Candidate Locator

Data Analysis Project

Course #: 094290

Submission Date: 11/04/2024

By:

Majd Bishara	**323958140**
Saleem Kheer-Eldeen	**212699581**
Tameer Milhem	**322729443**
Lana Haj	**206528382**

GitHub

Table of Contents

Introduction:

In a rapidly evolving world, staying updated with the latest trends in any industry, especially in the recruiting industry, has become increasingly challenging.
Finding a suitable candidate is no longer sufficient you need one who can effortlessly adapt to future and frequently changing requirements, a candidate who will give your competitors a run for their money.
By leveraging the power of LLMs and Machine Learning, we meticulously analyze job listings from your company and others in the industry. This enables us to provide candidates who not only meet your specific requirements but also align with the industry's evolving standards.

Data Collection and Integration:

Original Data

We used both datasets, a subset of the people dataset was used for training and the output of our model, the company's dataset was used to see which meta-industry each person works in.

Additional Data:

We decided to scrap job listings from LinkedIn.
A job listing includes the company that listed the job, the description for the job and other attributes like whether its full-time and so on.
We scrapped job listings for companies that have employees in our *people* dataset, since as we mentioned before scraping is an expensive operation.
For scraping we used selenium and the proxy provided by BrightData to avoid getting blocked by LinkedIn the code is provided in the notebooks.

Additional Data Usage:

We extracted the skills from the job description (explanation in LLM section) for the job listings and then grouped them by meta industry, we then choose the top 80 of the most appearing skills, and we used them as representative skills that are required in that meta industry.

Item definition & enrichment size:

We define an item as a job listing and they are identified with *job_id*.
Since we want to use the skills to represent a meta-industry, a large amount of data would yield a better representation and ideally yield better results for our solution.
Using scraping, we managed to reach 643 items, however, some meta-industries didn't have enough items and scraping as we mentioned is expensive, we searched the web for datasets that are similar to ours and luckily we found a dataset[1] with almost the same schema as our scrapped data, from this dataset we picked job listing for companies that are in our dataset to end up with a total of ~1500 items.

[1] https://www.kaggle.com/datasets/arshkon/linkedin-job-postings

Data Analysis:

Our project revolves around skills, meta-industries, and duration of work of the employees, we will analyze them and see how they interact with other variables so we can use them in our models.

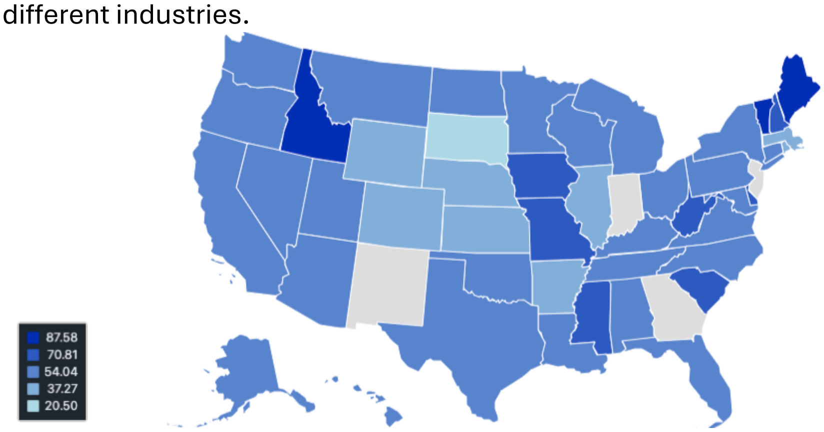
Variables:

