

# IT7993 Capstone Spring 2021

## Key Professional Dataset- Data Spider

### Abstract

The purpose of this project is to build a Web Crawler to extract personal information from a public website like Reddit and LinkedIn. We completed the Instagram crawling as a bonus for the project. The team will be using MySQL or any other open source relational database to organize the data and conduct a quantitative data analysis on it.

<https://sites.google.com/view/it7993-spring-2021/home>

### Introduction

Our goal is to establish a one stop shop for institutional asset management distribution intelligence; the one place to go for mandates, documents, and profiles of consultants, investors, and managers with key professional contact information.

### Materials and Methods

The Key Professionals dataset we call dataspider is aimed at delivering global coverage of investors and consultant key professional that are involved in making investment decisions beginning with US based firms. Jing Wang from eVestment has provided a list of 400 asset managers to crawl data, organize in a database and create appropriate visualization tools. By the end of this project, the goal is to provide coverage for the US Investors and Consultants found within eVestment platform.

- Asset manager list provided by Jing Wang
- Github list provided by Dr. Han
- Visual Studio Code
- Python
- MySQL Workbench/My SQL Installer for Window
- Jupyter
- Db4free.net
- PowerBI
- Tableau 2020.4

### Code Snippet

```

1 from linkedin_scraper import Company, actions
2 from selenium import webdriver
3 from selenium.webdriver import Chrome
4 from selenium.webdriver.common.keys import Keys
5 import mysql.connector
6 import time
7
8 # let Python know we using Chrome Driver
9 driver = webdriver.Chrome()
10
11 # must type your own LinkedIn email and password
12 email =
13 password =
14
15 # login use the info you provide above
16 actions.login(driver, email, password)
17
18 # scraping company
19 url = "https://www.linkedin.com/company/achmea-investment-management/"
20 company = Company(linkedin_url=url, name=True, specialties=True, company_size=True, driver=driver, scrape=False, go)
21
22 # sleep Chrome to wait 8 second for LinkedIn to load before scrape
23 time.sleep(8)
24

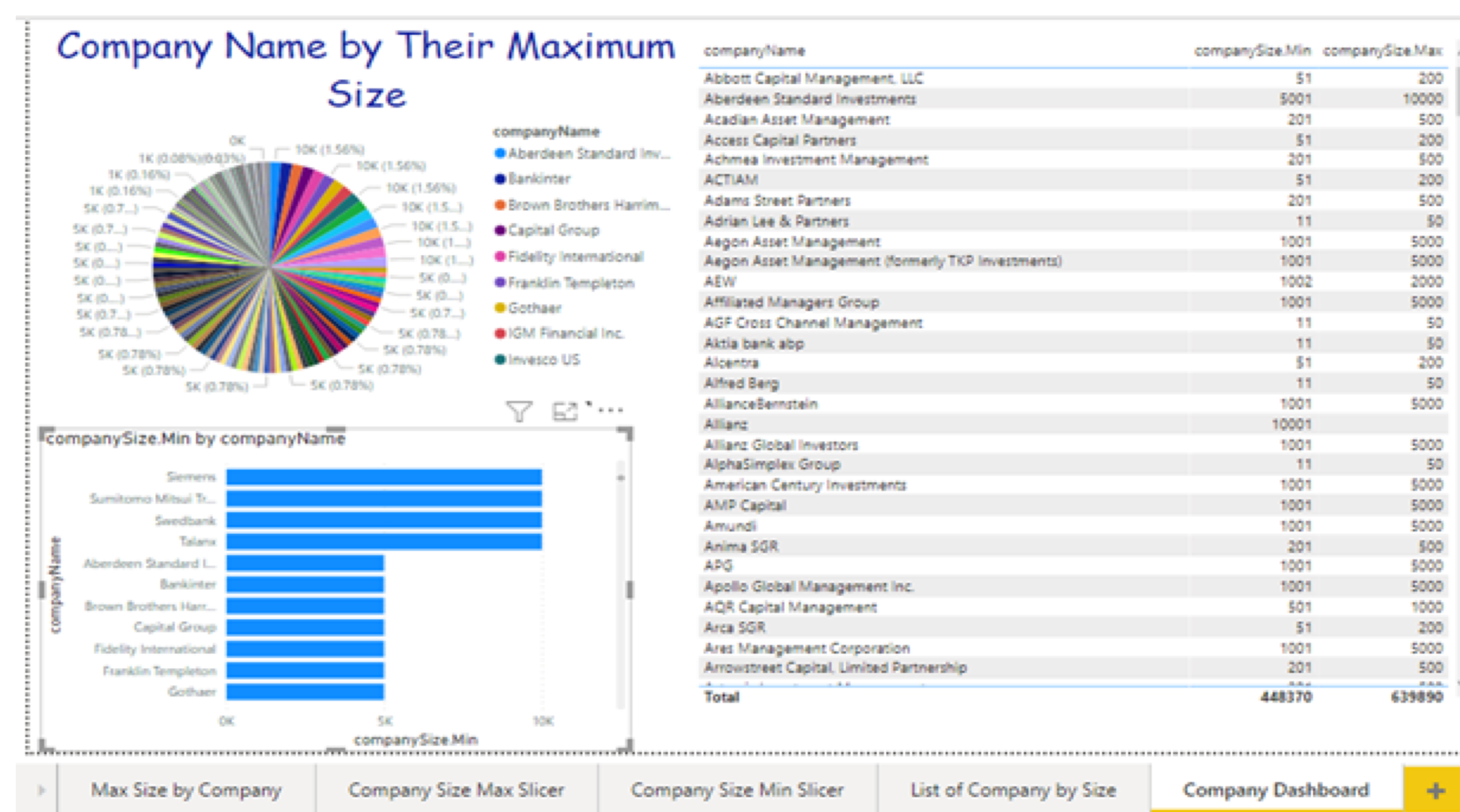
```

