AI–Powered Job Recommendation Engine

Marcel Gwerder, Andrew Jo, Orestis Kaitezidis, Kunle Lawal, Anubhav Rana, Laura Silva Jetter

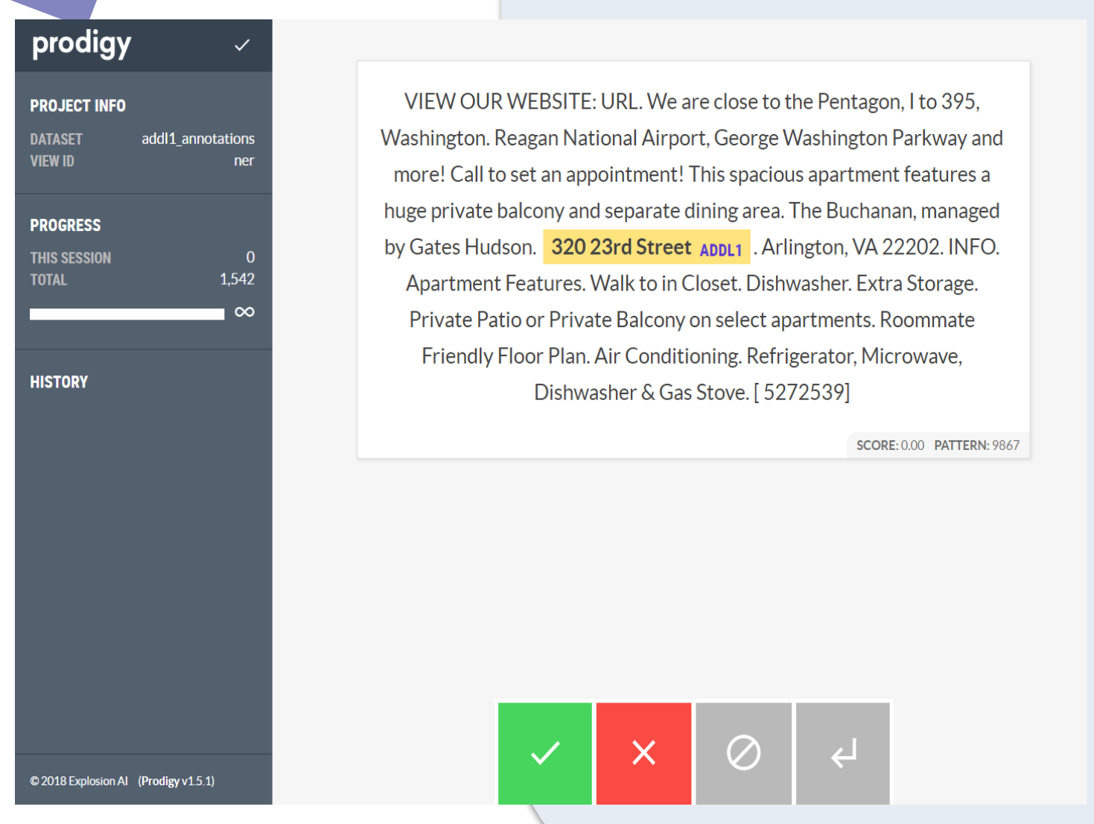
MOTIVATION

- Currently, job searching is a time consuming task where much manual input is needed and job titles can be misleading compared to the skills and experience required.
- As master level students seeking job opportunities, this problem affects us directly and for this reason a job recommendation engine platform was developed.
- This platform automates the job search by analyzing a resume and matching it to job postings to provide personalized job suggestions that match the user’s level of degree, skills and work experience. As a result, it provides more relevant and a wider range of results.