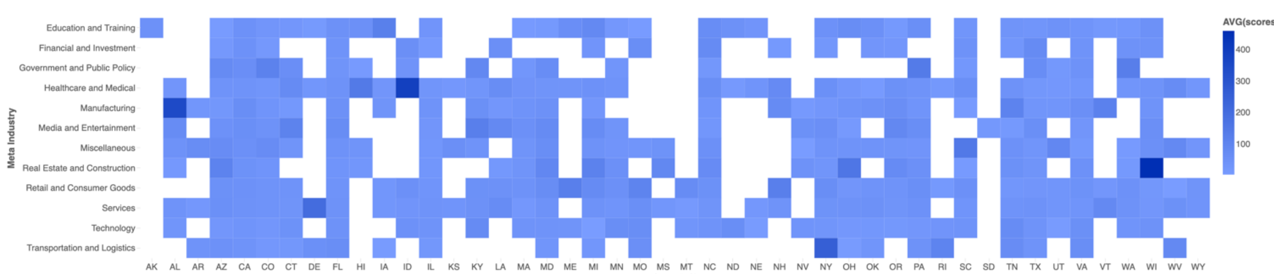
1. Location

Our data, sourced solely from the US, provides insight into the distribution of work durations for employees across different states, as well as variations in duration among meta-industries.

On average, work duration varies significantly between states. For instance, states like Idaho, Maine, and Vermont have the highest average duration of 80+ months, while South Dakota has the lowest at 20.5 months. (Interactive visualization available in the notebook.)

While there are some differences between meta-industries, the variations are relatively small. In California, the work duration appears uniform across all meta-industries. However, in certain states like Wisconsin (WI), there is a noticeable difference in duration among different industries.

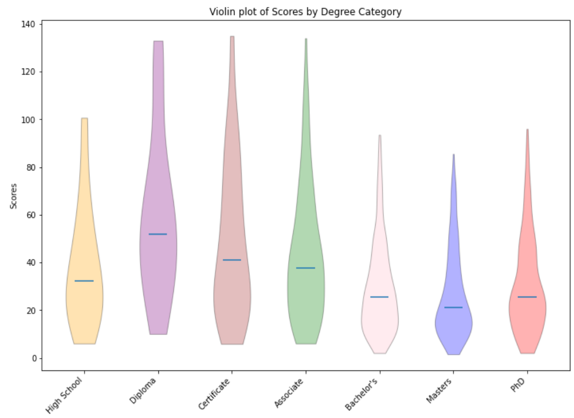




2. Degrees

After extracting the degree from the "experience" column and removing outliers, we analyzed the relationship between the employees' degree and their duration. The resulting graph shows that people with Diplomas, Certificates, and Associate's degrees have a higher duration than those with other degrees. Initially, this seemed counterintuitive since we expected individuals with higher degrees like Bachelor's and Master's to have longer durations. However, upon further examination of the data, we discovered that many individuals with Diploma, Certificates, and Associate's degrees hold stable long-term positions, such as managers and administrators.

To determine if there is statistical evidence that the different degrees are indeed different, we conducted Kruskal's test on every pair of degrees. The results will provide insights into the statistical significance of the observed differences in duration among various degrees.



Degree 1	Degree 2	Statistic	P-value	Reject?	Explanation
Associate	Bachelor's	25.11	0.00	✅	Different distribution
Associate	Certificate	0.52	0.47	❌	Similar distribution
Associate	Diploma	3.55	0.06	❌	Similar distribution
Associate	High School	0.66	0.42	❌	Similar distribution
Associate	Masters	35.59	0.00	✅	Different distribution

