Text Technology Project

Salary Information Collection

M.Sc Computer Science Batuhan Salmanoğlu
M.Sc Computational Linguistics Wen Wen

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

The story is ...

Name：*Naruto*

Work in New York
Middle company
Software engineer

Name：*Sakura*

Work in California
Small company
Machine Learning engineer

Name：*Sasuke*

Remote
Middle company
Data Scientist

With different backgrounds
Who will earn a higher salary?

Let's find out!

Collect:

- Scrape Webpages

Prepare

- Modify Data
- Insert to a NoSQL Database
- Query
- Encode in XML

Access

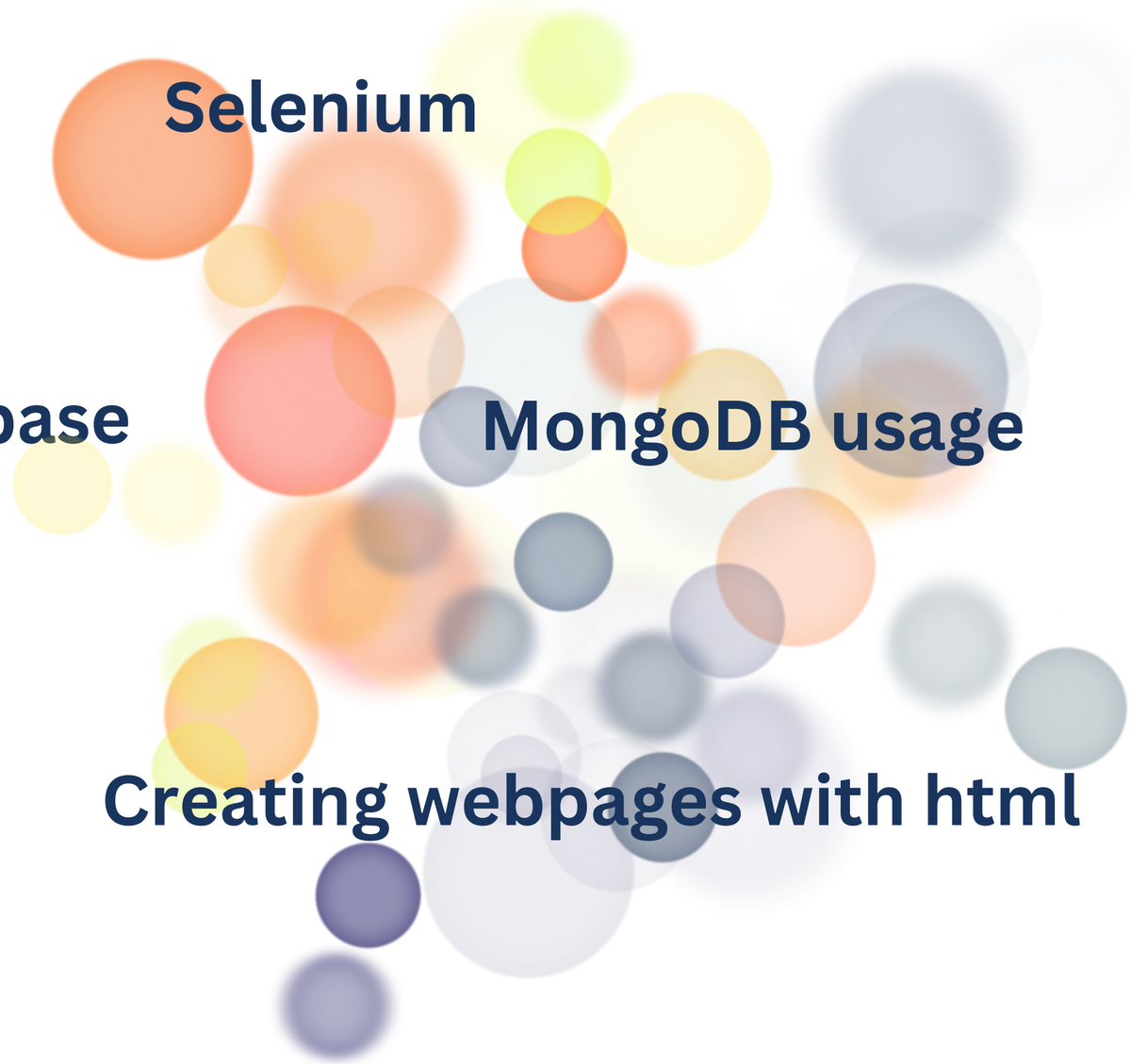
- Create XSLT file
- Present on a Webpage

Extension:

Selenium

MongoDB usage

Creating webpages with html



Collect

- **Scrape a webpage using XPath in the process, extract the result as JSON-Data**

