATION) NODE</u>
location_id:ID
location_name
:LABEL = "LOCATION"

Creating Nodes: Course

- From the Courses dataset
  - Created the COURSE node that will have Course Name, Difficulty Level and Course URL as the node properties
  - The original Skills column was not included, as it would be used to create SKILLS node (further step)



<u>(COURSE) NODE</u>
course_id:ID
course_name
course_difficulty_level
course_url
:LABEL = "COURSE"

Creating Nodes: Occupation & Career Outlook

- From the Occupation dataset
  - Created the career outlook node by finding unique career outlooks and giving them an ID
  - Remaining columns were kept as attributes for the Occupation Node

<u>(OCCUPATION) NODE</u>
occupation_id:ID
occupation_title
onet_code
occupation_synonyms
occupation_description
occupation_salary
:LABEL =
"OCCUPATION"

<u>(CAREER_OUTLOOK) NODE</u>
career_outlook_id:ID
career_outlook
:LABEL = "CAREER_OUTLOOK"

Creating Nodes: Skills

Extracting Course Skills for further matching with Listing Skills:

- Step 1: Extracted skills from Course Description using SPACY and STANZA together: ORG and PRODUCT: entities
- Step 2: Extracted skills from Course Title using SPACY and STANZA together: ORG and PRODUCT entities
- Step 3: Appended results of Step 1 and Step 2 to the list of originally provided Skills and removed duplicates
- Step 4: 'Exploded' the updated Skills column to have rows Course : Skill (one skill per row)

	course_id	course_name	course_skill
231	16	Python Programming Essentials	python programming
234	16	Python Programming Essentials	python
240	16	Python Programming Essentials	python syntax and semantics
545	40	Realtime OCR and Text Detection with Tensorflo...	python programming
769	57	Prediction and Control with Function Approxima...	python
...	...	...	...
48180	3512	Mining Data to Extract and Visualize Insights ...	python programming
48184	3512	Mining Data to Extract and Visualize Insights ...	python
48263	3517	Capstone Retrieving, Processing, and Visualizi...	python programming
48266	3517	Capstone Retrieving, Processing, and Visualizi...	python
48277	3517	Capstone Retrieving, Processing, and Visualizi...	python syntax and semantics

Creating Nodes: Skills

Extracting Course Skills for further matching with Listing Skills:

- Step 5: Removed duplicated skills (lower-cased)
- Step 6: Save into ‘intermediate’ table ‘Course Skills’ for further matching

	course_id	course_name	course_skill
231	16	Python Programming Essentials	python programming
234	16	Python Programming Essentials	python
240	16	Python Programming Essentials	python syntax and semantics
545	40	Realtime OCR and Text Detection with Tensorflo...	python programming
769	57	Prediction and Control with Function Approxima...	python
...	...	...	...
48180	3512	Mining Data to Extract and Visualize Insights ...	python programming
48184	3512	Mining Data to Extract and Visualize Insights ...	python
48263	3517	Capstone Retrieving, Processing, and Visualizi...	python programming
48266	3517	Capstone Retrieving, Processing, and Visualizi...	python
48277	3517	Capstone Retrieving, Processing, and Visualizi...	python syntax and semantics

	course_skill_id	course_skill_name
10	10	comedy
11	11	screenwriting
12	12	trelby
13	13	httpsvimeo.combbdc
14	14	active learning
15	15	drama
16	16	learner review
17	17	experiential learning active learning
18	18	ip
19	19	product development

Creating Nodes: Skills

Extracting Course Skills for further matching with Listing Skills:

