

# **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



Job Salary Data



DATASET



CSV File



Pandas Data Frame



JSON File



Our JSON Dataset is ready



# **Scraping Part Step by Step**

# Step 1: Clicking Job card

The screenshot shows the Glassdoor website interface. At the top, there is a navigation bar with the Glassdoor logo, a search bar containing 'machine learning engineer', a location selector, a search button, and a 'Sign In' link. Below the navigation bar, there are links for 'Jobs', 'Companies', 'Salaries', 'Careers', 'For Employers', and 'Post Jobs'. The main content area displays search filters like 'All Job Types', 'Posted Any Time', 'Salary Range', 'Location', and 'More'. A 'Create Job Alert' button is also present. The search results are listed under 'Most Relevant' with a count of 478 jobs. The first result is highlighted with a red box. It is for 'Deseret Digital Media' (3.9★) and the position is 'Jr Machine Learning Engineer (Hybrid - Remote, Utah)'. The job is located in Salt Lake City, UT, with an estimated salary range of '\$71K - \$108K'. The posting was 12 days ago. To the right of the job listing, there is a detailed description of the role, mentioning that the company is committed to being trusted voices of light and truth reaching hundreds of millions of people worldwide. It describes the junior machine learning engineer's responsibilities, which include building automated solutions to improve user experience, product features, client solutions, and business. There is a 'Show More' link for the responsibilities. Below the responsibilities, there is an 'Average Base Salary Estimate' chart showing a bar from '\$71K' to '\$108K' with a value of '\$87,493 /yr (est.)'. There are also 'Apply on employer site' and 'Save' buttons. Other job cards are visible below the first one, such as 'Machine Learning Engineer' at Silent Falcon UAS Technologies and 'AI / Machine Learning Engineer' at TSMC.

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

# Step 2: Scrolling down Job cards section

The screenshot shows the Glassdoor website interface for searching 'machine learning engineer' jobs. The search bar at the top contains 'machine learning engineer'. Below the search bar are filters for 'All Job Types', 'Posted Any Time', '\$86K-\$250K', '100 Miles', and 'More'. A 'Create Job Alert' button is also present. The main content area displays a list of job cards:

- Deseret Digital Media** 3.9★  
**Jr Machine Learning Engineer (Hybrid - Remote, Utah)**  
Salt Lake City, UT  
\$71K - \$108K (Glassdoor est.)
- Silent Falcon UAS Technologies** 4.2★  
**Machine Learning Engineer**  
Remote  
Easy Apply
- TSMC** 3.5★  
**AI / Machine Learning Engineer**  
Phoenix, AZ  
\$115K - \$167K (Glassdoor est.)
- Pinterest** 3.9★  
**Staff Machine Learning Engineer, Search Relevance**  
Remote

On the right side of the job card for Deseret Digital Media, there is a red box highlighting the 'Bookmark' icon. Below the job card, there is a summary of the company's mission, a detailed description of the job responsibilities, and a list of tasks. At the bottom, there is an 'Average Base Salary Estimate' bar chart showing a range from \$71K to \$108K, with a value of \$87,493 indicated.

```
card.location.once_scrolled_into_view
```

# Step 3: Clicking Show More Button

The screenshot shows the Glassdoor website interface. At the top, there is a search bar with 'machine learning engineer' typed in, a location input field, a green search button, and a 'Sign In' link. Below the header, there are navigation links for 'Jobs', 'Companies', 'Salaries', 'Careers', 'For Employers', and 'Post Jobs'. The main content area displays a list of job listings under the heading 'Most Relevant'. The first listing is for 'Deseret Digital Media' in Salt Lake City, UT, offering a Jr Machine Learning Engineer position (Hybrid - Remote) with a salary range of \$71K - \$108K. The second listing is for 'Silent Falcon UAS Technologies' in Salt Lake City, UT, offering a Machine Learning Engineer position (Remote) with an 'Easy Apply' option. The third listing is for 'TSMC' in Phoenix, AZ, offering an AI / Machine Learning Engineer position with a salary range of \$115K - \$167K. The fourth listing is for 'Pinterest' in San Francisco, CA, offering a Staff Machine Learning Engineer position (Search Relevance) with a salary range of \$115K - \$167K. On the right side of the page, there is a detailed view for the Deseret Digital Media job, showing the company's average base salary estimate of \$87,493 (est.) per year. A red box highlights the 'Show More' button next to the salary details.