Scrape Webpages

selenium + Xpath

https://www.glassdoor.com/

Job Salary Data

Pandas Data Frame

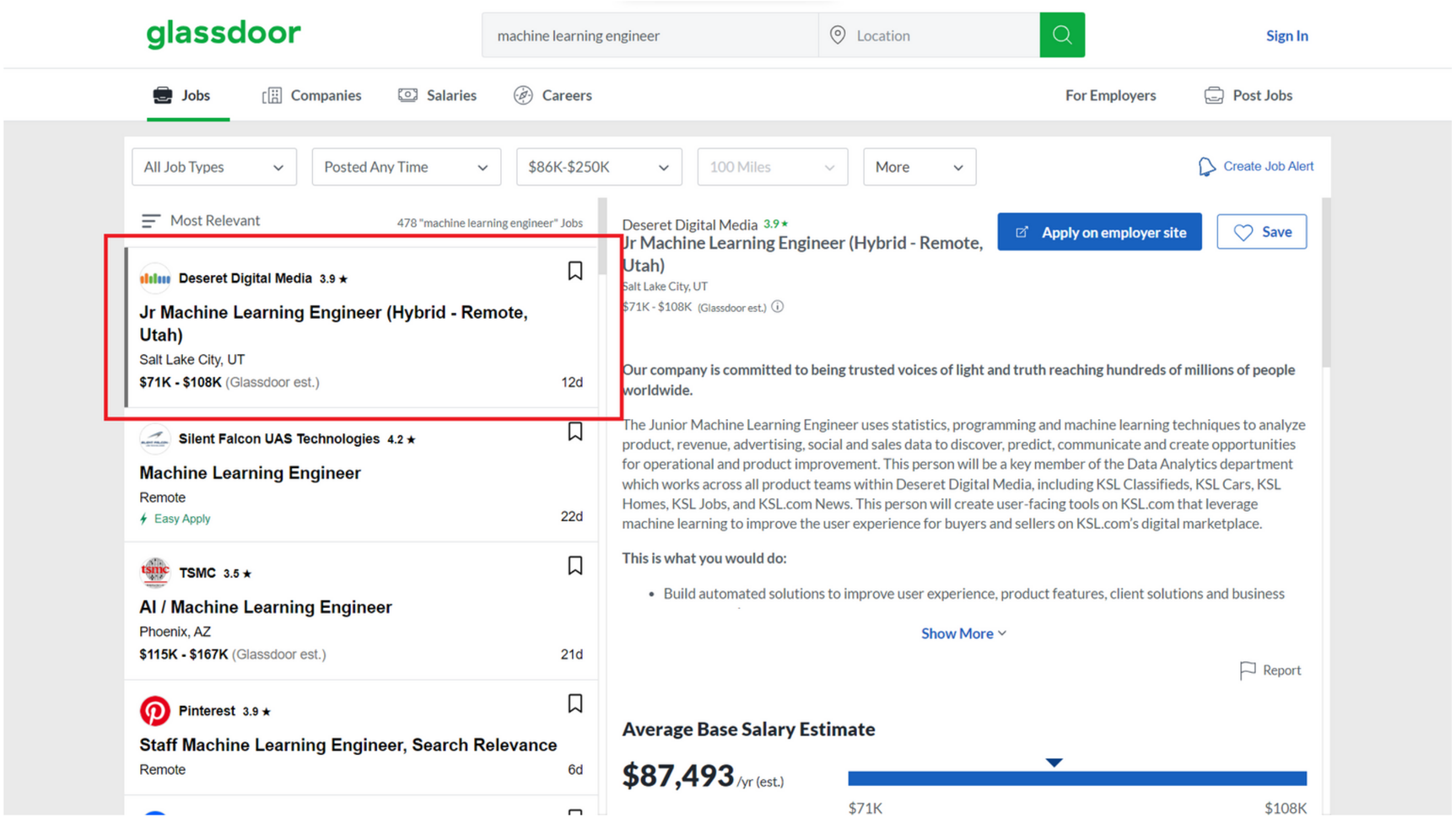
CSV File

JSON File

Our JSON Dataset is ready

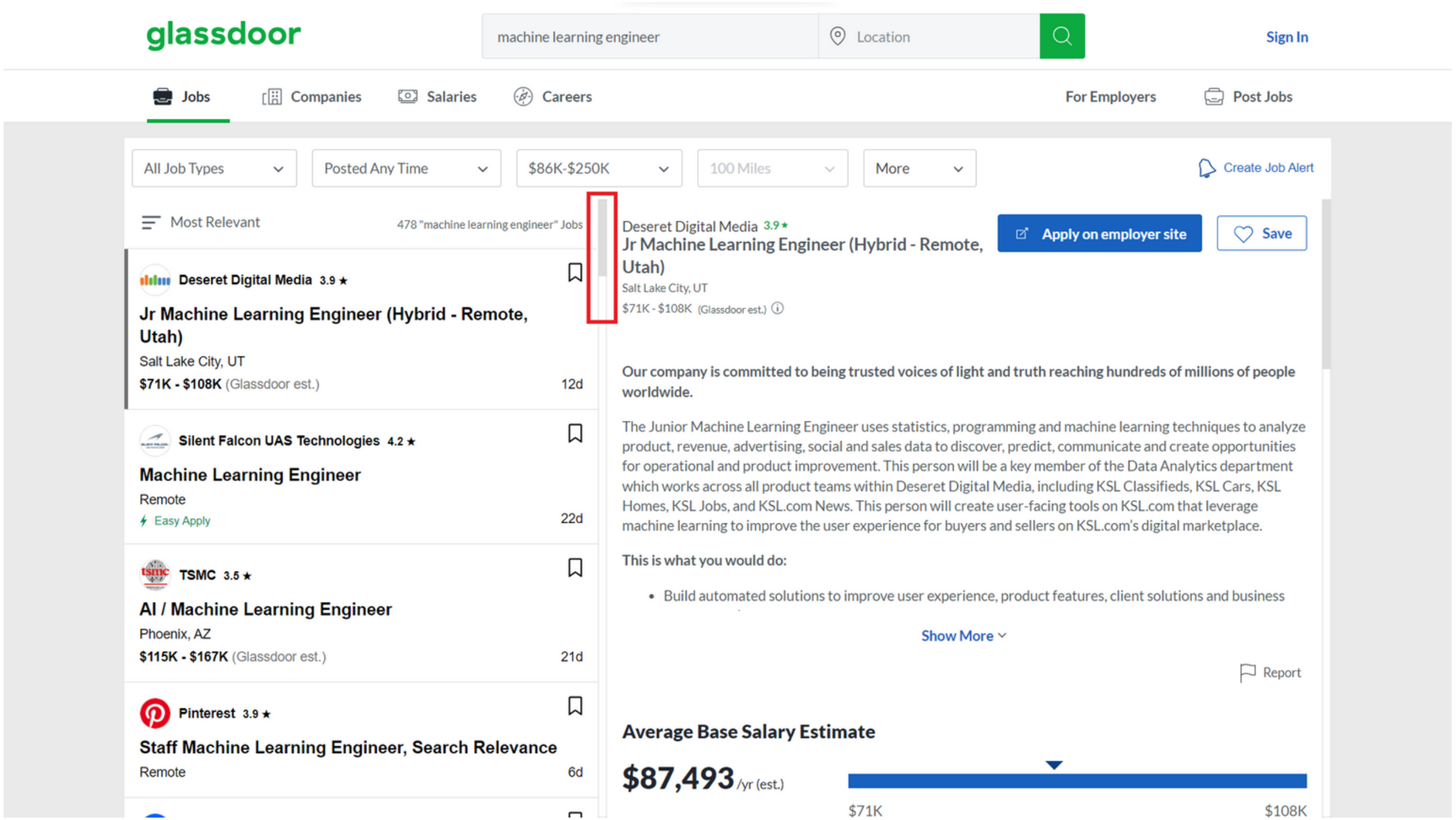