- Step 7: Created 'intermediate' table to map Courses Skills to Courses that they were extracted from:

	course_id	course_name	course_skill
231	16	Python Programming Essentials	python programming
234	16	Python Programming Essentials	python
240	16	Python Programming Essentials	python syntax and semantics
545	40	Realtime OCR and Text Detection with Tensorflo...	python programming
769	57	Prediction and Control with Function Approxima...	python
...	...	...	...
48180	3512	Mining Data to Extract and Visualize Insights ...	python programming
48184	3512	Mining Data to Extract and Visualize Insights ...	python
48263	3517	Capstone Retrieving, Processing, and Visualizi...	python programming
48266	3517	Capstone Retrieving, Processing, and Visualizi...	python
48277	3517	Capstone Retrieving, Processing, and Visualizi...	python syntax and semantics

	course_skill_id	course_skill_name
10	10	comedy
11	11	screenwriting
12	12	trelby
13	13	httpsvimeo.combbdc
14	14	active learning
15	15	drama
16	16	learner review
17	17	experiential learning active learning
18	18	ip
19	19	product development

	course_id	course_skill_id
10	0	1
11	1481	1
12	0	2
13	206	2
14	292	2
15	339	2
16	445	2
17	535	2
18	1067	2
19	1213	2

Now Course Skills is ready for matching with Listing Skills

Creating Nodes: Skills

Extracting Listing Skills for further matching with Course Skills:

- Step 1: If the original Skills column contained sentences instead list of skills, extracted skills from these sentences using SPACY and STANZA together: ORG and PRODUCT: entities
- Step 2: If after Step 1, we still didn't have any valid skills in the Skills column, applied skills extraction from Listing Description using SPACY and STANZA together: ORG and PRODUCT entities
- Step 3: Now we have a complete list of skills for each Listing, remove duplicates from each list
- Step 4: 'Exploded' the updated Skills column to have rows Listing : Skill (one skill per row)

	listing_title	listing_skill_name
9	Application Support Engineer	qa
9	Application Support Engineer	syfy
9	Application Support Engineer	apps for windows mobile
9	Application Support Engineer	webbased
9	Application Support Engineer	groovy
...	...	...
18	Sr. Systems Test Engineer (PERM)	linux
18	Sr. Systems Test Engineer (PERM)	windows
18	Sr. Systems Test Engineer (PERM)	java
18	Sr. Systems Test Engineer (PERM)	c c++
18	Sr. Systems Test Engineer (PERM)	load performance testing

Creating Nodes: Skills

Extracting Listing Skills for further matching with Course Skills:

- Step 5: Removed duplicated skills (lower-cased)
- Step 6: Save into 'intermediate' table 'Listing Skills' for further matching

	listing_title	listing_skill_name
9	Application Support Engineer	qa
9	Application Support Engineer	syfy
9	Application Support Engineer	apps for windows mobile
9	Application Support Engineer	webbased
9	Application Support Engineer	groovy
...	...	...
18	Sr. Systems Test Engineer (PERM)	linux
18	Sr. Systems Test Engineer (PERM)	windows
18	Sr. Systems Test Engineer (PERM)	java
18	Sr. Systems Test Engineer (PERM)	c c++
18	Sr. Systems Test Engineer (PERM)	load performance testing

	listing_skill_id	listing_skill_name
119	119	ios
120	120	androids
121	121	comcast
122	122	flash
123	123	summaryour
124	124	windows
125	125	oim
126	126	iam
127	127	scripting knowledge
128	128	oss

Now Listing Skills is ready to meld with Course Skills

Creating Nodes: Skills

Matching Listing Skills with Course Skills:

- Step 1: Matching Listing Skills to Course Skills using 3-Gram Tokenizer and Jaccard Similarity Measure with Threshold=0.6 (pruning was applied for performance, because skills names lengths vary a lot):

	listing_skill_id	listing_skill_name
119	119	ios
120	120	androids
121	121	comcast
122	122	flash
123	123	summaryour
124	124	windows
125	125	oim
126	126	iam
127	127	scripting knowledge
128	128	oss

	course_skill_id	course_skill_name
1410	1410	mainstreaming
1411	1411	demos
1412	1412	microcontroller computerscience softwaredevelo...
1413	1413	tensorflow lite
1414	1414	android
1415	1415	mobile operating systems
1416	1416	basic network security analysis
1417	1417	telnet
1418	1418	radius
1419	1419	wireshark

