UC-54 IT Alumni Database ETL System

Abstract

The IT department at KSU maintains an alumni database which stores graduates' career information. In the past, career information about alumni was manually extracted from LinkedIn.com and then loaded into a relational database which is a very labor-intensive process. In this project, we worked with a graduate student to create an automated extract, transform and load (ETL) system for the IT alumni database. A web crawler capable of extracting information from LinkedIn was used for this purpose.

Introduction

For this project, we used a web crawler that was provided by Hang, the graduate student we worked with, to extract information from LinkedIn. Originally, this information was stored in txt files. Later, it was realized that this information needed to be stored in JSON format.

After storing the information in JSON format, the challenging part was trying to interpret the data and store it in a way that SQL Server Management Studio would accept.

Research Question(s)

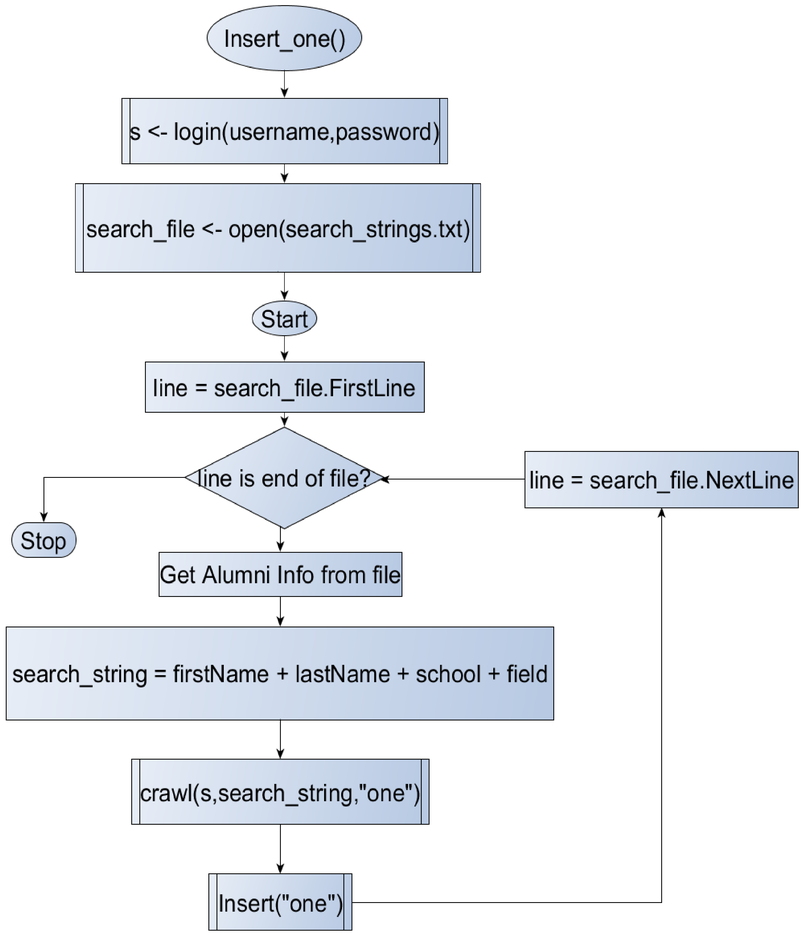
- Can we successfully pull career information about Kennesaw State graduates and store the information in a relational database?

Materials and Methods

Snippet of the Gantt chart we followed to complete the project.

					Milestone #1				Milestone #2			
Deliverable	Tasks	Complete%	Current Statu	Assigned To	01/28	02/04	02/11	02/18	02/25	03/04	03/11	03/18
Complete Database	Choose database enviornment	0%		Zack,Vy,Ricky, Desiree	5							
	Install the database server	0%		Zack,Vy,Ricky, Desiree								
	ERD conceptual design	0%		Vy		5						
	Logical design	0%		Desiree			4					
	Learning SQL Server	0%		Zack,Vy,Ricky, Desiree		8	8					
	Create the database schema (DDL)	0%		Vy			4	4				
	Execute the DDL on the database server	0%		Vy				4				
	Testing	0%		Zack,Vy,Ricky, Desiree			10	10				
Insert Record Scripting	Learning Python	0%		Zack,Vy,Ricky, Desiree	20	12	12	12	8			
	Be able to connect to database using Python	0%		Vy					4			
	Collect and parse data from excel spreadsheet	0%		Zack					8	8	6	
	Create logic to check if record exists in the database	0%		Desiree						6		
	Work on logic to insert new record into the database	0%		Ricky						20	20	20
	Test by using existing and new KSU Alumni Accounts	0%		Zack, Vy, Ricky, Desiree							10	10
Update Record Scripting	Research optimal amount of times to use web crawler	0%		Zack,Vy,Ricky, Desiree								
	Create script that will run web crawler	0%		Zack,Vy,Ricky, Desiree								
	Work on logic to check which items need updating	0%		Zack,Vy,Ricky, Desiree								
	Work on logic that will update records automatically	0%		Zack,Vy,Ricky, Desiree								
	Test by making changes to Linkedin accounts	0%		Zack, Vy, Ricky, Desiree								

Results




```
SELECT * FROM Jobs;
```

100 %

Results | Messages

	JOB_ID	TITLE	COMPANY	STARTDATE	ENDDATE	JOB_CATEGORY	ALUMNI_ID
1	543	Transcriptionist	Rev.com	2016-05-01	2018-04-01	NULL	ACoAABphCyMBLH61q2_vE-28_vYg6MYqaAGtXpw
2	544	Sales Associate	Sherlock's	2016-06-01	NULL	NULL	ACoAAB77dGoBh15bSXEUXv-gx1aKQHpM6fWUGx0
3	545	Valet Attendant	USA Parking System	2012-03-01	2016-03-01	NULL	ACoAAB77dGoBh15bSXEUXv-gx1aKQHpM6fWUGx0

Conclusions

In conclusion, we were able to pull information from LinkedIn and store the info into a relational database. There were some problems however. LinkedIn has different tier levels for your account. Sine we tested our program using a basic level account, we were unable to pull all the information that we wanted. Also, we found that making too many connection requests within a short amount of time can get your account temporarily suspended.

Acknowledgments

Hang Yu: yuhang517@gmail.com

Contact Information

rparks13@students.Kennesaw.edu Ricky Parks
vduong2@students.Kennesaw.edu Vy Duong
zdowning@students.Kennesaw.edu Zack Downing
dsmokes@students.Kennesaw.edu Desiree Smokes
lli13@Kennesaw.edu Lei Li (Owner)

References

https://www.linkedin.com/

Author(s) Ricky Parks, Vy Duong, Zack Downing, Desiree Smokes
Advisor(s) Dr. Ming Yang