APPROCHES

NLP

- In the modeling stage, Natural Language Processing was used in order to identify and extract named entities in the job postings and resumes such as the skills, degree and years of experience.


prodigy
✓
PROJECT INFO
DATASET addl1_annotations
VIEW ID ner
PROGRESS
THIS SESSION 0
TOTAL 1,542
∞
HISTORY
© 2018 Explosion AI (Prodigy v1.5.1)
SCORE: 0.00 PATTERN: 9867
✓
✕
⊘
↵

VIEW OUR WEBSITE: URL. We are close to the Pentagon, I to 395, Washington. Reagan National Airport, George Washington Parkway and more! Call to set an appointment! This spacious apartment features a huge private balcony and separate dining area. The Buchanan, managed by Gates Hudson. 320 23rd Street ADDL1 . Arlington, VA 22202. INFO. Apartment Features. Walk to in Closet. Dishwasher. Extra Storage. Private Patio or Private Balcony on select apartments. Roommate Friendly Floor Plan. Air Conditioning. Refrigerator, Microwave, Dishwasher & Gas Stove. [5272539]

- Prodigy, an annotation tool powered by active learning, was used where the model was trained with a few hundred labeled examples which were manually annotated.
- The model was improved over the initial manual annotations by teaching the model through a accept/reject procedure which improves the model accuracy.

Data Processing

- The feature extraction code was ran on the resumes, and as part of the data preprocessing step on all job-postings.
- In order to have fast access to the data for our user-facing application, the preprocessed data will be stored in MongoDB database platform.

Matching

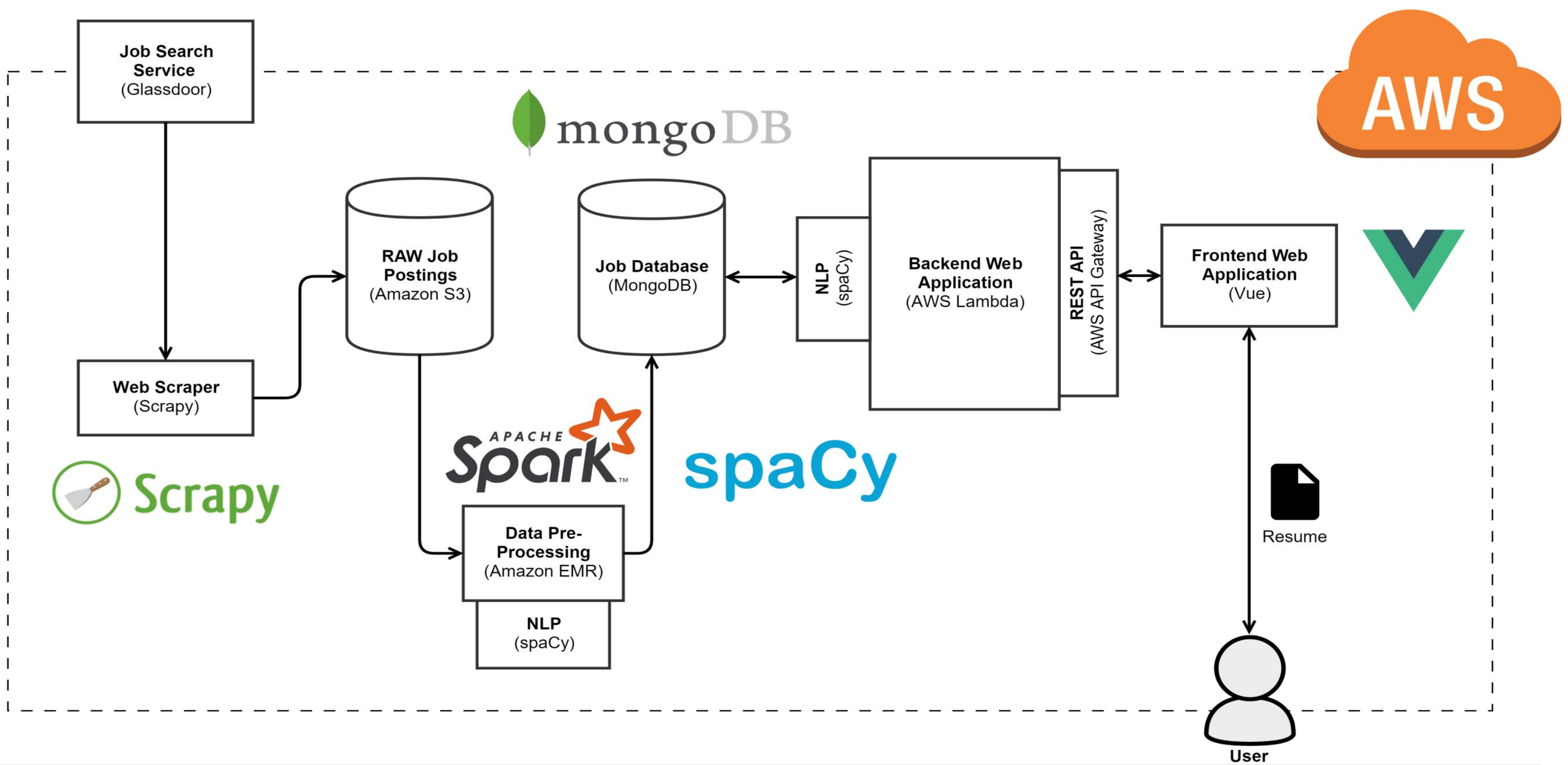
- To reduce the number of jobs that need to be scanned, an initial filtering is done using the “Skills Index” to find jobs with at least one matching skill.
- The education levels and experience for each job are assigned integer values and these values are mapped to the respective level for both, jobs and resumes. The difference in these values are used to compute the similarity by:
    - 1 - difference(Levels)max(Level)-min(Level).
- The Jaccard index for the skills is calculated and we experiment with the weighting on the different entity similarities when calculating final similarity.