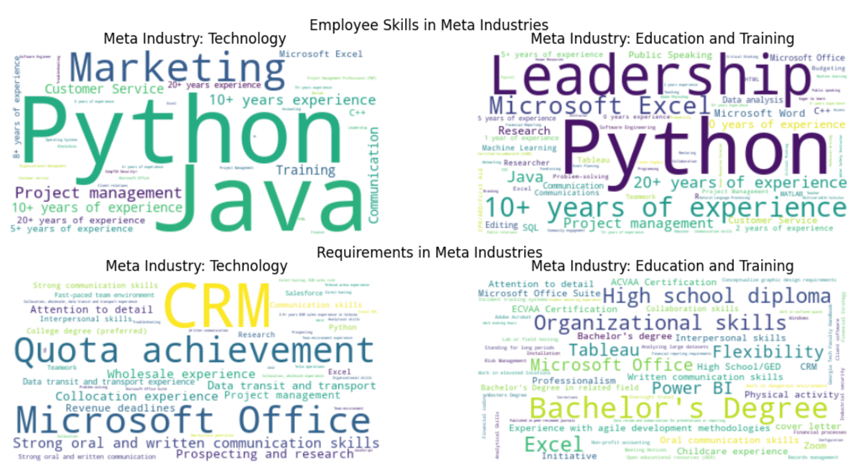
3. Skills

Skills are an integral part of our solution; we extracted the skills from job descriptions and about sections of employees using filtering and Large Language Models (LLMs) like Gemini.

To visualize the required and possessed skills, we created word clouds for each meta industry.

The results are intuitive, aligning with expectations. For instance, employees in the technology industry often possess skills like Python and Java, while manufacturing industry requirements include driving-related skills (Entire visualization).

However, some results did not meet our expectations. For example, we anticipated more programming-oriented requirements in the technology industry.



AI Methodologies:

LLM:

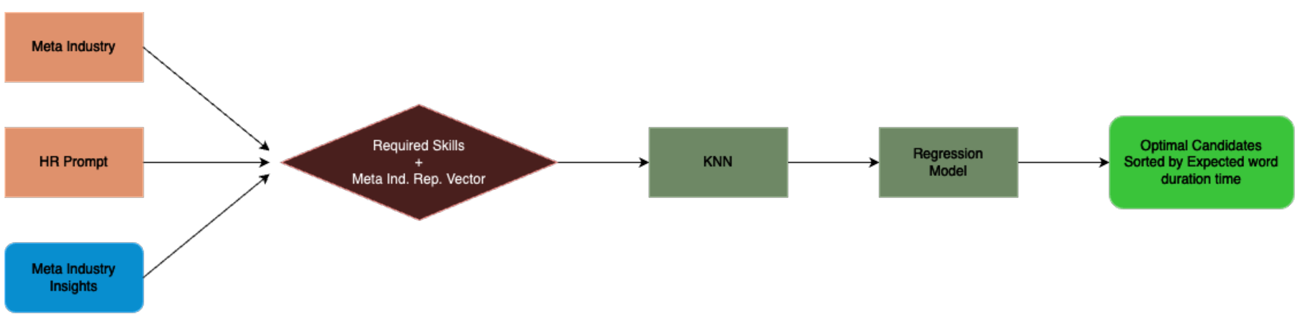
Our solution relies on many things and one of which are the skills/requirements of the employees and those representatives of the meta-industry, to extract these skills we used the Gemini API[2], Google's free LLM, in the preprocessing with specific prompts *see more at 1*
Evaluating the LLM is challenging given the scale of our operation, as it necessitates an understanding of both input and output. Nonetheless, we conducted our own evaluations by randomly selecting items. We observed that the LLM performed admirably in most cases; however, there were rare instances where it deviated from the prompt and generated random outputs. Fortunately, such occurrences were manageable.

KNN + Regression:

To find candidates that meet the prompt's requiremenets we used KNN with cosine similarity as the distance metric, and we choose top K most similar vectors, which were then fed to the regression model to predict the duration of work.

The Pipeline:

Pipline representation:



The pipeline combines the meta industry in 2 ways, first we fetch a representitve feature vector that represents the meta industry's requirements, and secondly we fetch insights about the meta industry and combine it with the HR prompt for it to be embedded in the same vector, then we sum the skills vector with the meta industry's vector to which we then apply KNN and regression to get the desired output.

Evaluation & Results:

KNN:

KNN was implemeneted with cosine similiartity, so when we query we find the K most similar candidates, since this is an unsuprvised task we calculated the mean of the similiarties to evaluate the results (tangible results and more in the 2).