Scraping Part Step by Step

Step 1: Clicking Job card



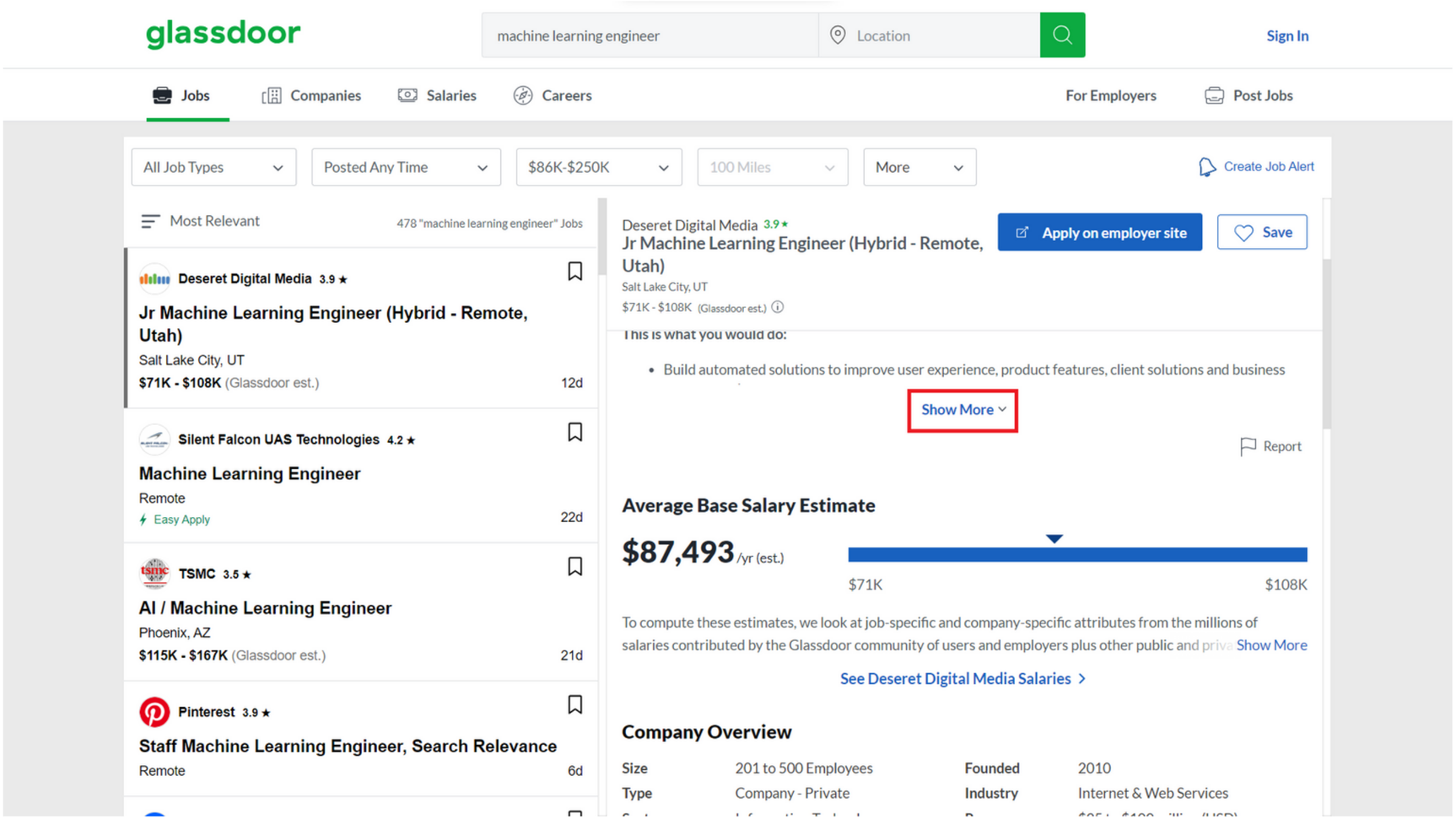

```
job_cards = driver.find_elements(By.XPATH, "//article[@id='MainCol']//ul/li[@data-adv-type='GENERAL']")
```