Creating Nodes: Skills

Matching Listing Skills with Course Skills:

- Step 2: After matching, we want to preserve all matched Course Skills, so we make a list of matched skills as an attribute and include all 'aliases' there.

	listing_skill_id	listing_skill_name	course_skill_names	course_skill_ids
0	1	lan	[lan]	[7652]
0	6	selenium	[selenium]	[70]
0	7	unix	[unix]	[1720]
0	10	development	[developmental, redevelopment, drug development]	[4405, 14793, 196]
0	11	relational databases	[relational database, relational database syst...	[936, 3409]

Creating Nodes: Skills

Matching Listing Skills with Course Skills:

- Step 3: Create SKILL node table:

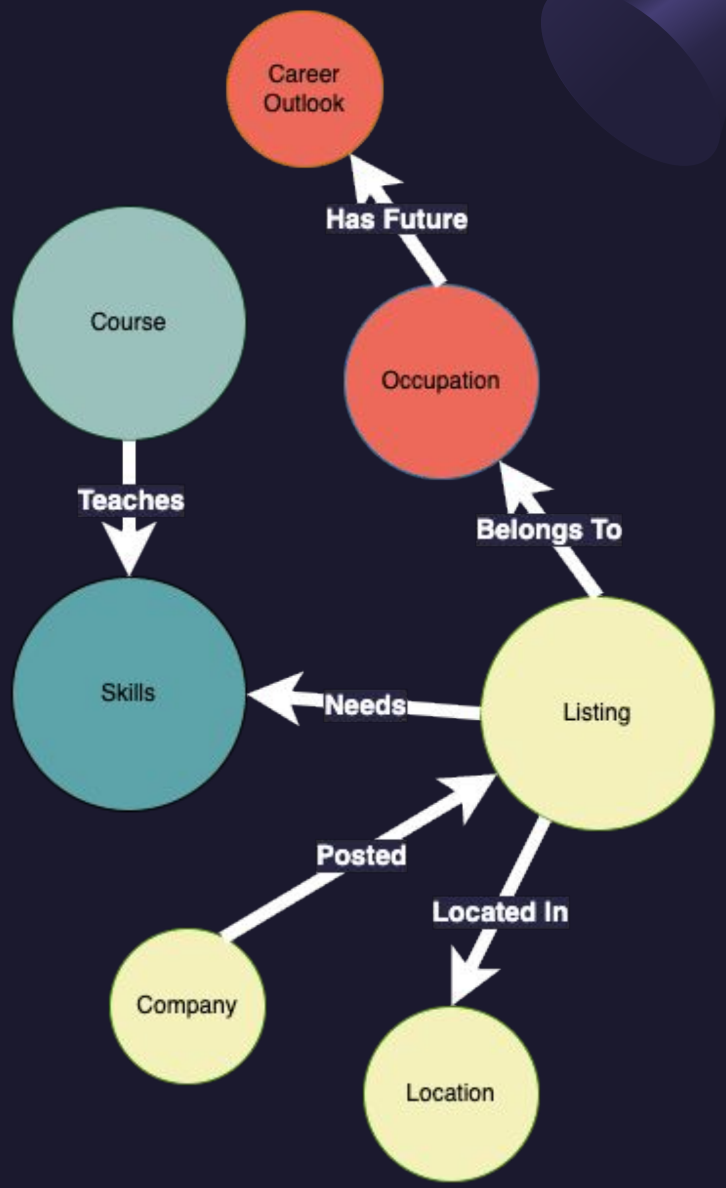
	skill_id:ID	skill_name	aliases[]	:LABEL
119	119	ios	ios	SKILL
120	120	androids	androids	SKILL
121	121	comcast	comcast	SKILL
122	122	flash	flash	SKILL
123	123	summaryour	summaryour	SKILL
124	124	windows	windows os;windows	SKILL
125	125	oim	oim	SKILL
126	126	iam	iam	SKILL
127	127	scripting knowledge	scripting knowledge	SKILL
128	128	oss	oss	SKILL
129	129	sales engineer	sales engineer	SKILL