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

# Step 4: Scrolling down Job Description Section

The screenshot shows the Glassdoor website interface. At the top, there is a navigation bar with the Glassdoor logo, a search bar containing 'machine learning engineer', a location selector, a search button, and a 'Sign In' link. Below the navigation bar, there are links for 'Jobs', 'Companies', 'Salaries', 'Careers', 'For Employers', and 'Post Jobs'. The main content area displays a list of job listings under the heading 'Most Relevant'. The first listing is for 'Deseret Digital Media' in Salt Lake City, UT, offering a Jr Machine Learning Engineer position (Hybrid - Remote, Utah) with a salary range of \$71K - \$108K (Glassdoor est.). The second listing is for 'Silent Falcon UAS Technologies' in Remote, offering a Machine Learning Engineer position with an 'Easy Apply' option and a salary range of \$71K - \$108K (Glassdoor est.). The third listing is for 'TSMC' in Phoenix, AZ, offering an AI / Machine Learning Engineer position with a salary range of \$115K - \$167K (Glassdoor est.). The fourth listing is for 'Pinterest' in Remote, offering a Staff Machine Learning Engineer, Search Relevance position with a salary range of \$71K - \$108K (Glassdoor est.). On the right side of the job listing for Deseret Digital Media, there is a 'Create Job Alert' button, an 'Apply on employer site' button, and a 'Save' button. A red vertical rectangle highlights the scroll bar on the right side of the page. At the bottom of the page, there is a section titled 'Average Base Salary Estimate' showing '\$87,493 /yr (est.)' with a horizontal bar chart ranging from '\$71K' to '\$108K'.

```
show_more.location_once_scrolled_into_view
```

# Step 5: Scrapping Data

The screenshot shows the Glassdoor homepage with a search bar for 'machine learning engineer'. Below the search bar are filters for job type, posted time, salary range, location, and more. A 'Create Job Alert' button is also present. The main content displays a list of job listings, each with a company logo, name, rating, location, and salary range. To the right of the job list is a 'Company Overview' section for 'Pryon Incorporated' with details like size (1 to 50 employees), founded year (2017), industry (Enterprise Software & Network Solutions), and sector (Information Technology). Below this is a 'Pryon Incorporated Ratings' section showing a 5.0 rating, 100% recommend rate, 100% approve rate, and a profile for Igor Jablokov.

Size	1 to 50 Employees	Founded	2017
Type	Company - Private	Industry	Enterprise Software & Network Solutions
Sector	Information Technology	Revenue	Unknown / Non-Applicable

Pryon Incorporated Ratings	
5.0 ★★★★★	Career Opportunities ★★★★★ 5.0
100 % Recommend to a friend	Comp & Benefits ★★★★★ 5.0
100 % Approve of CEO	Culture & Values ★★★★★ 5.0
Igor Jablokov 2 Ratings	Senior Management ★★★★★ 5.0
	Work/Life Balance ★★★★★ 5.0

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

Chrome otomatik test yazılımı tarafından kontrol ediliyor.



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:



Digital Diagnostics, Inc. 4.5★  
Machine Learning Engineer

Remote

2d

 Apply on employer site

Save

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



## 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,

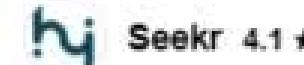


Digital Diagnostics is looking for a Machine Learning Engineer to join our growing team! As a Machine Learning Engineer, you will assist with our R&D and algorithmic development efforts. You will also participate in the development of the novel machine learning and computer vision algorithms which form 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.