DATA

- Glassdoor is partially built as a single page application (SPA) and uses an internal JSON API to build the detail view of each job posting.
- The project team uses that internal API to get detailed structured data for the job postings by simply incrementing an auto increment ID in the URL.
- While Glassdoor block several AWS IPs it does not seem to have a strong protection against scraping. Several thousand requests per second from a non AWS IP work just fine.
- The data is temporal and very structured, however the part primarily used by the project team, the job description, is unstructured text.
- The following data was collected:
    - About 2 million job postings
    - 14GB of JSON documents

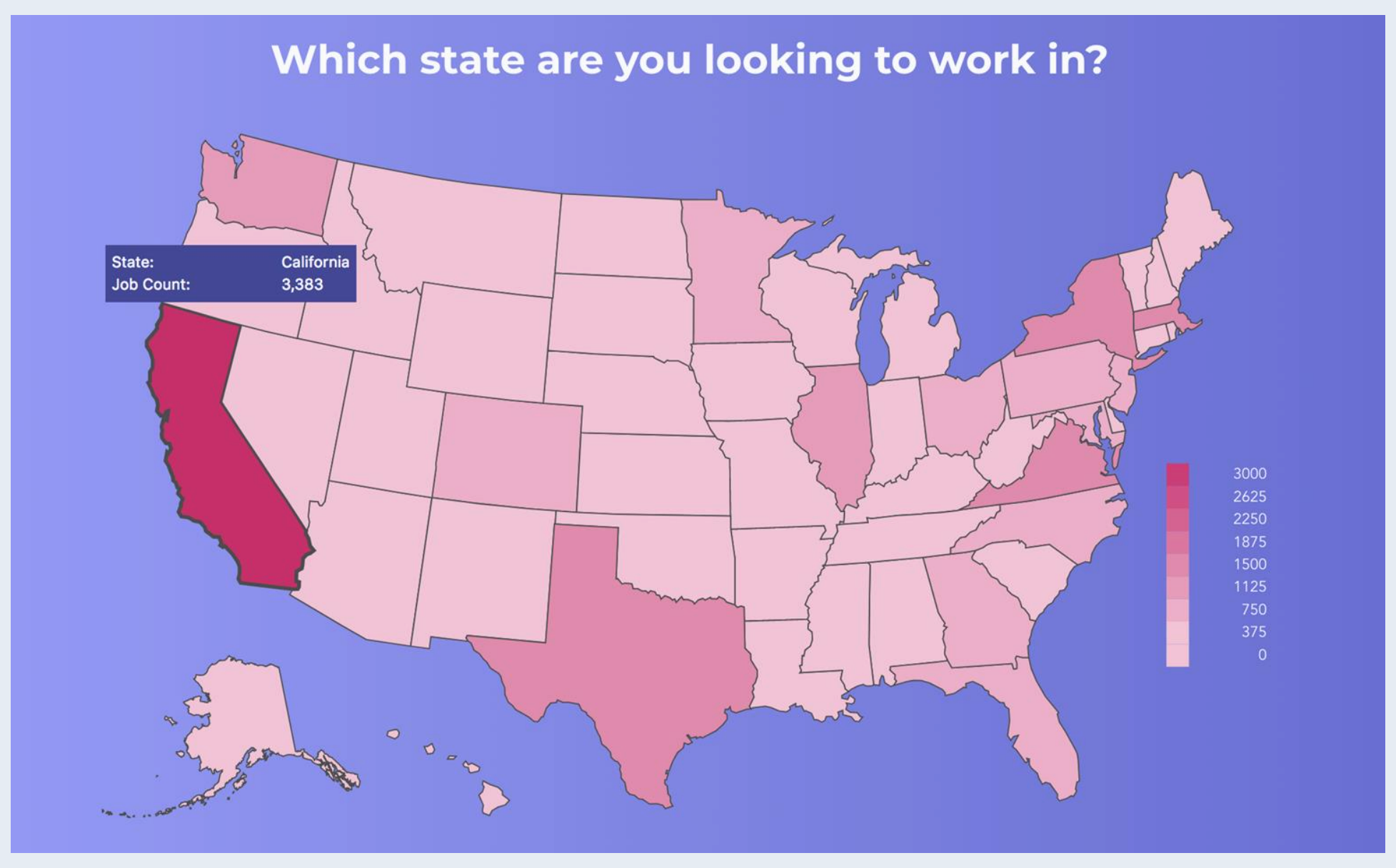
ARCHITECTURE

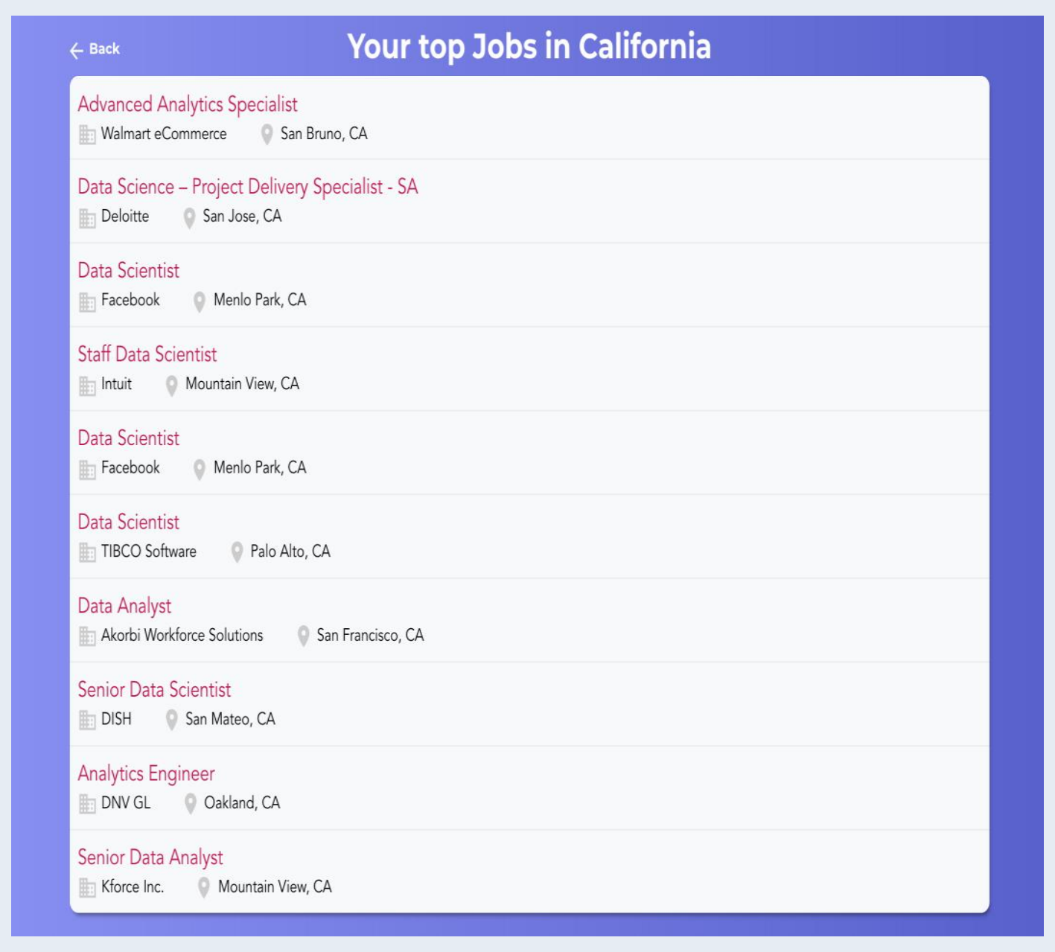
- The entire architecture is built in the AWS cloud based on various managed and unmanaged services. Job postings are scraped from Glassdoor using Scrapy and stored in JSON lines format on Amazon S3.
- In order to query the data in a performant manner, it is preprocessed using Apache Spark on Amazon EMR. During the preprocessing step, the data is cleaned up and the natural language processing tasks are run on the job description in order to extract the features needed.
- The resulting data is loaded into a MongoDB instance running on Amazon EC2. The data is exposed to the frontend application through an AWS Lambda function and AWS API Gateway.
- At the end of the pipeline, the user accesses a Vue web application through Amazon CloudFront.


Job Search Service (Glassdoor)
mongoDB
AWS
Web Scraper (Scrapy)
RAW Job Postings (Amazon S3)
Job Database (MongoDB)
NLP (spaCy)
Backend Web Application (AWS Lambda)
REST API (AWS API Gateway)
Frontend Web Application (Vue)
Scrapy
Apache Spark
spaCy
Data Pre-Processing (Amazon EMR)
NLP (spaCy)
Resume
User