Step 2: Scroling down Job cards section



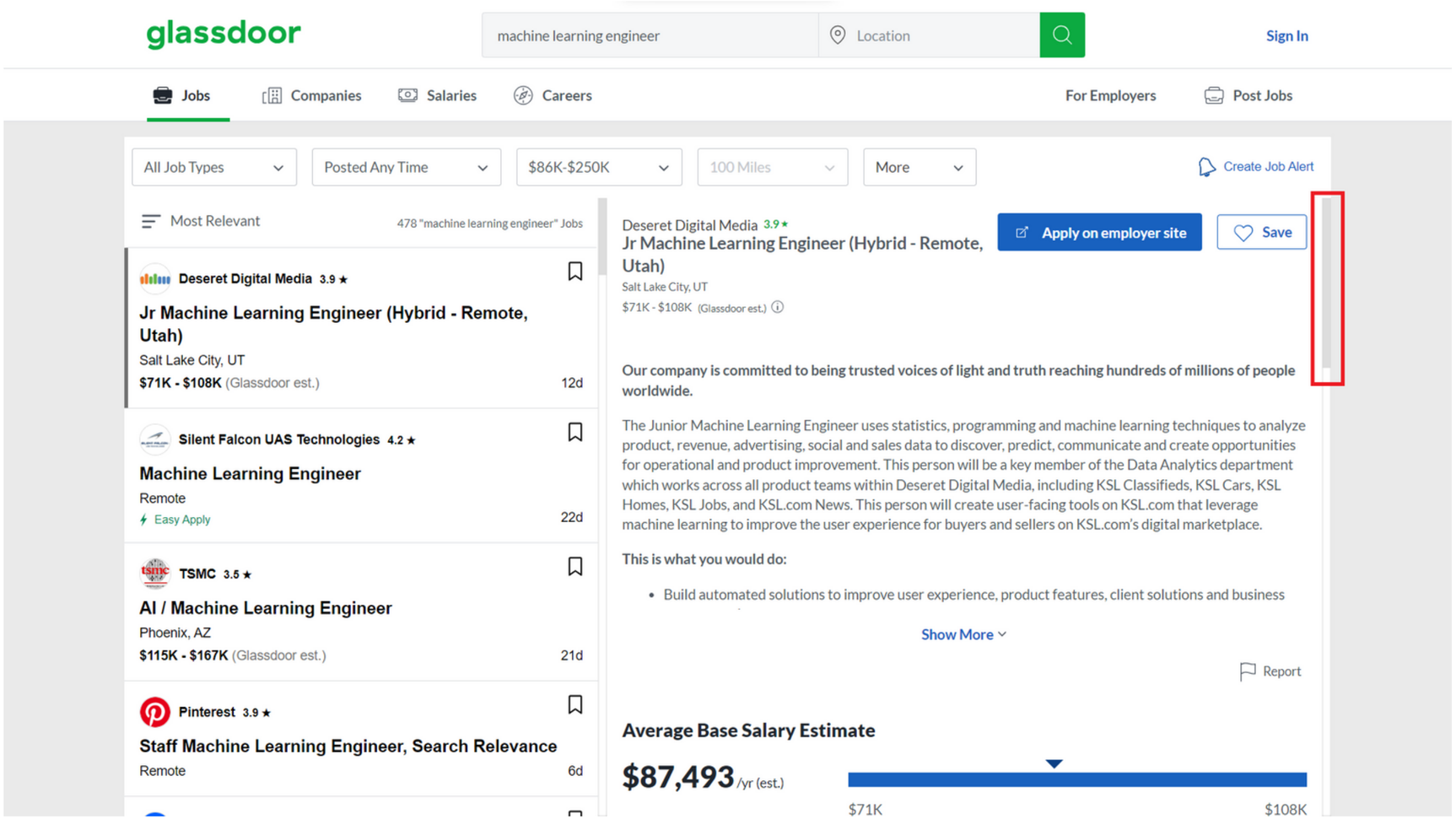

```
card.location_once_scrolled_into_view
```