[2] https://ai.google.dev/

Regression Model:

For the regression model we used the RMSE loss, we trained a RandomForestRegressor with 10 trees on 0.7 of the employees dataset, and 0.3 for validation, and the best we got is RMSE = 46.93

Limitations and Reflections:

Our project involved scraping and interfacing with large language models (LLMs). However, we faced limitations due to the time-consuming nature of scraping and the use of an API for LLM interaction, which hindered the processing of large amounts of data.

These constraints necessitated careful data selection. We employed sampling techniques and made assumptions to simplify operations.

Despite these challenges, the project provided valuable lessons in improvisation and maximizing results within constraints.

We believe that additional data and time could further enhance the outcomes. For instance, the vectors representing meta-industries would more accurately reflect their characteristics with more data.

Conclusions:

We provided recruiters an easy, rigid and time saving product that first and foremost helps them find the candidates according to their requirement while also automatically keeping up with the industry's changing requirements.
After testing the model on different prompts and meta-industries we can see that it performed very well, and it fetched the relevant candidates correctly.
We look forward to seeing the results when it's deployed in the real world, and after getting feedback from recruiters.

Appendix:

1. We used the *Steerability*[3] method where we prompt the LLM with a context/role, usually a prompt telling it what to do and what to return when given an input, and then providing it with the input, usually a field like *job_description*, example:

   Context:

   *Given a job requirement text, return a Python list containing the listed requirements for the job. Be concise and abstract, listing software names individually. List degree of applicability, DONT use the words proficiency, ability or **knowledge***

   Input:

   3-8+ years of solid, B2B sales experience in a telecom arena with collocation, wholesale experience and data transit and transport.

   Output:

   ['Telecom sales (3-8+ years)', 'wholesale', 'data transit and transport']

   Using it with conjunction of filtering and aggregating as to not overwhelm the LLM with long pieces of text, and providing it with the appropiate context managed to give great results, yet unfortunatly it also sometimes returned bad results, we used some functions to evaluate the responses, to check if the responses fit our standards, and whether we need to reprompt it, however, this didn't help improve the results drastically.

2. Example run:

Prompt: "I'm looking for an employee with many years of experience in drawing and art, especially in digital art, for an animation job"

Meta-Industry: "Technology"

Return:

id	Skills	Work_duration
trevordavisdesign	['Alumni Tutor at Animation Mentor', 'Figma, 'Unreal Engine', 'Animation Studio', 'Animation Studio II', 'Art History Pre-20th Century', 'Basic Drawing', 'Digital & Traditional Illustration Studio', 'Digital Modeling Studio', 'Illustration Studio', 'Interactive Design Studio', 'Intro to Animation Studio', 'Rapid Visualization Studio', 'Studies in Fiction Writing']	22.857
…	…	…

KNN score: 0.8884262

RMSE: 85.67

[3] https://www.latentview.com/glossary/steerability/

3. Another mention that caught our eyes is this candidate:
   Id: savannah-aaron
   Skills: ['BS in Digital Arts, Stetson University', 'Photoshop, Lightroom, Illustrator', 'Python, Java', 'Style, color and texture expertise', 'Mockups and drawings proficiency']
   Duration: 32.54

   The green highlighted skills match the prompt, and since the industry is technology, the blue highlighted text matches the type of employees that work in this industry in the sense that those employees have knowledge in programming languages.

4. We generated some prompts (provided in the notebook) for every meta industry and calculated the KNN score and got the following:
   Mean distance for prompt 'Healthcare and Medical': 0.906
   Mean distance for prompt 'Government and Public Policy': 0.901
   Mean distance for prompt 'Education and Training': 0.911
   Mean distance for prompt 'Services': 0.9208
   Mean distance for prompt 'Healthcare and Medical': 0.911
   Mean distance for prompt 'Government and Public Policy': 0.903
   Mean of all industries: 0.9107

   To clarify, given a prompt and meta-industry we run it through the model and get K candidates, we calculate the mean distances (in our case similarity), and this is the number shown above.