VISUALIZATION

Explore jobs that match your skills





Your top Jobs in California

Advanced Analytics Specialist
Walmart eCommerce — San Bruno, CA

Data Science – Project Delivery Specialist - SA
Deloitte — San Jose, CA

Data Scientist
Facebook — Menlo Park, CA

Staff Data Scientist
Intuit — Mountain View, CA

Data Scientist
Facebook — Menlo Park, CA

Data Scientist
TIBCO Software — Palo Alto, CA

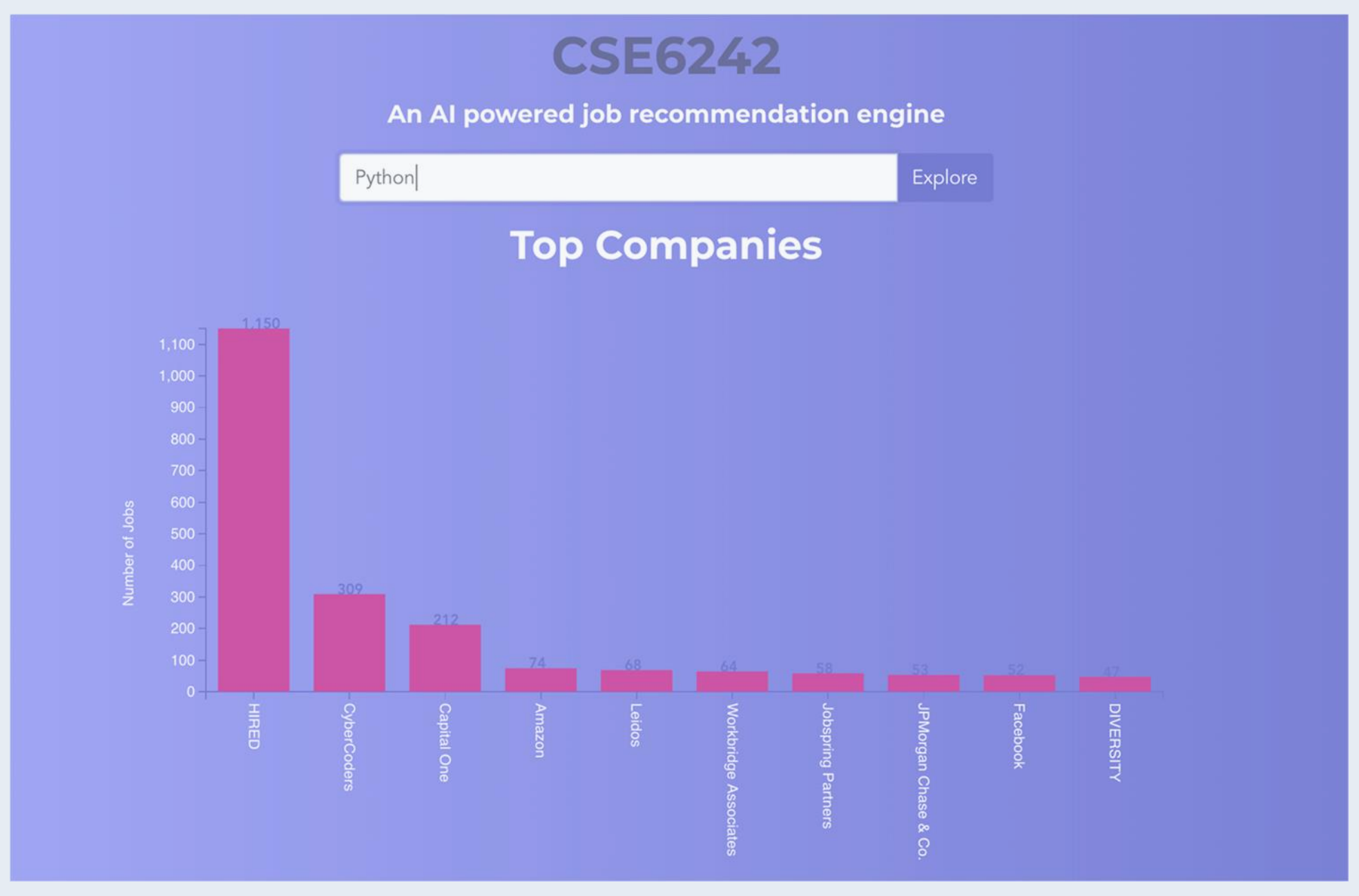
Data Analyst
Akorbi Workforce Solutions — San Francisco, CA