Step 3: Clicking Show More Button



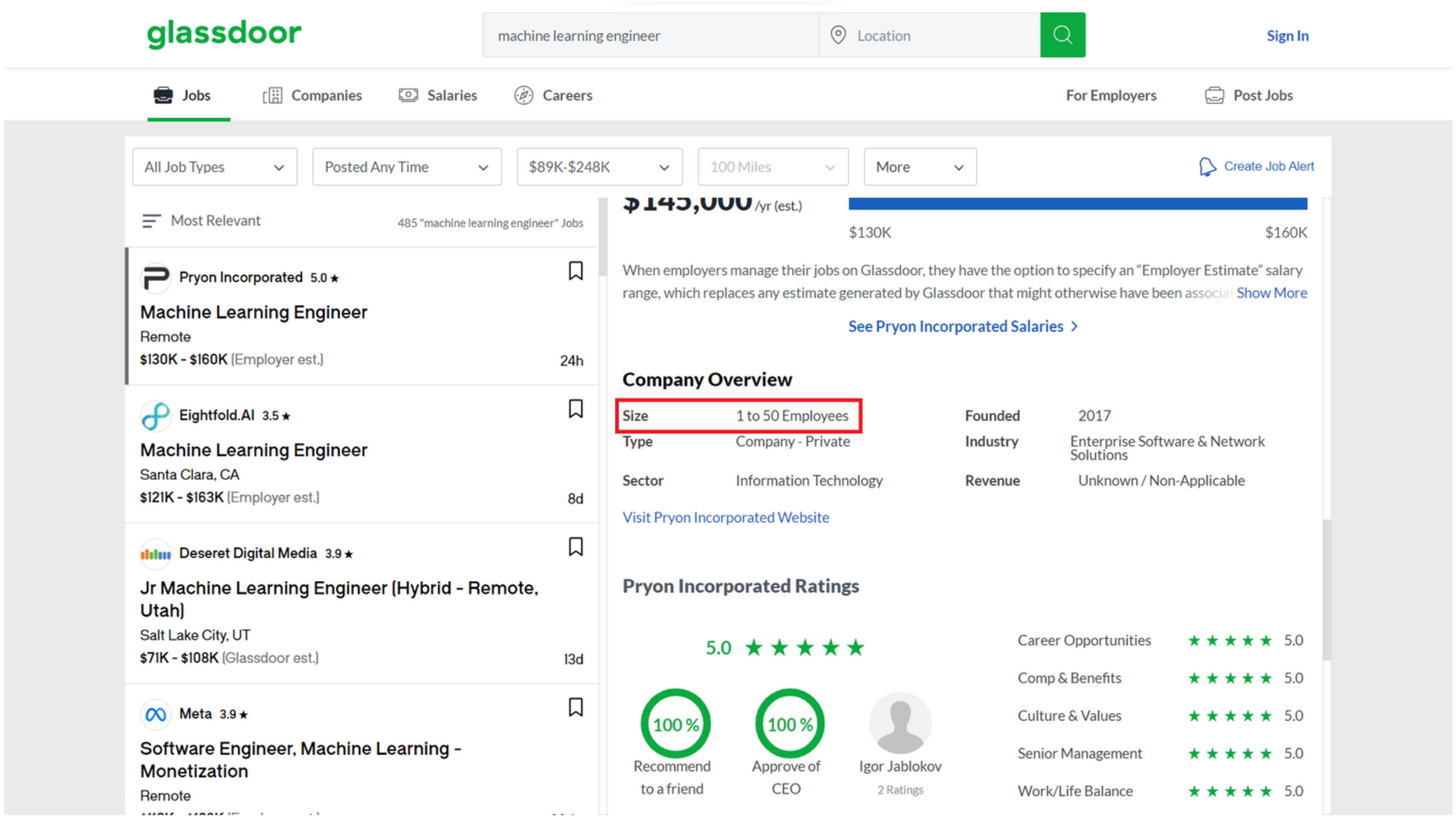

```
show_more = driver.find_element(By.XPATH, "//div[@class='p-std css-1k5huso e856ufb0']")
```

Step 4: Scroling down Job Description Section




```
show_more.location_once_scrolled_into_view
```

Step 5: Scrapping Data




```
size_elem = driver.find_element(By.XPATH, "//div[@id='CompanyContainer']//span[text()='Size']//following-sibling::*")
size.append(size_elem.text)
```

glassdoor

machine learning engineer | Location | Sign In

Jobs | Companies | Salaries | Careers | For Employers | Post Jobs

All Job Types | Posted Any Time | $88K-$250K | 100 Miles | More | Create Job Alert

Digital Diagnostics, Inc. 4.5 ★

Machine Learning Engineer

Remote

Ph.D. in a technical field, or MS in a technical field with 2+ years of industry experience, or 5+ years of industry experience. What We Are Looking For...

2d

Apollo.io 4.0 ★

Staff Machine Learning Engineer

Remote

PhD in Computer Science or related field with a focus on **machine learning** Experience with **Machine Learning** software tools and libraries (e.g., Scikit-learn,...

30d+

Seekr 4.1 ★

Machine Learning Engineer

Carlsbad, CA

$160K - $220K (Employer est.)

Easy Apply

Digital Diagnostics, Inc. 4.5 ★

Machine Learning Engineer

Remote

Apply on employer site | Save

Location – Chicago, IL | Coralville, IA | or Remote-US

Digital Diagnostics is looking for a Machine Learning Engineer to join our growing team! As a Machine Learning Engineer, you will assist with our R&D and algorithmic development efforts. You will also participate in the development of the novel machine learning and computer vision algorithms which form the core of our medical device products, design efficient experiments to analyze current algorithm performance, report on results and assist in determining future direction for algorithmic enhancement.

What We Are Looking For

- Ph.D. in a technical field, or MS in a technical field with 2+ years of industry experience, or 5+ years of industry experience.
- Outstanding analytical and problem-solving skills.

Show More

Report

Prepare

- **Insert the result to a NoSQL Database, do some queries and write a grammar to encode the result in XML**

Modify Data

Not uniform

Missing data

String
Not number

Not easy to query

	job title String	location String	salary estimate Mi	company_size Mi
1	"Software Developer"	"Remote"	null	"201 to 500 Emplo
2	"JUNIOR SOFTWARE DEVELOPER"	"New York, NY"	"$86,891 /yr (est.	"1 to 50 Employees"
3	"Software Developer I"	"Waco, TX"	"$75,820 /yr (est.)"	"501 to 1000 Employees"
4	"Jr. Software Developer"	"New York, NY"	"$107,145 /yr (est.)"	"51 to 200 Employees"
5	"Software Developer 2"	"Diamondhead, MS"	"$27.00 /hr (est.)"	"51 to 200 Employees"
6	"Software Developer 1"	"Diamondhead, MS"	"$19.11 /hr (est.)"	"51 to 200 Employees"
7	"Junior Software Developer"	"Remote"	"$82,000 /yr (est.)"	"10000+ Employees"
8	"Junior Software Developer"	"Boston, MA"	"$88,500 /yr (est.)"	"1 to 50 Employees"
9	"Junior Software Developer"	"Long Beach, CA"	"$4,584 /mo (est.)"	"5001 to 10000 Employees"
10	"Software Developer"	"Remote"	"$92,040 /yr (est.)"	"201 to 500 Employees"
11	"ReactJS/NodeJS Software De…	"Remote"	null	"Unknown"
12	"Senior Software Developer …	"Remote"	null	"1001 to 5000 Employees"
13	"Software Developer (Remote…	"Remote"	"$91,400 /yr (est.)"	"10000+ Employees"