<u>(SKILL) NODE</u>
skill_id:ID
skill_name
aliases[]
:LABEL = "SKILL"

Creating Relations

- The notebooks create an individual CSV for each relation type
  - Also stored into the "Output Dataset" folder
  - We followed a predetermined schema to keep names consistent

<u>(NAME OF RELATION) RELATION</u>
:START_ID = start_node_id
:END_ID = end_node_id
:TYPE = "NAME_OF_RELATION"



Creating relation: Belongs To

- Steps to match Listing Title and Occupation
  - Turn everything lowercase and remove certain characters (leave in +)
  - Occupation Title often had two titles in one
    - “Zoologist and Wildlife Biologists”
    - Split into two rows, “Zoologist” and “Wildlife Biologists” with same ID

	occupation_id	onet_code	occupation_title
1013	1013	43-9022.00	Word Processors and Typists
1014	1014	27-3043.00	Writers and Authors
1015	1015	19-1023.00	Zoologists and Wildlife Biologists

	occupation_id	occupation_title
1014	1014	authors
1015	1015	zoologists
1015	1015	wildlife biologists

  - Explode the occupation_synonym array and have the synonyms also represent occupation titles and appended the results

	occupation_id	onet_code	occupation_title	occupation_synonyms
1013	1013	43-9022.00	Word Processors and Typists	['Clerk Specialist', 'Clerk Typist', 'Keyboard...
1014	1014	27-3043.00	Writers and Authors	['Advertisement Agency Copywriter (Ad Agency C...
1015	1015	19-1023.00	Zoologists and Wildlife Biologists	['Aquatic Biologist', 'Conservation Resources ...

	occupation_id	occupation_title
9743	1015	conservation resources management biologist
9744	1015	fish and wildlife biologist
9745	1015	fisheries biologist
9746	1015	fisheries management biologist
9747	1015	habitat biologist
9748	1015	migratory game bird biologist
9749	1015	wildlife biologist
9750	1015	zoologist

Creating relation: Belongs To

- For every listing title, loop through every occupation title to find the highest 3-gram Jaccard similarity score
  - Helped account for miss-spellings or slight changes such as Software Engineer 3 vs Software Engineer
  - White space and other n-grams were not as successful
    - relatively short titles

	occupation_id	occupation_title	listing_title	listing_id	jaccard_3_gram_score
70	596.0	design engineer (design eng)	senior devops engineer (contract)	70.0	0.234043
71	628.0	planning engineer	capacity planning engineer - 11350	71.0	0.400000
72	244.0	information architect	data center virtualization architect	72.0	0.319149
73	575.0	management analyst	sr. information risk management analyst	73.0	0.428571
74	417.0	grant coordinator	account coordinator ii	74.0	0.375000
75	563.0	supply chain analyst	technical lead supply chain - 12241	75.0	0.250000
76	889.0	software developer	c++ software developer for multi-asset risk sy...	76.0	0.303571
77	261.0	senior adults director	senior mysql dba	77.0	0.131579
78	191.0	network engineer	manager of is network engineers	78.0	0.405405
79	599.0	program manager	(us)-program manager senior	79.0	0.424242

Creating relation: Belongs To

- If the highest similarity score for the listing title was < 0.3, then the cleaned listing skills were used to generate a request from the ONET API to find the most similar occupation match

	occupation_id	occupation_title	listing_title	listing_id
3	889.0	Software Developers	java developer (mid level)- ft- great culture,...	3.0
15	889.0	Software Developers	java architect - denver, co - fulltime	15.0
32	889.0	Software Developers	core java developer with distributed computing	32.0
34	1000.0	Web and Digital Interface Designers	mobile automation tester , rate :open negotiab...	34.0
44	1001.0	Web Developers	senior. net developer (temp-to-perm)	44.0
48	889.0	Software Developers	java full stack engineer (angular js is must)	48.0
54	889.0	Software Developers	sr service delivery systems administrator (dev...	54.0
57	727.0	Pharmacy Aides	scientific software specialist and ba	57.0