30d+

### 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](#)

## Machine Learning Engineer

Carlsbad, CA

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

Easy Apply

Report

# **Prepare**

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

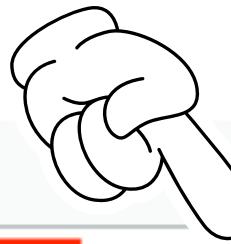
# Modify Data

## Not uniform



	job title String
1	"Software Developer"
2	"JUNIOR SOFTWARE DEVELOPER"
3	"Software Developer I"
4	"Jr. Software Developer"
5	"Software Developer 2"
6	"Software Developer 1"
7	"Junior Software Developer"
8	"Junior Software Developer"
9	"Junior Software Developer"
10	"Software Developer"
11	"ReactJS/NodeJS Software Dev."
12	"Senior Software Developer ..."
13	"Software Developer (Remote)"

## Missing data



	location String
1	"Remote"
2	"New York, NY"
3	"Waco, TX"
4	"New York, NY"
5	"Diamondhead, MS"
6	"Diamondhead, MS"
7	"Remote"
8	"Boston, MA"
9	"Long Beach, CA"
10	"Remote"
11	"Remote"
12	"Remote"
13	"Remote"

Not easy to query

## String Not number



	salary estimate Mi...
1	null
2	"\$86,891 /yr (est.)"
3	"\$75,820 /yr (est.)"
4	"\$107,145 /yr (est.)"
5	"\$27.00 /hr (est.)"
6	"\$19.11 /hr (est.)"
7	"\$82,000 /yr (est.)"
8	"\$88,500 /yr (est.)"
9	"\$4,584 /mo (est.)"
10	"\$92,040 /yr (est.)"
11	null
12	null
13	"\$91,400 /yr (est.)"

Not uniform

## Not easy to query



	company_size Mix
1	"201 to 500 Employees"
2	"1 to 50 Employees"
3	"501 to 1000 Employees"
4	"51 to 200 Employees"
5	"51 to 200 Employees"
6	"51 to 200 Employees"
7	"10000+ Employees"
8	"1 to 50 Employees"
9	"5001 to 10000 Employees"
10	"201 to 500 Employees"
11	"Unknown"
12	"1001 to 5000 Employees"
13	"10000+ Employees"

# Modified data

job_title String	location String	monthly_salary Mixed	company_size String
"Software engineer"	"New York"	7240.916666666667	"XS"
"Software engineer"	"Texas"	6318.33333333333	"S"
"Software engineer"	"New York"	8928.75	"XS"
"Software engineer"	"Remote"	6833.33333333333	"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"	6833.33333333333	"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

