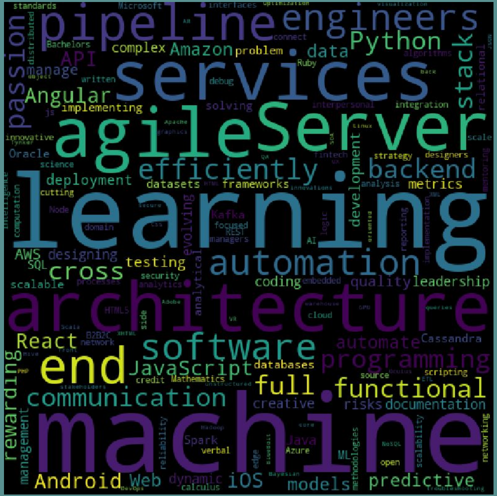
Skills and Jobs Match using NLP

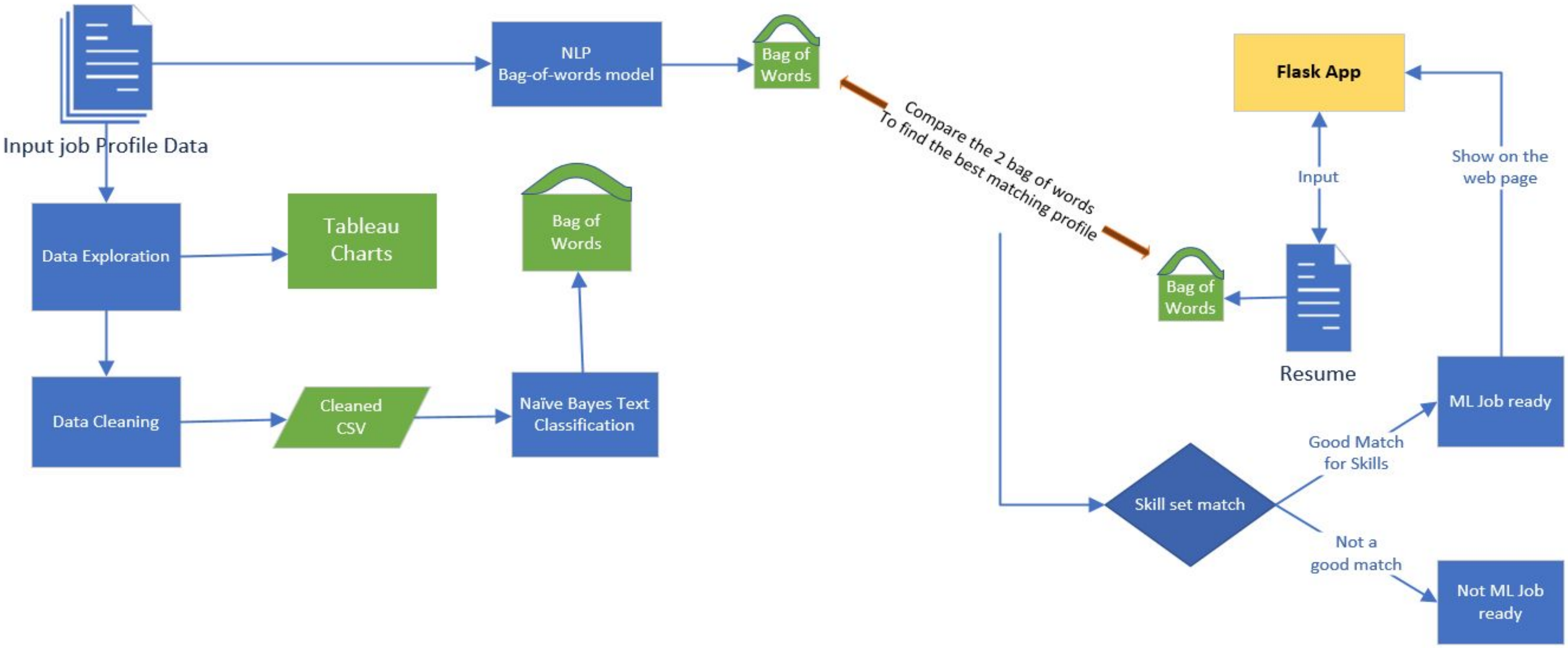
Contributors: Bhumika Sharma
Cong Tran
Denis Cohen
Chris Janssen
Imran Bawany
Vijay Gangaprasad