- Appended the two dataframes together

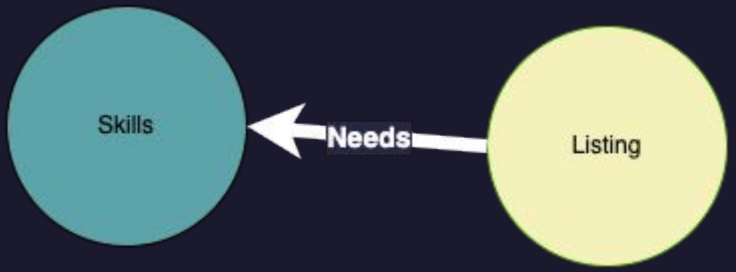
(BELONGS TO) RELATION
:START_ID = listing_id
:END_ID = occupation_id
:TYPE = "BELONGS_TO"

Creating Relation: # Has Future

(HAS_FUTURE) RELATION
:START_ID = occupation_id
:END_ID = career_outlook_id
:TYPE = "HAS_FUTURE"

Creating Relation: Needs

(NEEDS) RELATION
:START_ID = listing_id
:END_ID = skills_id
:TYPE = "NEEDS"



Creating Relation: Teaches

(TEACHES) RELATION

:START_ID = course_id

:END_ID = skills_id

:TYPE = "TEACHES"

Creating Relation: Located In & Posted

(LOCATED_IN) RELATION
:START_ID = listing_id
:END_ID = location_id
:TYPE = "LOCATED_IN"

(HAS_FUTURE) RELATION
:START_ID = company_id
:END_ID = listing_id
:TYPE = "HAS_FUTURE"

Export to Neo4j

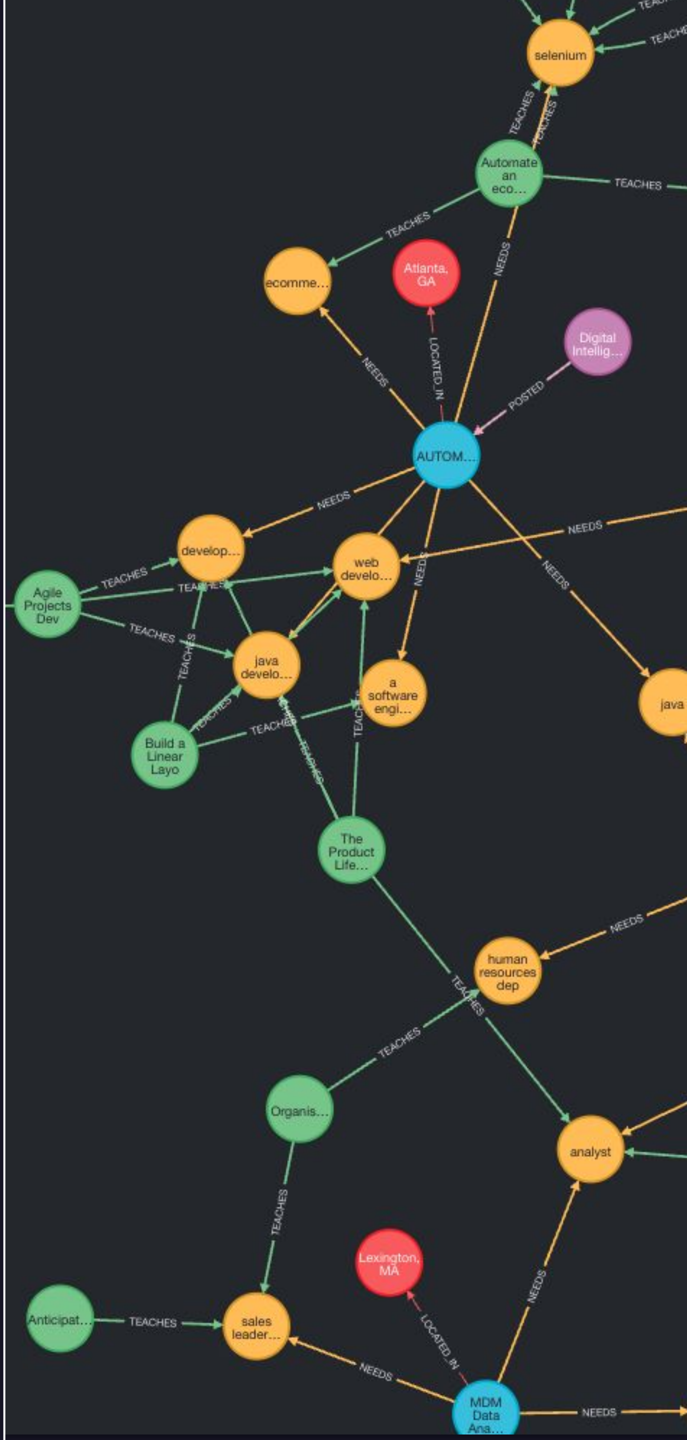
- Take all the saved Nodes and Relationship tables
- Increment :IDs of each table such that nodes' IDs never repeat
- Export to the DB using neo4j-admin command

