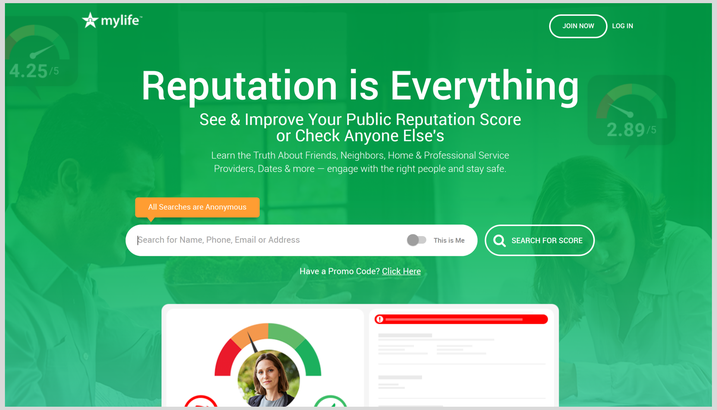
WEB SCRAPING AND DATABASE INTERACTION

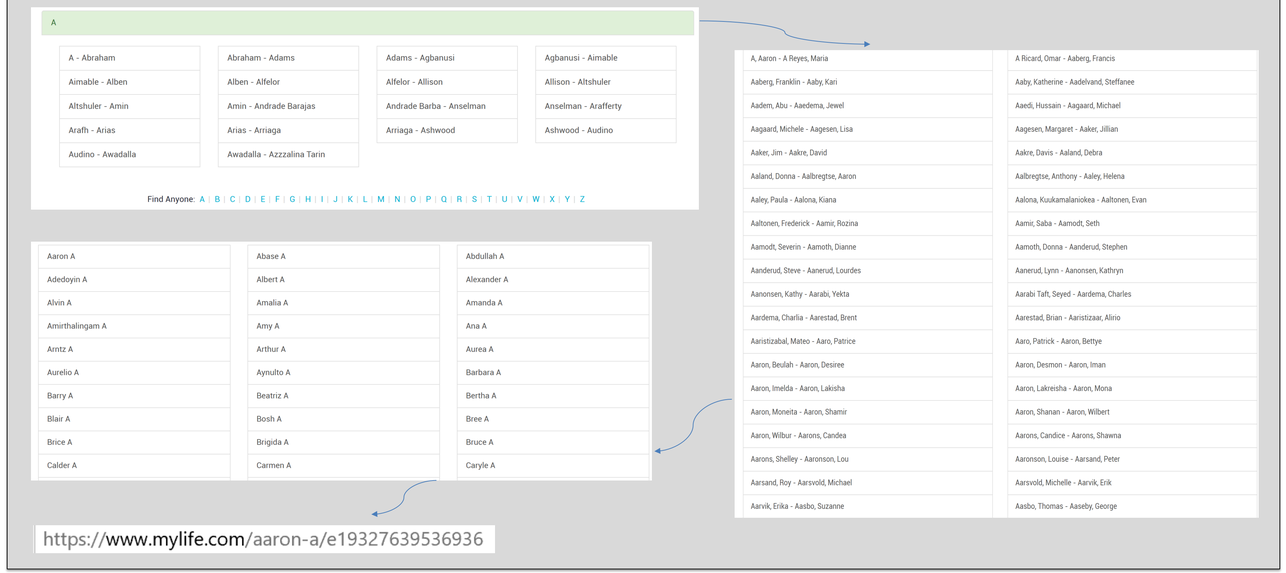
By Luke Nitschke, Evan Shaw, Raunak Srivastava

INTRODUCTION

- TMap: a small company who uses technology and targeted marketing to identify and engage qualified employment at scale.
- Develop a prototype that scrapes websites and databases (mylife)
- Obtain important, relevant information about various working individuals with the intention of matching them with employers seeking to fill positions.

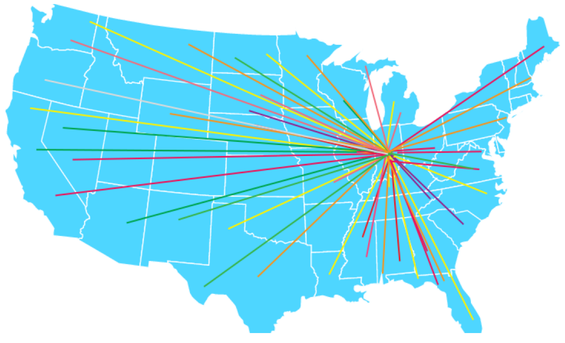


METHODS: SEQUENCE OF PAGES THAT WE SCRAPED



Used MongoDB to store results.

TMap is trying to bring talent with connections to Indiana back to Indiana.



CHALLENGES

1. Coming into this project, no one on our team knew any of the tools that we used in this project (JavaScript, Node.js, MongoDB), but we picked up the skills through online tutorials and articles provided by our mentor Reuben and learning through practice.
2. Our original web scraping prototype led us to get blocked by the website, since we made so many calls to the webpages. To counter this, we utilized a proxy service to route our requests through their servers and use an alternate IP address to continue to collect data.
3. We needed to merge our coding style with Reuben's coding style so that he could implement our code into his existing algorithm.

METHODS

- This project has two parts:

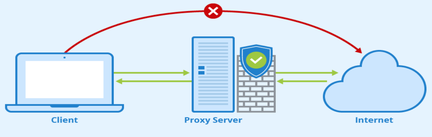
1. Recursively scrape the mylife.com directories to find and store all the unique identifiers corresponding to each person in the database.

Each identifier is formatted like this: https://www.mylife.com/aaron-a/e19327639536936

2. Use the list of identifiers to scrape and store relevant information of potential employees and their relationships to other potential employees

We used Node.js as the primary tool to scrape data from MyLife.com

We utilized a proxy service to bypass getting blocked by the website.



CONCLUSION

- Logic within our code can be tweaked and refined to scrape other databases as well
- Now able to work with new software that we could not before
- Solid knowledge base will be helpful with future projects and other educational endeavors

ACKNOWLEDGEMENTS

We'd like to thank Reuben Wilson for all of his support during this project. He was always reachable through Slack, and provided us with the proper guidance and encouragement during our weekly meetings. We would also like to thank Evan Hock for presenting this project opportunity to us.