Goal of Project

1. Help Job seekers to find how their skills stack up against the most sought after Machine Learning skills and tools and also recommend job postings that matches their skills reflected in their resume.
2. We focused on Machine Learning type job postings as the basis of the application but the functionality can extend to all types of postings.
3. We used NLP techniques to build our matching process with machine learning libraries in Python.

NLP Text Classification



Data

Source

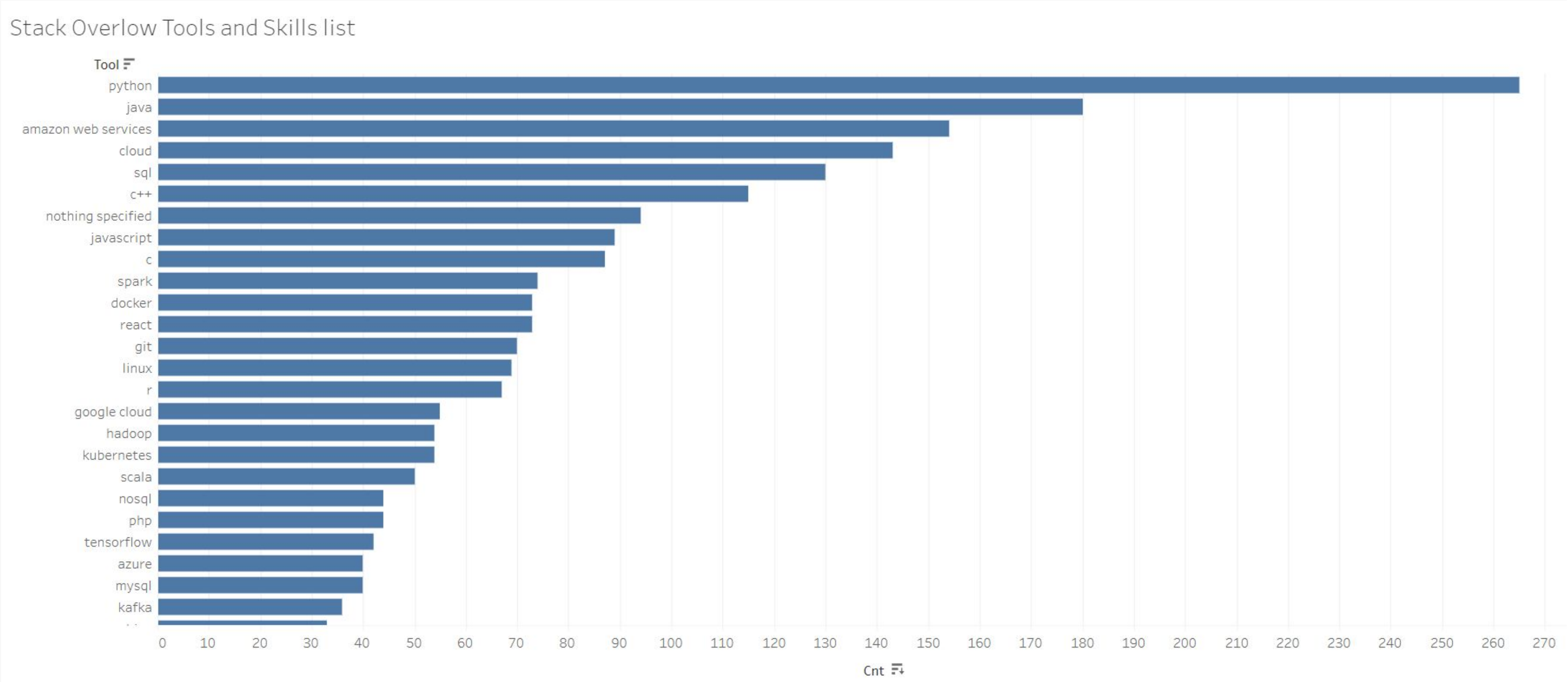
Retrieved data set of job postings from Stack Overflow using the RSS Feed downloaded as an XML file and converted to a CSV.