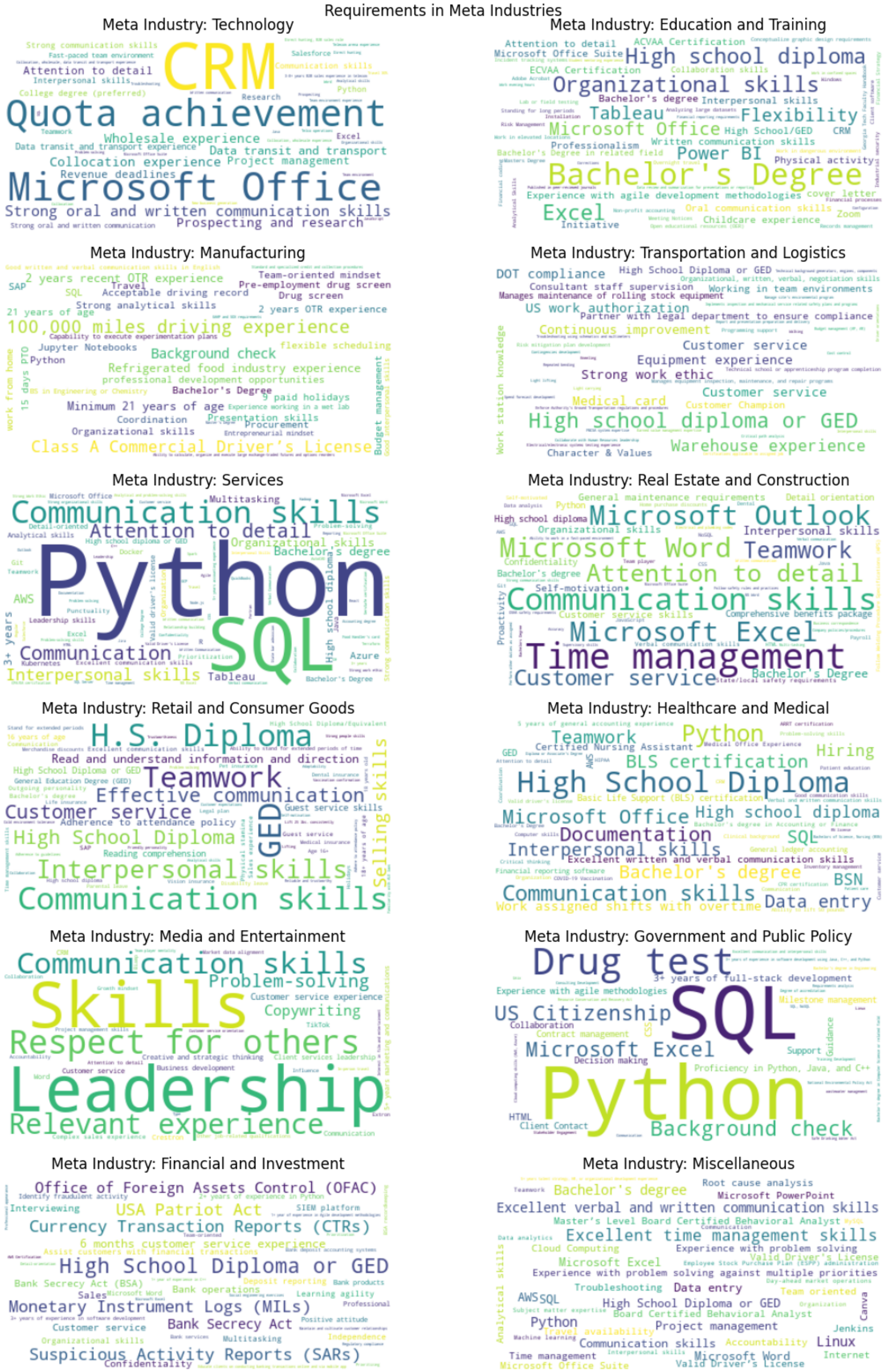
5. Job Listing's schema, relevant columns highlighted.

```
root
 |-- job_id: long (nullable = true)
 |-- company_id: double (nullable = true)
 |-- title: string (nullable = true)
 |-- description: string (nullable = true)
 |-- max_salary: double (nullable = true)
 |-- med_salary: double (nullable = true)
 |-- min_salary: double (nullable = true)
 |-- pay_period: string (nullable = true)
 |-- formatted_work_type: string (nullable = true)
 |-- location: string (nullable = true)
 |-- applies: double (nullable = true)
 |-- remote_allowed: boolean (nullable = false)
 |-- views: double (nullable = true)
 |-- job_posting_url: string (nullable = true)
 |-- formatted_experience_level: string (nullable = true)
 |-- skills_desc: string (nullable = true)
 |-- sponsored: long (nullable = true)
 |-- work_type: string (nullable = true)
 |-- currency: string (nullable = true)
 |-- compensation_type: string (nullable = true)
```