Not easy to query

Not uniform

Modified data

job_title String	location String	monthly_salary Mixed	company_size String
"Software engineer"	"New York"	7240.916666666667	"XS"
"Software engineer"	"Texas"	6318.333333333333	"S"
"Software engineer"	"New York"	8928.75	"XS"
"Software engineer"	"Remote"	6833.333333333333	"L"
"Software engineer"	"Massachusetts"	7375	"XS"
"Software engineer"	"California"	88500	"M"
"Software engineer"	"Remote"	7670	"S"
"Software engineer"	"Remote"	7616.666666666667	"L"
"Software engineer"	"Virginia"	6833.333333333333	"L"
"Software engineer"	"Texas"	6006.5	"XS"
"Software engineer"	"Ohio"	6304.5	"M"
"Software engineer"	"California"	75654	"L"
"Software engineer"	"Virginia"	7119.75	"XS"

Insert to a NoSQL Database

Insert to a NoSQL Database

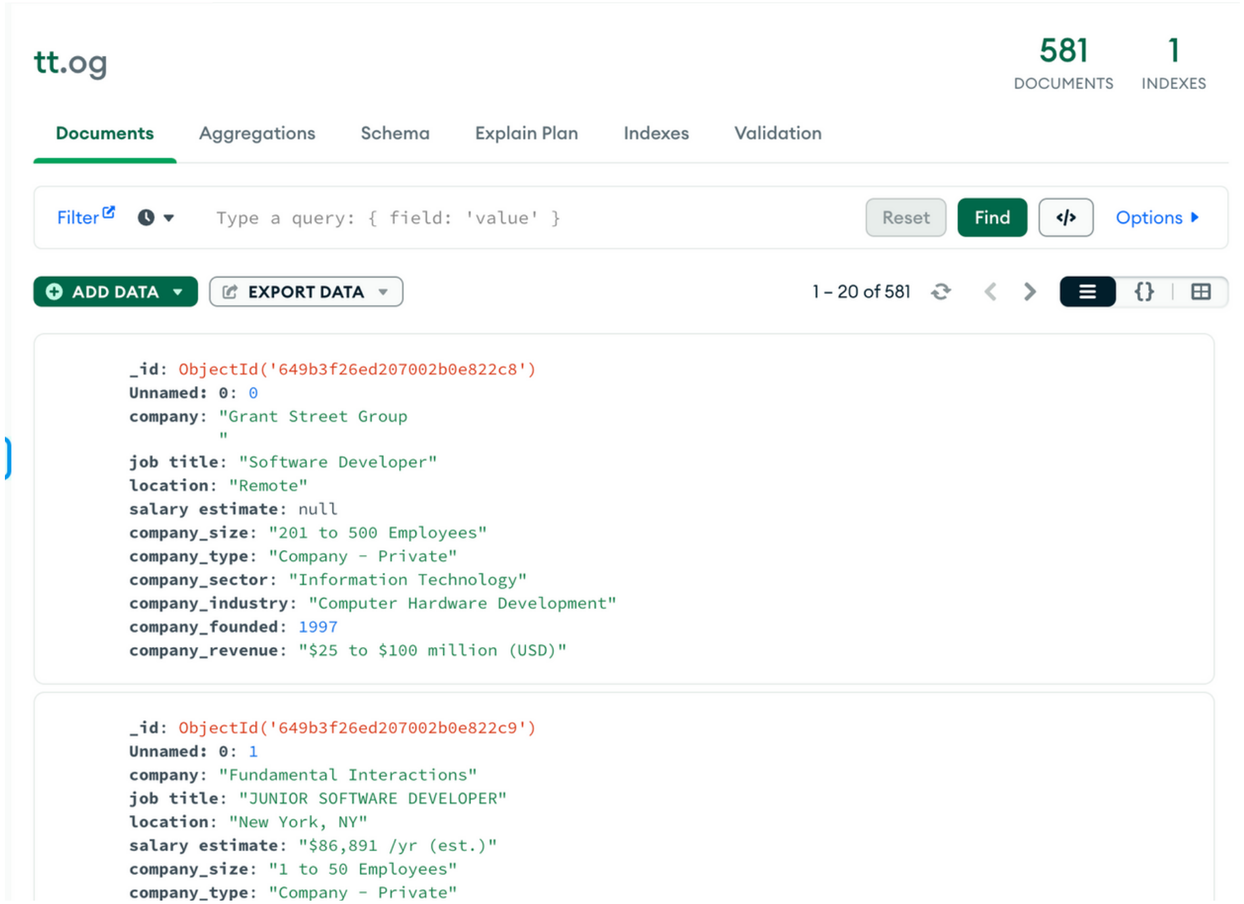
JSON file

+

MongoDB

ADD DATA

MongoDB Compass



Queries

Compass Filter Queries

Examples:

