Jobs Database Progress Report

YUNAN SHAO
1818832

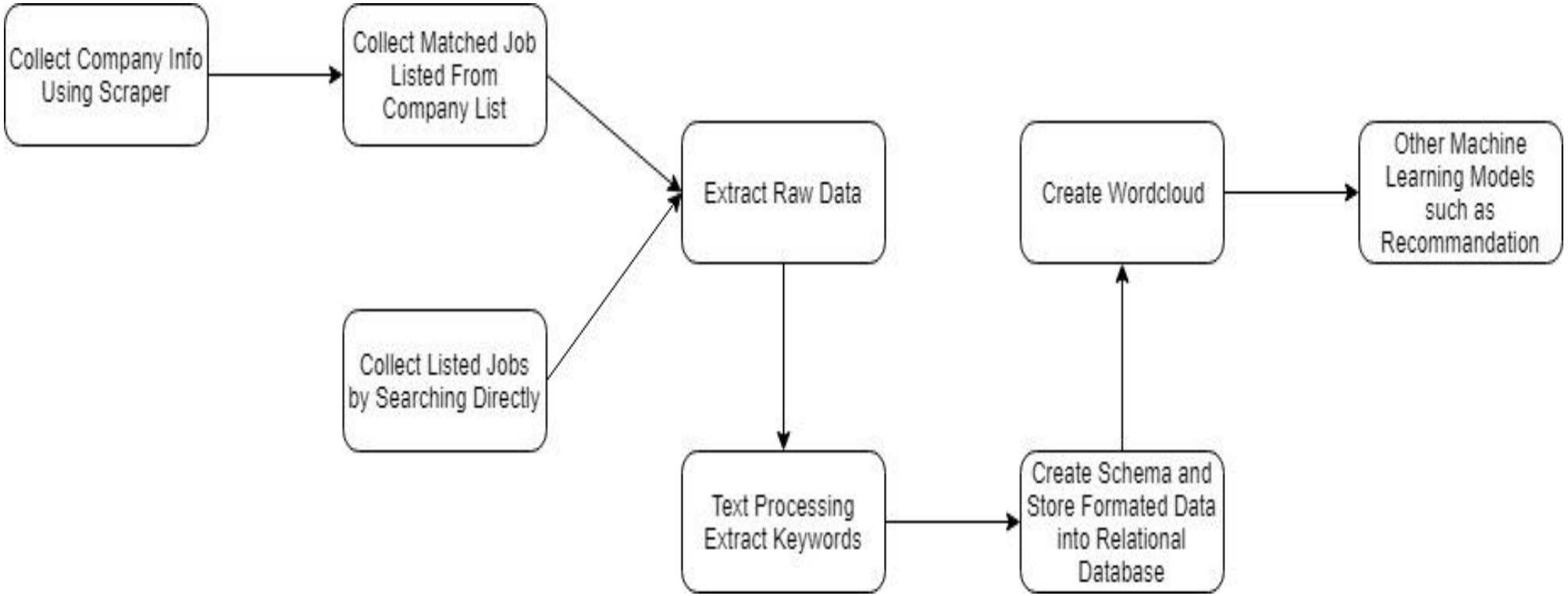
Jobs Database - AI Skunkworks Project Goals

- Collecting data from company websites, media pages or other websites such as Indeed, LinkedIn, Glassdoor etc. and store into relational database.
- Matching the job seekers with the right employers and second, provide guidance to aspiring job seekers on the skills that are in demand so that they can build them to stay relevant in the job market.

Tasks
(From Project Description)

- Collect company website and media page links
- Extract content/data from collected data
- Store data into relational database
- Creating word cloud based on the collected data

Current Work Flow



Task Check List

- Scraper for getting company list – Complete
- Scraper for social media page – Skipped
- Scraper for getting job list from company list – Complete
- Scraper for getting job list by directly searching – Complete
- Design schema and store data in to relational database – In Progress
- Text processing & Word cloud – In Progress
- Try other machine learning model – Not Started

Current Status & Problems

- Selenium is used for all scrapers
- Media page links don't have enough data related to jobs and skills
- Directly searched jobs work better
- Current process of text processing:
  - Tokenize the job description
  - Remove stop words
  - Counting skill related terms with manually created dictionaries

Remaining Tasks

- Data Storing – Currently using csv to store collected data, should be easy after I decide what fields I need to create or keep
- Text Processing – Currently counting words with manually created dictionaries. Trying to use pre-defined vocabulary dictionaries with NLP to collect all the tech terms.
- Other machine learning models
- Modify scrapers for more websites – Current data source: Angel.co, Glassdoor.com
- Take a look at the social media pages again, maybe there are some useful contents but need different method to process
- Code cleaning, optimization
- Paper, Report, etc.