course_id:ID	course_name	course_difficulty_level	course_url	:LABEL
0	Write A Feature Length Screenplay For Film Or ...	Beginner	https://www.coursera.org/learn/write-a-feature...	COURSE
1	Business Strategy Business Model Canvas Analys...	Beginner	https://www.coursera.org/learn/canvas-analysis...	COURSE
2	Silicon Thin Film Solar Cells	Advanced	https://www.coursera.org/learn/silicon-thin-fi...	COURSE
3	Finance for Managers	Intermediate	https://www.coursera.org/learn/operational-fin...	COURSE
4	Retrieve Data using SingleTable SQL Queries	Beginner	https://www.coursera.org/learn/single-table-sq...	COURSE
...	...	...	...	...
3517	Capstone Retrieving, Processing, and Visualizi...	Beginner	https://www.coursera.org/learn/python-data-vis...	COURSE
3518	Patrick Henry Forgotten Founder	Intermediate	https://www.coursera.org/learn/henry	COURSE
3519	Business intelligence and data analytics Gener...	Advanced	https://www.coursera.org/learn/business-intell...	COURSE
3520	Rigid Body Dynamics	Beginner	https://www.coursera.org/learn/rigid-body-dyna...	COURSE
3521	Architecting with Google Kubernetes Engine Pro...	Intermediate	https://www.coursera.org/learn/deploying-secur...	COURSE

skill_id:ID	skill_name	aliases[]	:LABEL
0	ecommerceretail qa	ecommerceretail qa	SKILL
1	lan	lan	SKILL
2	peoplesoft	peoplesoft	SKILL
3	bourne shell scripting	bourne shell scripting	SKILL
4	groovy	groovy	SKILL
...	...	...	...
29418	nosqldatabase	nosqldatabase	SKILL
29419	programmingdevelopment	programmingdevelopment;program development	SKILL
29420	programming on win xp788.1	programming on win xp788.1	SKILL
29421	skills win32 programming expertcc++ programming	skills win32 programming expertcc++ programming	SKILL
29422	cc	cc	SKILL