### Database

MySQL Workbench query results:

companyID	personID	companyName	companySize	name	current_position	from_date	to_date	location
31	1	John Hancock Investment Management	501-1,000 employees	Matthew Mokin CFA	Company Name John Han...			
34	2	China Asset Management Co., Ltd.	501-1,000 employees	Qingze (David) Lin	Company Name China Ass...	Jul 2004	May 2011	Shanghai
35	3	Svebank	10,001 employees	Tomas Bekuts, CFA	Fund Manager	11 yrs		2 yrs 8 mos
36	4	AMP Capital	1,001-5,000 employees	Claire (Stanfort) Tak...	Company Name AMP Capital			
37	5	Bridgewater Associates	1,001-5,000 employees	Daniel Hofman	Company Name Bridgewater...	Jan 2010	Mar 2010	Westport, CT
37	6	Bridgewater Associates	1,001-5,000 employees	Kevin Brennan	Company Name Bridgewater...	Jul 2001	Aug 2001	Washington, District of Col...
80	7	Achmea Investment Management	201-500 employees	Lucas Bounhus	Senior Portfolio Manager			
80	8	Achmea Investment Management	201-500 employees	Amr Saleem, CFA	Senior Portfolio Manager ...			
39	11	Harris Associates	51-200 employees	Ben Nelson	Sr Investment Analyst	Mar 1998	Jan 2000	Greater Los Angeles Area
39	12	Harris Associates	51-200 employees	Jeremy Thomas	Investment Research Ana...			
40	13	SURA Asset Management	5,001-10,000 employees	Jose R. Carlos Justo	Equity Portfolio Manager	4 yrs	mos	2 yrs
40	14	SURA Asset Management	5,001-10,000 employees	Ignacio Calle Cuartas	CEO			
41	15	ClearBridge Investments	201-500 employees	Daniel P. Finnegan	Company Name ClearBrid...	Sep 2007	Jul 2011	Stanford, CT
41	16	ClearBridge Investments	201-500 employees	Aram Green	Portfolio Manager			

### Visualization



### Conclusions

The project is to create a data repository and visualization for 400 asset managers that are provided by our sponsor. The objective was achieved through the Data Spider system that we created. The system crawls different social media such as LinkedIn, Reddit, and Instagram. The crawled data then persists into a centralized MySQL database. A visualization is then built on top utilizing Tableau dashboard. Since we do have a final product that is working as intended, the project is successful.

### Acknowledgments

Dr. Meng Han - Professor  
Jing Wang - Sponsor

### Contact Information

Janel Westmoreland  
[westmo6@students.kennesaw.edu](mailto:westmo6@students.kennesaw.edu)  
Kajal Vaghani  
[kvaghani@students.kennesaw.edu](mailto:kvaghani@students.kennesaw.edu)  
Ritu Choudhary  
[rchoudh2@students.kennesaw.edu](mailto:rchoudh2@students.kennesaw.edu)  
Vy Duong  
[vduong.it7@gmail.com](mailto:vduong.it7@gmail.com)  
Nyong Nkereuwem  
[johnsonjuniornn@gmail.com](mailto:johnsonjuniornn@gmail.com)

### References

The Source for Institutional Intelligence. eVestment. (2021, April 9). <https://www.evestment.com/>.

Boe, B. (n.d.). PRAW: The Python Reddit API Wrapper. Retrieved from <https://praw.readthedocs.io/en/latest/>

Greening, C. (2021). instascrape. Retrieved from <https://chris-greening.github.io/instascrape/>

Sham, J. (2021, April 10). Scrapes user data from LinkedIn. Retrieved from <https://pypi.org/project/linkedin-scraper/>