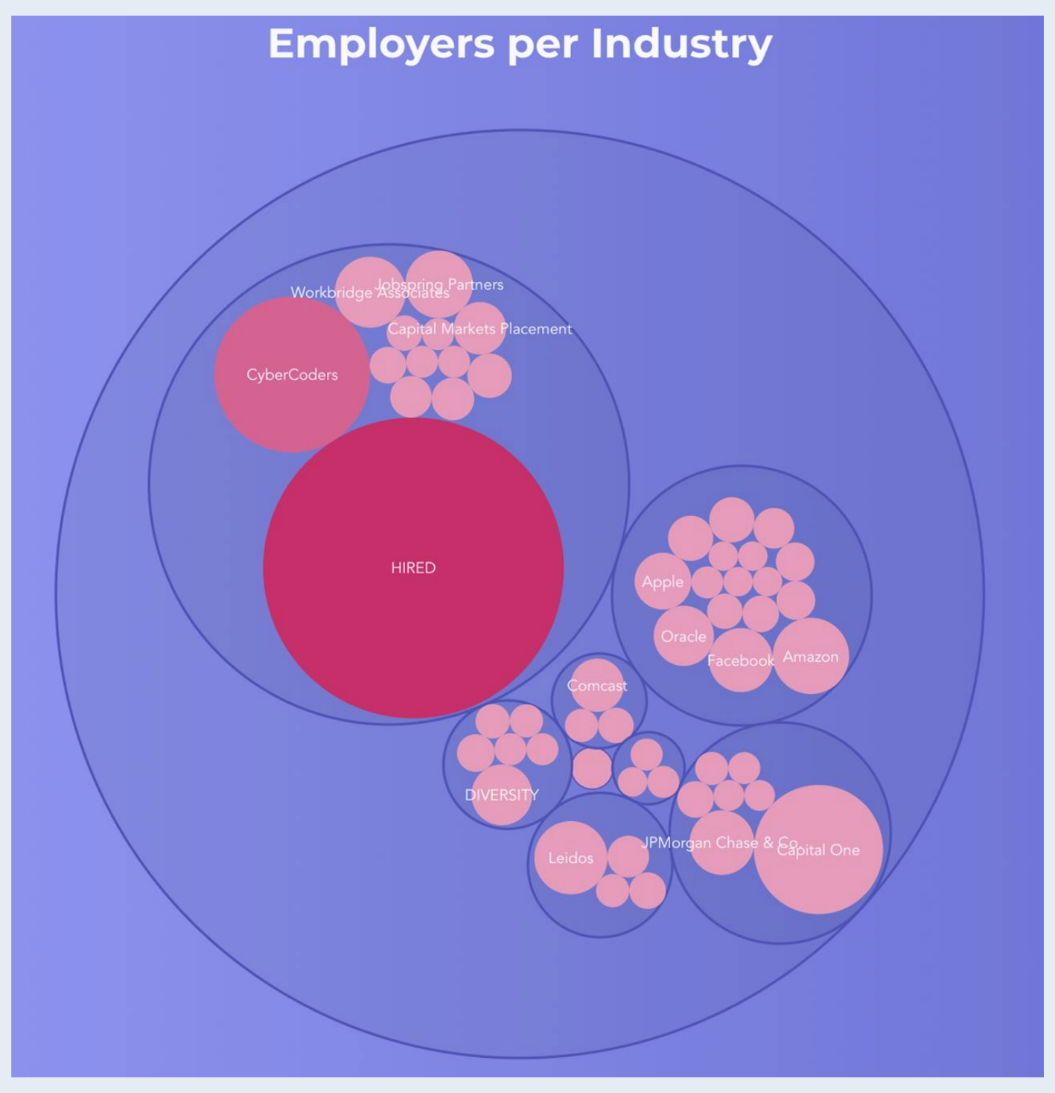
Senior Data Scientist
DISH — San Mateo, CA

Analytics Engineer
DNV GL — Oakland, CA

Senior Data Analyst
Kforce Inc. — Mountain View, CA

Explore jobs through a specific skill





EVALUATION

- A test dataset is used to evaluate the extracted entities and a confusion matrix to calculate performance measures.
- To evaluate the quality of our job recommendations, two phases of survey are conducted to our classmates. In the first phase of the survey, we evaluate users’ satisfaction with the current existing methods in job searching. We then gauge their perception of the application in the second phase, and compare the differences between the two phases.