Match

*# Company size **equals** the specified value "L(large)"*

{ company_size : { $eq : "L" } }

#Job title is "Software engineer"

{ job_title : { $eq : "Software engineer" } }

Compass Filter Queries

Exclusion

*# location **is not equal to** the specified value "Remote"*

{ location: { $ne: "Remote" } }

Comparison

*# Salary is **over** $10,000/mo*

{ monthly_salary: {$gt: 10000} }

*#Salary is **less than or equal to** $5,000/mo*

{ monthly_salary: {$lte: 5000} }

Compass Filter Queries

Union

*#Company size is either "S" **or** "M"*

{company_size :{$in: ["S", "M"]}}

***#**Company size is either "S" **or** "M" **OR** remotely working*

{$or: [{company_size : { $in: ["S", "M"] } }, {"location": {$eq : "Remote"}}]}

Intersection

*#Company size is "M" **AND** monthly salary is over 10,000*

{company_size: {$eq: 'M'},monthly_salary: {$gt: 10000}}

Queries

Name：Naruto

Work in New York
Middle company
Software engineer

```
{location: {$eq:'New York'},
company_size: {$eq: 'M'},
job_title: {$eq: 'Software engineer'}}
```

Name：Sakura

Work in California
Small company
Machine Learning engineer

```
{location: {$eq:'California'},
company_size: {$eq: 'S'},
job_title: {$eq: 'Machine Learning engineer'}}
```

Name：Sasuke

Remote
Middle company
Data Scientist

```
{location: {$eq:'Remote'},
company_size: {$eq: 'M'},
job_title: {$eq: 'Data Scientist'}}
```

Queries

JSON file

+

MongoDB.

Queries

Selected data

Selected data in JSON file

Encode in XML

Encode in XML

JSON file

`xml.etree.ElementTree`

+ code

XSD file

Validate

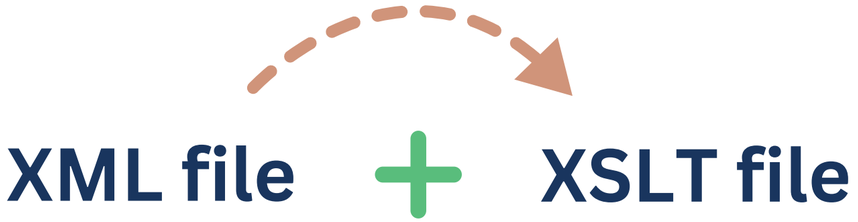
XML file

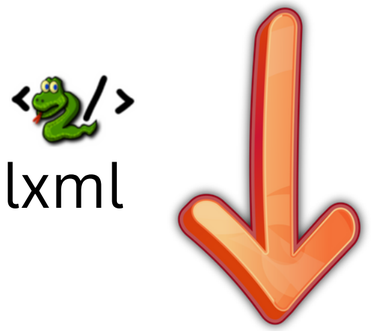
```xml
<?xml version='1.0' encoding='utf-8'?>
<data>
  <item>
    <_id>
      <oid>649b6d7fed207002b0e8336d</oid>
    </_id>
    <company>D. E. Shaw &amp;amp; Co., L.P.</company>
    <location>New York</location>
    <company_size>M</company_size>
    <company_type>Company - Private</company_type>
    <company_sector>Financial Services</company_sector>
    <company_industry>Investment &amp;amp; Asset Management</company_industry>
    <company_founded>1988</company_founded>
    <company_revenue>Unknown / Non-Applicable</company_revenue>
    <monthly_salary>16666.666666666668</monthly_salary>
    <number>56</number>
    <job_title>Software engineer</job_title>
    <experience_level>None</experience_level>
  </item>
</data>
```

