



### 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
- Juypter
- Db4free.net
- PowerBl
- Tableau 2020.4

# **IT7993 Capstone Spring 2021 Key Professional Dataset- Data Spider**

# **Code Snippet**

× 3	File Edit Selection View Go	Run Terminal Help	linked	in_scrap.py - IT7993 - Visual Studio		
ம	EXPLORER ····	reddit_scrap_loop.py	🔹 linkedin_scrap.py 🗙	🔹 insta.py		
	✓ OPEN EDITORS	linkedin_scrap.py >				
0	reddit_scrap_loop.py	1 from linkedin_s	craper import Compan	y, actions		
$\sim$	🗙 💠 linkedin_scrap.py	2 from selenium i	mport webdriver			
0.	🜩 insta.py	4 from selenium.	ebdeiver import Chro	import Keys		
~ <u>F</u> ~	V IT7993	5 import mysql.co	nnector	ampore keys		
	> .vscode	6 import time				
>	chromedriver.exe	7				
~~~	Data scraper for reddit_it799	8 # let Python kr	www.e.using Chrome D	)river		
00	insta_testing1.py	9 driver = webdri	iver.Chrome()			
	🗇 insta.py	10 11 # must type you	un own linkedin email	and password		
_	linkedin_education.py	12 email=	Now Tinkedin enali	and passiona		
	🗇 linkedin_person.py	13 password	transforme pi			
	linkedin_scrap_loop.py	14				
	linkedin_scrap.py	15 # login use the	e info you provide ab	ove		
	linkedin_testing.py	16 actions.login(c	lriver, email, passwo	ord)		
	💷 linkedin2.xlsx	1/ 18 # Scrapping cor	00201/			
	💷 linkedin4.xlsx	19 url = "https://	/www.linkedin.com/com	panv/achmea-investment-m		
	🗇 person.py	20 company - Compa	my(linkedin_url-url,	name-True, specialties-		
	R_testing.py	21				
	reddit_scrap_loop.py	22 #Need Chrome to	wait 8 second for 1	linkedin to load before s		
	Reddit_scrap_with_loop_it79	23 time.sleep((8)				
	Reddit_scrap.py	24				
	Reddit_testing4.py	PROBLEMS OUTPUT DEBU	G CONSOLE TERMINAL			
	🕼 reddit.xlsx	Windows PowerShell				
_	💷 Reddit2.xlsx	Copyright (C) Microsoft	Corporation. All righ	ts reserved.		
8	Reddits_testing3.py					
	redditscore.xlsx	Try the new cross-platf	orm PowerShell https:/	/aka.ms/pscore6		
\$63	Redditscore1.xlsx	PS C:\Users\rituc\Webcr	awling\IT7993> []			
000	> OUTLINE					
Pyth	ion 3.9.2 64-bit ⊗ 0 ∆ 0					
	O Type here to search		H4 II 🥋 💳			

# Database

📉 MySQL Workbench						
☆ capstone ×						
File Edit View Query D	)ataba	se Server	Tools S	cripting Help		
8 8 6 6 6	6	a 🗗 🖸	₩ <u>₽</u>			
Navigator	r_sc	crap I_c	ompany $ imes$	I_education I_person i_comp	any r_score	
SCHEMAS 🚸		1 🖬   🕖	1 👰	🕛   🏡   📀 💿 🔞   Limit to 500	10 rows 🝷   🎪   🕩	9. N 7
Q Filter objects		1 • SELE	CT * FRO	M linkedin.l_company;		
▶ 🗐 instagram		2				
V 🗧 linkedin		3 • SELE	CT c.com	panyID, p.personID, c.company№	Name, c.companySize	, p.name, p.curre
		4 FROM	linkedi	n.l company c		
► I_company		5 JOIN	linkedi	n.l person p		
I_education		6 01	c.compa	nyName = p.company;		
I_person	<					
person				-	<del></del>	
468 VIEWS	Day				Mann Call Contract: 10	
Stored Procedures	Re	sult Grid   🏭	Filter	Rows:	Wrap Cell Content: 1A	
Stored Procedures	Res	companyID	ersonID	companyName	Wrap Cell Content: <u>‡A</u> companySize	name
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> </ul>	Res	companyID	PersonID	CompanyName John Hancock Investment Management Chies Accest Management Could the	Wrap Cell Content: <u>1A</u> companySize 501-1,000 employees	name Matthew Miskin CFA
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> </ul>	Re:	companyID 31 34	Filter	Rows: Export:	Wrap Cell Content: <u>‡A</u> companySize 501-1,000 employees 501-1,000 employees 10,001 employees	name Matthew Miskin CFA Qingze (David) Lin
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	Re:	companyID 31 34 35 26	Pilter personID 1 2 3	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         10,001 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	Res	sult Grid   ]] companyID 31 34 35 36 37	Filter personID 1 2 3 4	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal.