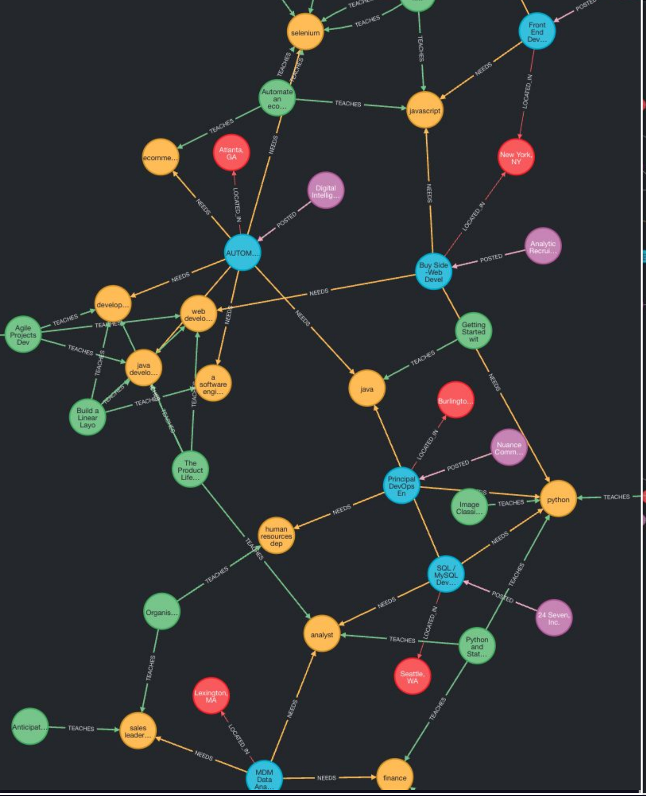
	:START_ID	:END_ID	:TYPE
0	0	20261	TEACHES
1	330	20261	TEACHES
2	1906	20261	TEACHES
3	2424	20261	TEACHES
4	2445	20261	TEACHES
...	...	...	...
37364	3516	23830	TEACHES
37365	3516	1597	TEACHES
37366	3516	25	TEACHES
37367	3516	3126	TEACHES
37368	3518	24427	TEACHES