job_id	compa...	title	description	max_s...	med_s...	min_s...	pay_p...	format...	location	applies	remote...	views	job_po...	format...	skills_...	sponsored	work_
3757940104	553718	Hearing Care P...	Overview He...	null	5250	null	MONTHLY	Full-time	Little Ri...	null	false	9	https://...	Entry level	null	0	FULL_TI
3757940025	2192142	Shipping & Rec...	Metalcraft of ...	null	null	null	null	Full-time	Beaver ...	null	false	null	https://...	null	null	0	FULL_TI
3757938019	474443	Manager, Engin...	The TSUBAKI...	null	null	null	null	Full-time	Bessem...	null	false	null	https://...	null	Bachelo...	0	FULL_TI
3757938018	18213359	Cook	descriptionTi...	null	22.27	null	HOURLY	Full-time	Aliso Vi...	null	false	1	https://...	Entry level	null	0	FULL_TI
3757937095	437225	Principal Cloud...	Job Summar...	275834	null	205956	YEARLY	Full-time	United ...	null	true	null	https://...	Mid-Se...	null	0	FULL_TI
3757937037	13727	Territory Mana...	Location: Re...	null	null	null	null	Full-time	United ...	null	true	16	https://...	Mid-Se...	null	0	FULL_TI
3757937004	10515052	Auto Body Tec...	Company: Ge...	null	null	null	null	Full-time	Daytona...	null	false	1	https://...	Entry level	null	0	FULL_TI
3757936167	2915	ACME D8- Asst...	The First Assi...	null	null	null	null	Full-time	Sussex, NJ	null	false	2	https://...	Mid-Se...	null	0	FULL_TI
3757936097	18213359	Dishwasher	descriptionTi...	null	19.3	null	HOURLY	Full-time	Aliso Vi...	null	false	null	https://...	Entry level	null	0	FULL_TI
3757936026	634806	Instrumentatio...	Instrumentati...	null	null	null	null	Contract	United ...	12	true	59	https://...	Entry level	null	0	CONTRA
3757935384	232541	Power Utility Di...	Power Utility ...	null	null	null	null	Full-time	Hammo...	null	false	null	https://...	null	Educati...	0	FULL_TI

6. Skills & Requirements for meta-industries:

Requirements in Meta Industries



7. Link to Data

The following link takes you to a folder of csv files, we have the scrapped data called "job_postings", we also have csv files that were "cached" since they need time to extract.
Link

8. Embedding

To embed the skills, we used a fine-tuned embedding model, that was specifically trained on data from LinkedIn, which helped us achieve great results, a showcase of the embedding can be found in the end of the Data Analysis Notebook.
The embedding can be found here

References:

- https://www.kaggle.com/datasets/arshkon/linkedin-job-postings
- https://ai.google.dev/
- https://www.latentview.com/glossary/steerability/
- https://sparknlp.org/2023/09/15/distilbert_base_uncased_linkedin_domain_adaptation_en.html