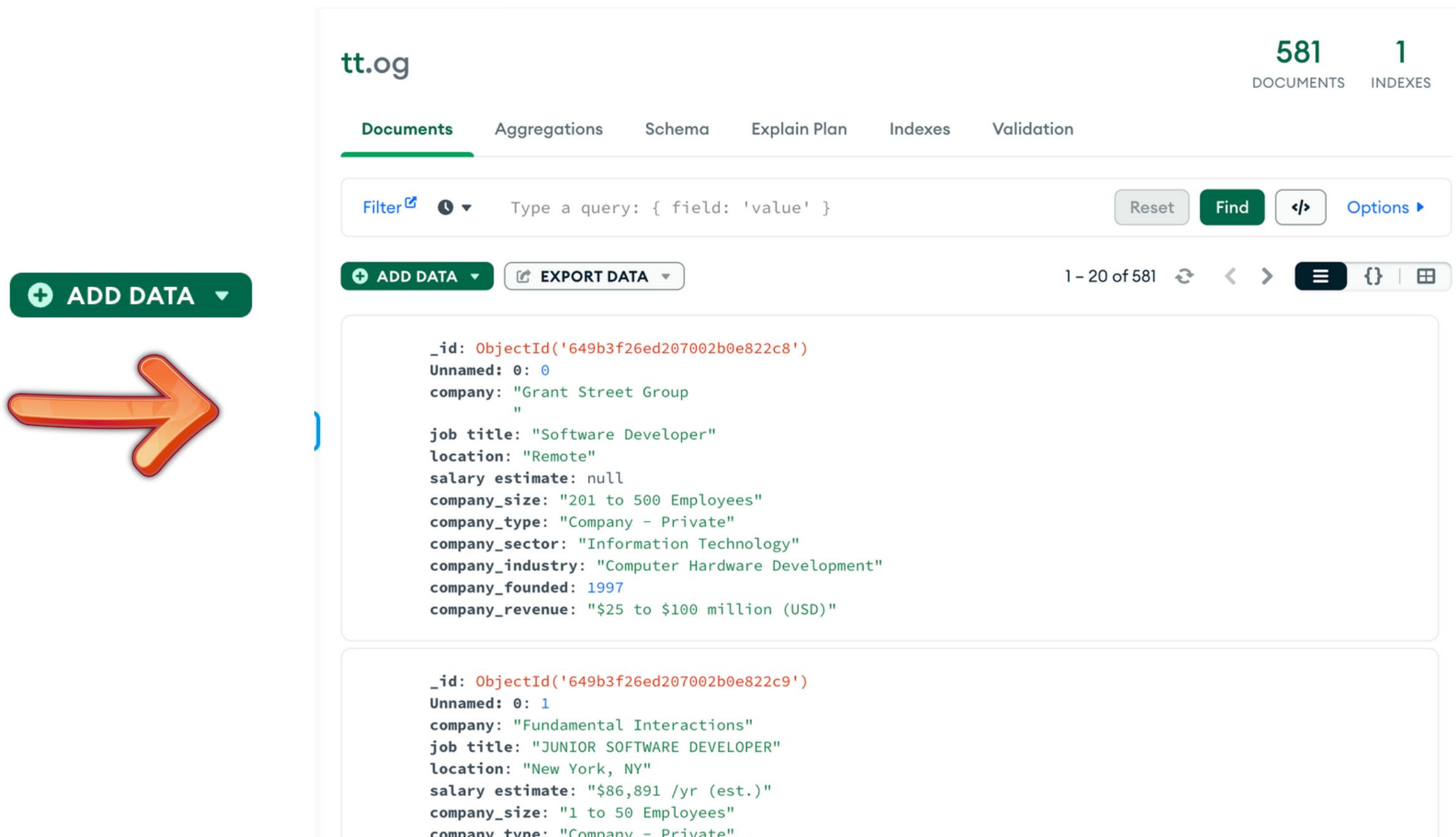
# Insert to a NoSQL Database

## MongoDB Compass

JSON file



 MongoDB®



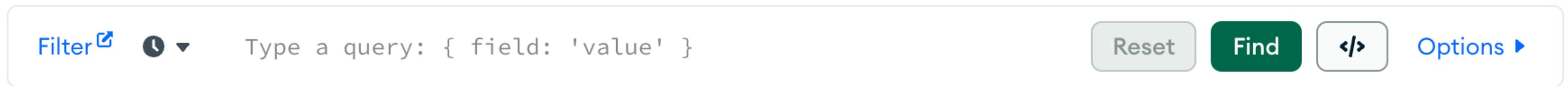
The screenshot shows the MongoDB Compass interface for a collection named "tt.og". The top right corner displays "581 DOCUMENTS" and "1 INDEXES". The main area is titled "Documents" and contains two documents listed vertically. Each document is represented by a code block:

```
_id: ObjectId('649b3f26ed207002b0e822c8')
Unnamed: 0: 0
company: "Grant Street Group"
"
job_title: "Software Developer"
location: "Remote"
salary_estimate: null
company_size: "201 to 500 Employees"
company_type: "Company - Private"
company_sector: "Information Technology"
company_industry: "Computer Hardware Development"
company_founded: 1997
company_revenue: "$25 to $100 million (USD)"

_id: ObjectId('649b3f26ed207002b0e822c9')
Unnamed: 0: 1
company: "Fundamental Interactions"
job_title: "JUNIOR SOFTWARE DEVELOPER"
location: "New York, NY"
salary_estimate: "$86,891 /yr (est.)"
company_size: "1 to 50 Employees"
company_type: "Company - Private"
```