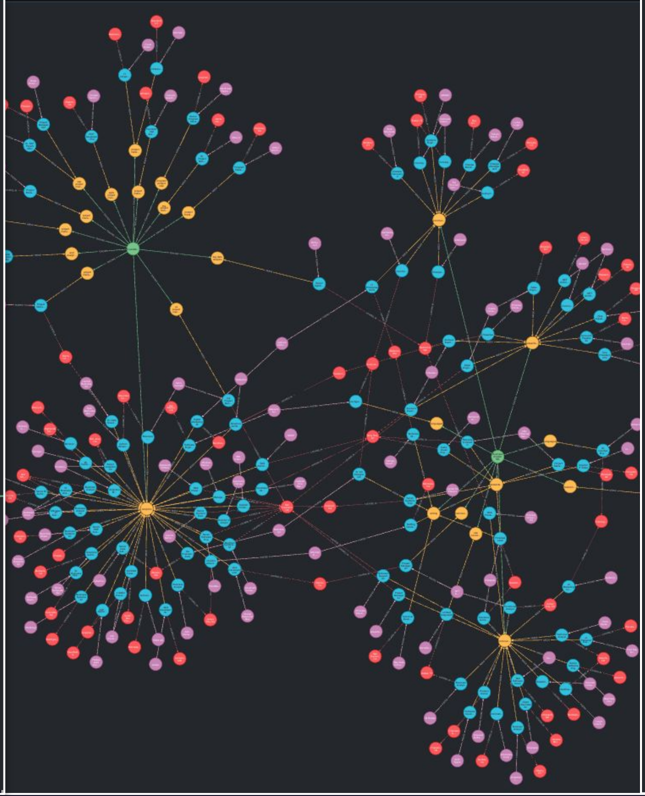
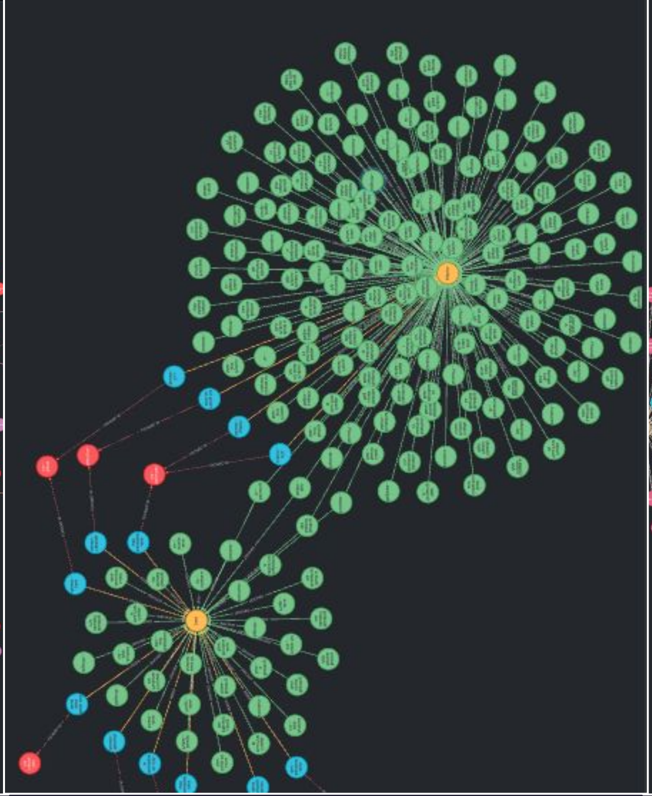
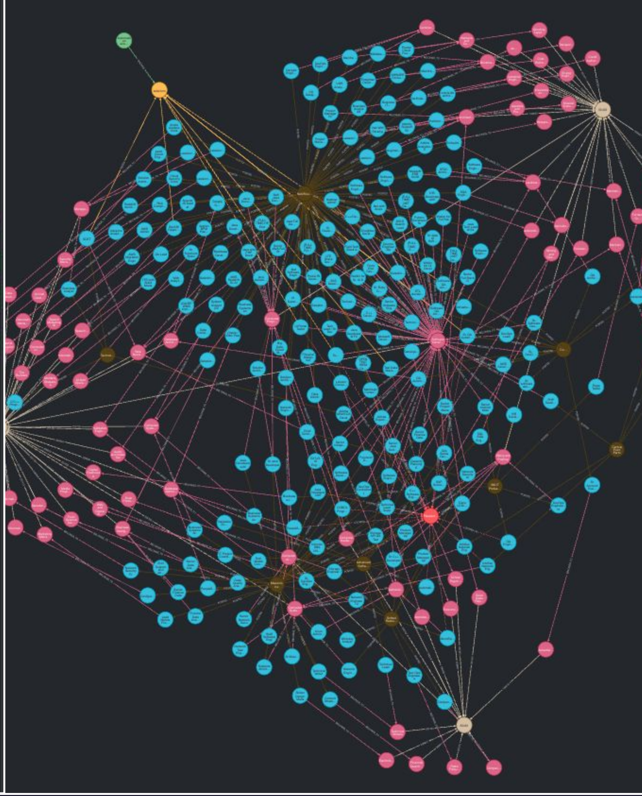
55453 nodes, 206814 relationships, 167584 properties



Tools Used

- Spacy + Stanza for extracting skills entities
- PY_stringmatching for skills and job-occupations matching
- Python Requests library to download the ONET data
- Google Colab to run and share notebooks
- ~~AWS RDS Postgres DB~~



Demo in Neo4j

Problems and Next Steps

- Pruning the skills columns more and more accurate extracting with Stanza and Spacy
- Consolidate the skills list (more efficiently combine skills like "quality assurance" and "qa") by removing the need for the aliases property with PY_stringmatching
- Use such metrics as F1 Score and Precision to quantify the string matching quality
- Acquire more information about the companies (e.g. working conditions, turnover rate, etc.), and add these as properties to each company node
- Create and implement an "APPLICANT" node to represent a person and their skillset, and find results via Neo4j catered to them
- Build more complex and sophisticated queries for analytics, especially because the real world application will be constantly updated with new job listings and courses

Thank You

Please let us know if you have any questions.