Access

- **Use XSLT to present the results on a webpage**

Present on a Webpage



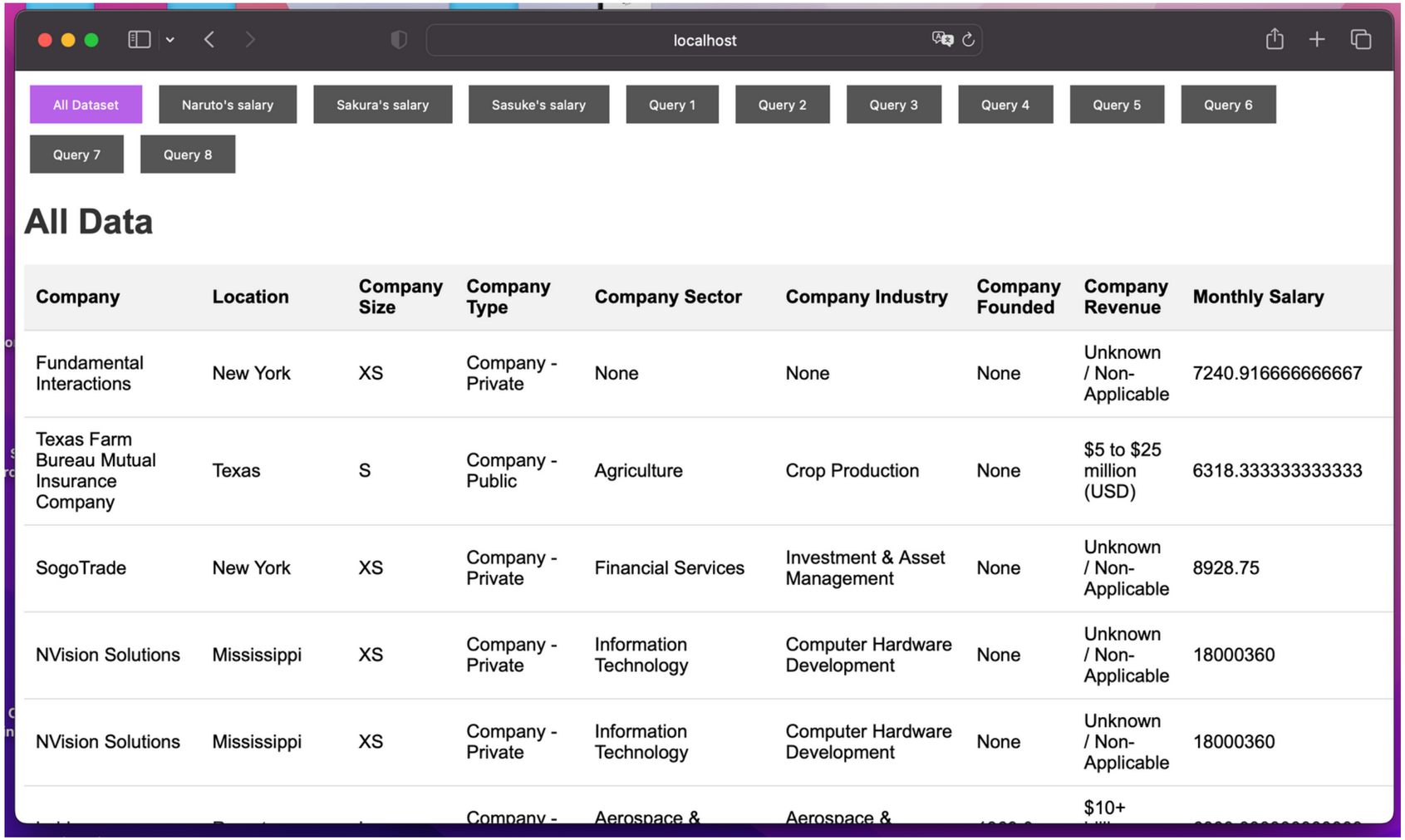


file:///Users/wenwen/Desktop/TT_project/Home_page/tt_project_dat

Job Listings

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Nu
Fundamental Interactions	New York	XS	Company - Private	None	None	None	Unknown / Non-Applicable	7240.916666666667	1
Texas Farm Bureau Mutual Insurance Company	Texas	S	Company - Public	Agriculture	Crop Production	None	$5 to $25 million (USD)	6318.333333333333	2
SogoTrade	New York	XS	Company - Private	Financial Services	Investment & Asset Management	None	Unknown / Non-Applicable	8928.75	3
NVision Solutions	Mississippi	XS	Company - Private	Information Technology	Computer Hardware Development	None	Unknown / Non-Applicable	18000360	4
NVision Solutions	Mississippi	XS	Company - Private	Information Technology	Computer Hardware Development	None	Unknown / Non-Applicable	18000360	5
Leidos	Remote	L	Company - Public	Aerospace & Defense	Aerospace & Defense	1969.0	$10+ billion (USD)	6833.333333333333	6

The webpage is like....



According to the database, the winner should be...

localhost

All Dataset | Naruto's salary | Sakura's salary | Sasuke's salary | Query 1 | Query 2 | Query 3 | Query 4 | Query 5 | Query 6

Query 7 | Query 8

Naruto's salary

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number	Job Title	Experien Level
D. E. Shaw & Co., L.P.	New York	M	Company - Private	Financial Services	Investment & Asset Management	1988	Unknown / Non-Applicable	16666.666666666668	56	Software engineer	None
D. E. Shaw & Co., L.P.	New York	M	Company - Private	Financial Services	Investment & Asset Management	1988	Unknown / Non-Applicable	16666.666666666668	70	Software engineer	None
D. E. Shaw & Co., L.P.	New York	M	Company - Private	Financial Services	Investment & Asset Management	1988	Unknown / Non-Applicable	16666.666666666668	95	Software engineer	None

localhost

All Dataset | Naruto's salary | Sakura's salary | Sasuke's salary | Query 1 | Query 2 | Query 3 | Query 4 | Query 5 | Query 6

Query 7 | Query 8

Sakura's salary

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number	Job Title	E L
Eightfold.AI	California	S	Company - Private	Information Technology	Enterprise Software & Network Solutions	2016	Unknown / Non-Applicable	11833.333333333334	22	Machine Learning engineer	N
Neuralink	California	S	Company - Private	Pharmaceutical & Biotechnology	Biotech & Pharmaceuticals	2016	Unknown / Non-Applicable	16175	48	Machine Learning engineer	N
Neuralink	California	S	Company - Private	Pharmaceutical & Biotechnology	Biotech & Pharmaceuticals	2016	Unknown / Non-Applicable	16175	65	Machine Learning engineer	N

localhost

All Dataset | Naruto's salary | Sakura's salary | Sasuke's salary | Query 1 | Query 2 | Query 3 | Query 4 | Query 5 | Query 6

Query 7 | Query 8

Sasuke's salary

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number	Job Title	Experience Level
Title Resource Group	Remote	M	Subsidiary or Business Segment	Insurance	Insurance Carriers	None	$1 to $5 billion (USD)	6250	5	Data Scientist	None

Name：*Naruto*

Work in New York
Middle company
Software engineer

Name：*Sakura*

Work in California
Small company
Machine Learning engineer

Name：*Sasuke*

Remote
Middle company
Data Scientist

References

Data source: Glassdoor

Selenium

Xpath In Selenium

MongoDB

Query Documents

XML Schema Tutorial

XSLT-Transformation

Thank you !