# 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

+ ADD DATA ▾



Queries

 MongoDB®

Selected data  
in JSON file



EXPORT DATA ▾



Selected data



# Encode in XML

# Encode in XML

JSON file



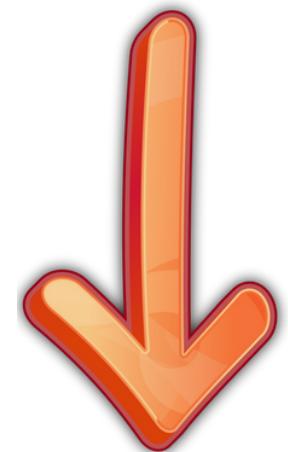
`xml.etree.ElementTree`

+ code

XML file



XSD file



Validate

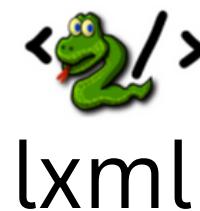
```
<?xml version='1.0' encoding='utf-8'?>
<data>
  <item>
    <_id>
      <oid>649b6d7fed207002b0e8336d</oid>
    </_id>
    <company>D. E. Shaw & Co., L.P.</company>
    <location>New York</location>
    <company_size>M</company_size>
    <company_type>Company - Private</company_type>
    <company_sector>Financial Services</company_sector>
    <company_industry>Investment & Asset Management</company_industry>
    <company_founded>1988</company_founded>
    <company_revenue>Unknown / Non-Applicable</company_revenue>
    <monthly_salary>16666.66666666668</monthly_salary>
    <number>56</number>
    <job_title>Software engineer</job_title>
    <experience_level>None</experience_level>
  </item>
</data>
```

# Access

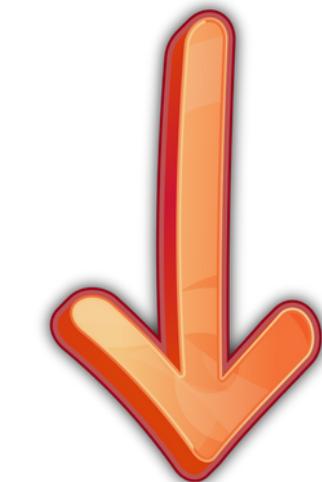
- Use XSLT to present the results on a webpage