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	Res	sult Grid   ]] companyID 31 34 35 36 37 37	Pilter personID 1 2 3 4 5 6	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	▶ ■	sult Grid   ]] companyID 31 34 35 36 37 37 37 80	Filter personID 1 2 3 4 5 6 7	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	▶ ■	sult Grid   ]] companyID 31 34 35 36 37 37 80 80	Filter personID 1 2 3 4 5 6 7 8	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CEA
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul>	▶ ■	sult Grid   ]] companyID 31 34 35 36 37 37 80 80 80 39	Filter personID 1 2 3 4 5 6 7 8 11	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           201-500 employees         51-200 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul> Administration Schemas	▶ ■	sult Grid   ]] companyID 31 34 35 36 37 37 37 80 80 80 39 39	Filter personID 1 2 3 4 5 6 7 8 11 12	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames
Administration Schemas	▶	sult Grid   ]] companyID 31 34 35 36 37 37 37 80 80 80 39 39 40	Filter personID 1 2 3 4 5 6 7 8 11 12 13	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees           5,001-10,000 employees         500 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames Josã @ Carlos Justo
Administration Schemas	Re:	sult Grid   ]] companyID 31 34 35 36 37 37 37 80 80 80 80 39 39 40 40	Filter personID 1 2 3 4 5 6 7 8 11 12 13 14	Rows: Export:	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees           5,001-10,000 employees         5,001-10,000 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames José Carlos Justo Ignacio Calle Quartas
Administration Schemas Information No object selected	▶ ■	sult Grid       companyID 31 34 35 36 37 37 37 80 80 80 39 39 40 40 41	Filter personID 1 2 3 4 5 6 7 8 11 12 13 14 15	Rows:       Export: Ex	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees           5,001-10,000 employees         5,001-10,000 employees           201-500 employees         5,001-10,000 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames José Carlos Justo Ignacio Calle Cuartas Daniel P. Einpegan
Administration Schemas Information No object selected	Res	sult Grid   ]] companyID 31 34 35 36 37 36 37 37 80 80 80 39 39 39 40 40 40 41 41	Filter personID 1 2 3 4 5 6 7 8 11 12 13 14 15 16	Rows:       Export:         companyName         John Hancock Investment Management         China Asset Management Co., Ltd.         Swedbank         AMP Capital         Bridgewater Associates         Bridgewater Associates         Achmea Investment Management         Harris Associates         SURA Asset Management         SURA Asset Management         ClearBridge Investments         ClearBridge Investments	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees           5,001-10,000 employees         5,001-10,000 employees           201-500 employees         201-500 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames José Carlos Justo Ignacio Calle Cuartas Daniel P. Finnegan Aram Green