# Present on a Webpage

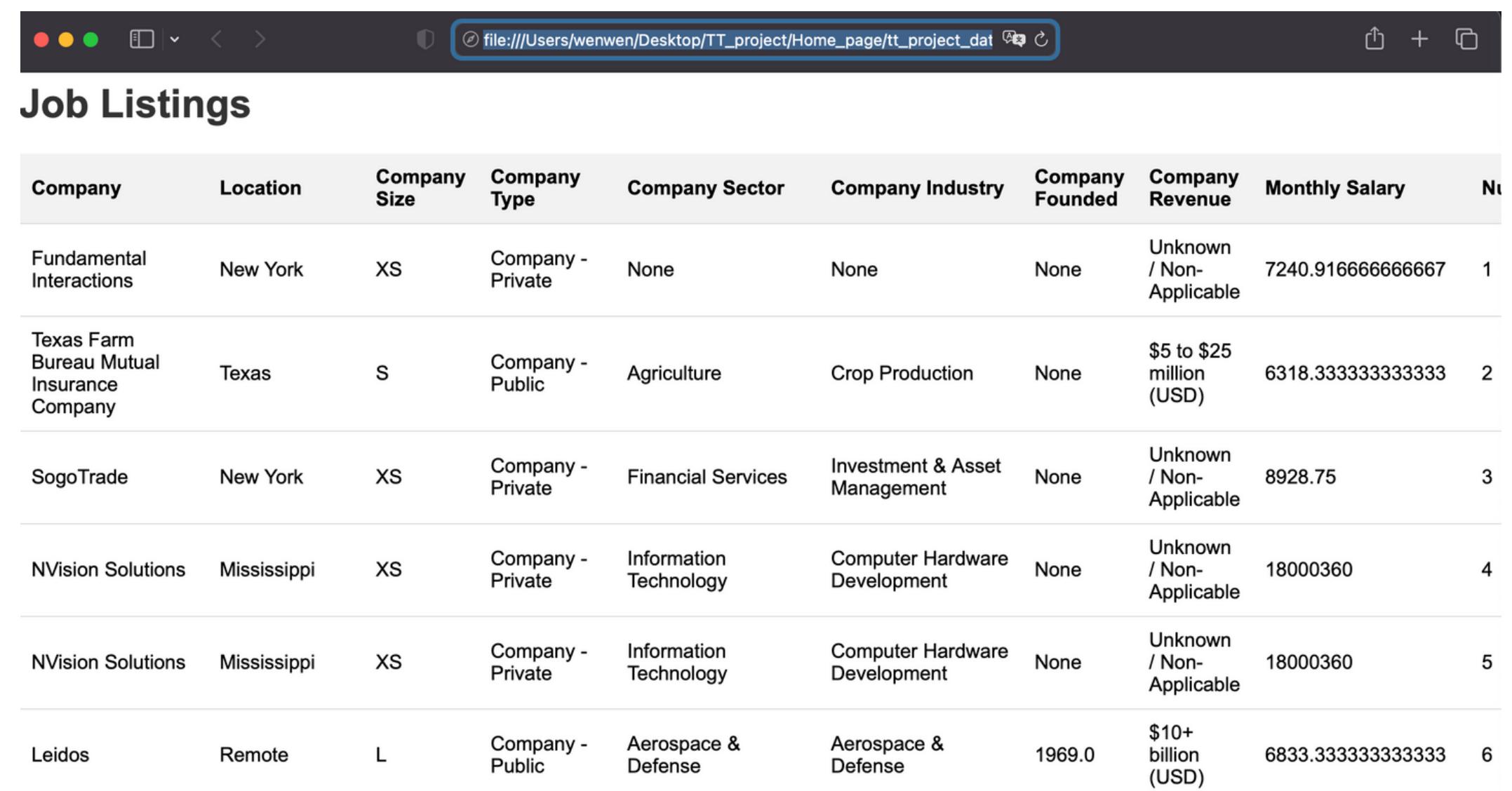
XML file + XSLT file



lxml



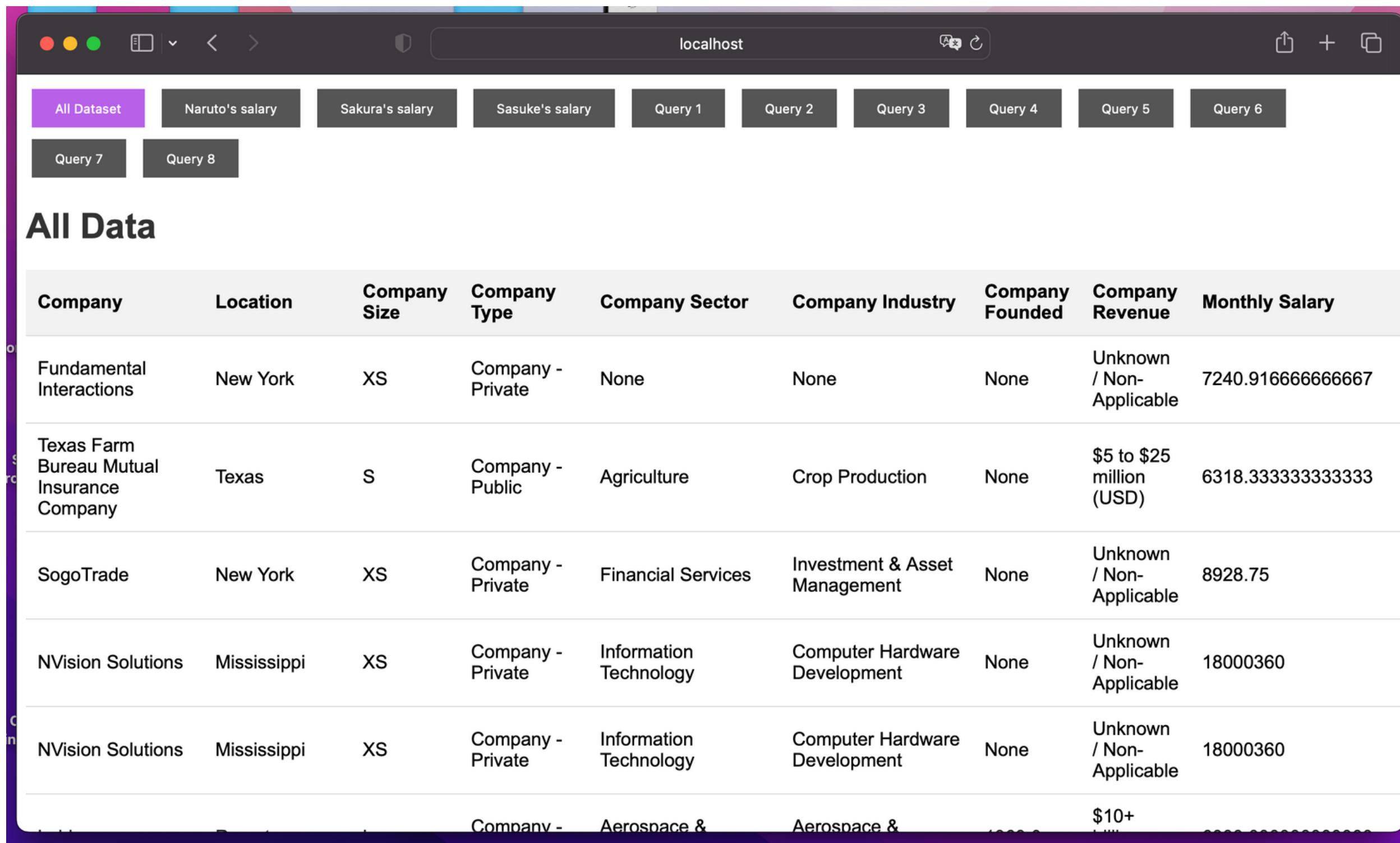
HTML file



A screenshot of a web browser window displaying a table titled "Job Listings". The table has 11 columns: Company, Location, Company Size, Company Type, Company Sector, Company Industry, Company Founded, Company Revenue, Monthly Salary, and Number. The data is as follows:

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number
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....



A screenshot of a web browser window titled "localhost". The window contains a navigation bar with buttons for "All Dataset", "Naruto's salary", "Sakura's salary", "Sasuke's salary", "Query 1" through "Query 8". Below this is a section titled "All Data" containing a table with 10 rows of company information.

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary
Fundamental Interactions	New York	XS	Company - Private	None	None	None	Unknown / Non-Applicable	7240.916666666667
Texas Farm Bureau Mutual Insurance Company	Texas	S	Company - Public	Agriculture	Crop Production	None	\$5 to \$25 million (USD)	6318.33333333333
SogoTrade	New York	XS	Company - Private	Financial Services	Investment & Asset Management	None	Unknown / Non-Applicable	8928.75
NVision Solutions	Mississippi	XS	Company - Private	Information Technology	Computer Hardware Development	None	Unknown / Non-Applicable	18000360
NVision Solutions	Mississippi	XS	Company - Private	Information Technology	Computer Hardware Development	None	Unknown / Non-Applicable	18000360
			Company -	Aerospace &	Aerospace &		\$10+	

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





A screenshot of a web browser window titled "localhost". The address bar shows "localhost". Below the title bar are several buttons: a back arrow, a forward arrow, a shield icon, a refresh icon, and a search icon. To the right of the search icon are three icons: a thumbs up, a plus sign, and a square.

The main content area contains a series of buttons arranged horizontally. From left to right, they are: "All Dataset" (grey), "Naruto's salary" (purple, indicating it is selected), "Sakura's salary" (grey), "Sasuke's salary" (grey), "Query 1" (grey), "Query 2" (grey), "Query 3" (grey), "Query 4" (grey), "Query 5" (grey), "Query 6" (grey), "Query 7" (grey), and "Query 8" (grey).

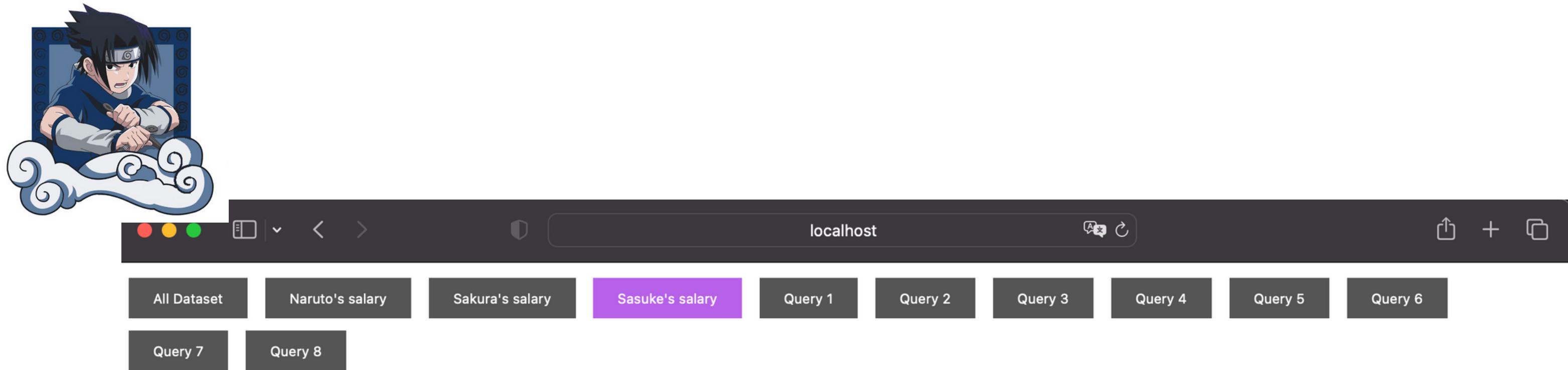
## Naruto's salary

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number	Job Title	Experience 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



## Sakura's salary

Company	Location	Company Size	Company Type	Company Sector	Company Industry	Company Founded	Company Revenue	Monthly Salary	Number	Job Title	Employee ID
Eightfold.AI	California	S	Company - Private	Information Technology	Enterprise Software & Network Solutions	2016	Unknown / Non-Applicable	11833.33333333334	22	Machine Learning engineer	N-00000000000000000000000000000000
Neuralink	California	S	Company - Private	Pharmaceutical & Biotechnology	Biotech & Pharmaceuticals	2016	Unknown / Non-Applicable	16175	48	Machine Learning engineer	N-00000000000000000000000000000001
Neuralink	California	S	Company - Private	Pharmaceutical & Biotechnology	Biotech & Pharmaceuticals	2016	Unknown / Non-Applicable	16175	65	Machine Learning engineer	N-00000000000000000000000000000002



## 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 !**