<ul> <li>Views</li> <li>Stored Procedures</li> <li>Functions</li> <li>reddit</li> <li>sakila</li> <li>sys</li> <li>world</li> </ul> Administration Schemas Information No object selected	Re:	sult Grid       companyID 31 34 35 36 37 37 37 80 80 39 39 40 40 41 41 41	<ul> <li>Filter</li> <li>personID</li> <li>1</li> <li>2</li> <li>3</li> <li>4</li> <li>5</li> <li>6</li> <li>7</li> <li>8</li> <li>11</li> <li>12</li> <li>13</li> <li>14</li> <li>15</li> <li>16</li> </ul>	Rows:       Export:         companyName         John Hancock Investment Management         China Asset Management Co., Ltd.         Swedbank         AMP Capital         Bridgewater Associates         Bridgewater Associates         Achmea Investment Management         Harris Associates         SURA Asset Management         SURA Asset Management         ClearBridge Investments         ClearBridge Investments	Wrap Cell Content:         IA           companySize         501-1,000 employees           501-1,000 employees         10,001 employees           10,001 employees         1,001-5,000 employees           1,001-5,000 employees         1,001-5,000 employees           201-500 employees         201-500 employees           51-200 employees         51-200 employees           5,001-10,000 employees         5,001-10,000 employees           201-500 employees         201-500 employees	name Matthew Miskin CFA Qingze (David) Lin Tomas Bakutis, CFA Claire (Staniforth) Tal. Daniel Hochman Kevin Brennan Lucas Bouwhuis Amir Saleem, CFA Ben Nielson Jeremy Thames José Carlos Justo Ignacio Calle Cuartas Daniel P. Finnegan Aram Green

# Visualization





					-	Ć	כ	×
						۵		
								~
rrent	t position, p.from da	te, p.to d	date, p.lo	ocation				
			, p					
								. ×
								_
	current_position	from_date	to_date	location				^
1	Company Name John Han	NULL	NULL	NULL				
	Company Name China Ass	Jul 2004	May 2011	Shanghai				
1	Fund Manager	11 yrs		2 yrs 8 mos				
al	Company Name AMP Capital	NULL	NULL	NULL				
	Company Name Bridgewat	Jan 2010	Mar 2010	Westport, CT				
	Company Name Bridgewat	Jun 2001	Aug 2001	Washington, District Of Col				
	Senior Portfolio Manager	NULL	NULL	NULL				
	Senior Portfolio Manager	NULL	NULL	NULL				
	Sr Investment Analyst	Mar 1998	Jan 2000	Greater Los Angeles Area				
	Investment Research Ana	NULL	NULL	NULL				
)	Equity Portfolio Manager	4 yrs	mos	2 yrs				
as	CEO	NULL	NULL	NULL				
	Company Name ClearBrid	Sep 2007	Jul 2011	Stamford, CT				
	Portfolio Manager	NULL	NULL	NULL				

Read Only

	companySize.Min	companySize.Max	^
ement, LLC	51	200	
vestments	5001	10000	
ment	201	500	
5	51	200	
lanagement	201	\$00	
	51	200	
	201	500	
	11	50	
nent	1001	5000	
nent (formerly TKP investments)	1001	5000	
	1002	2000	
quo	1001	5000	
Inement	11	50	
	11	50	
	\$1	200	
	11	50	
	1001	5000	
	10001		
1	1001	5000	
-	11	50	
stments	1001	5000	
	1001	5000	
	1001	5000	
	201	500	
	1001	5000	
ment inc.	1001	5000	
ant .	501	1000	
	51	200	
moration	1001	5000	
mited Partnershin	201	500	
	444	244	~
	448370	639890	
List of Company by Size	Company Dach	hoard t	





### 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**

Janell Westmoreland jwestmo6@students.kennesaw.edu Kajal Vaghani kvaghani@students.kennesaw.edu **Ritu Choudhary** rchoudh2@students.kennesaw.edu Vy Duong vduong.it7@gmail.com Nyong Nkereuwem 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 <a href="https://praw.readthedocs.io/en/latest/">https://praw.readthedocs.io/en/latest/</a> Greening, C. (2021). instascrape. Retrieved from https://chrisgreening.github.io/instascrape/ Sham, J. (2021, April 10). Scrapes user data from LinkedIn. Retrieved

from <a href="https://pypi.org/project/linkedin-scraper/">https://pypi.org/project/linkedin-scraper/</a>