Exploration and Cleansing

Used Tableau to build chart to visualize the data. Visualizations included Job Title by Company, Job count by State, and Job count by Company. Data cleaning is done using Python libraries like Pandas.

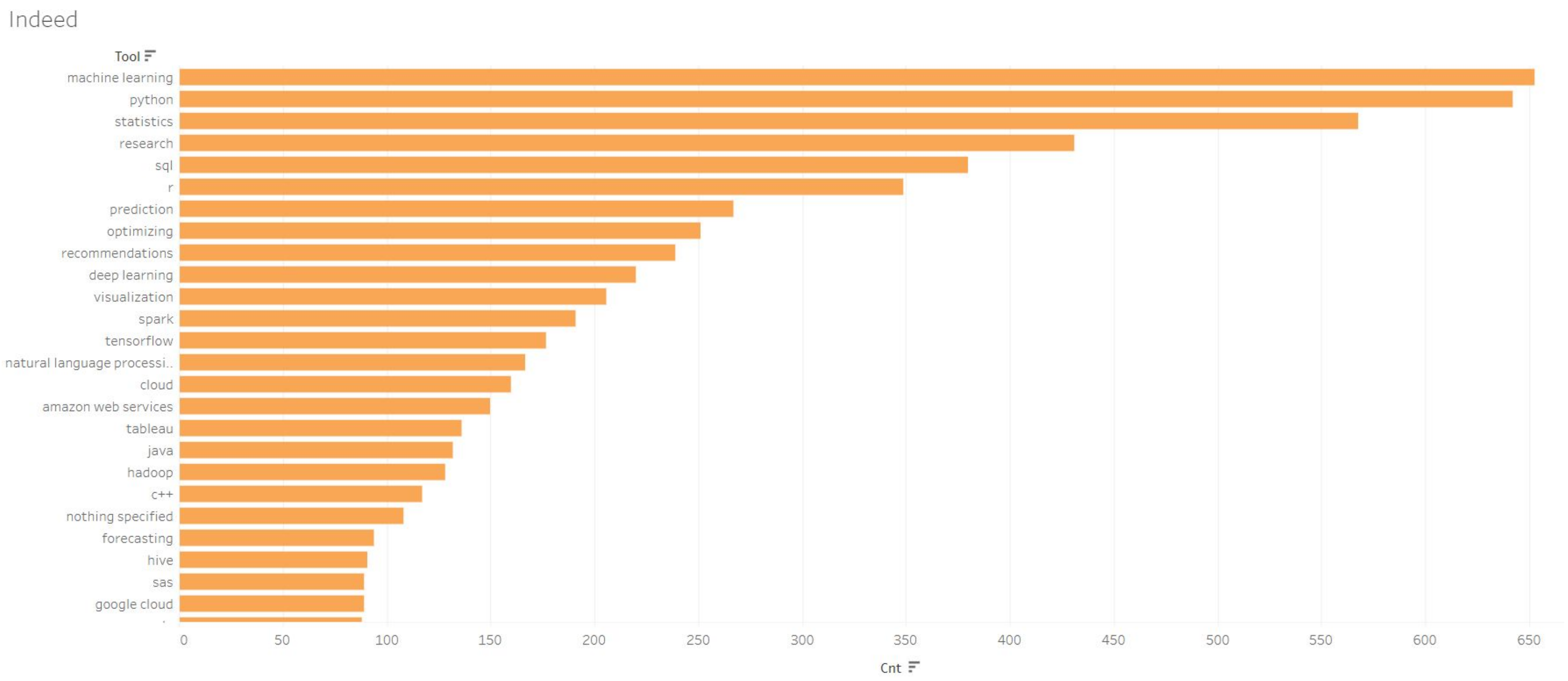
Issues

The biggest issue was finding a dataset of current job postings. Many of the job posting websites (Indeed, Monster, etc) did not have an accessible API.



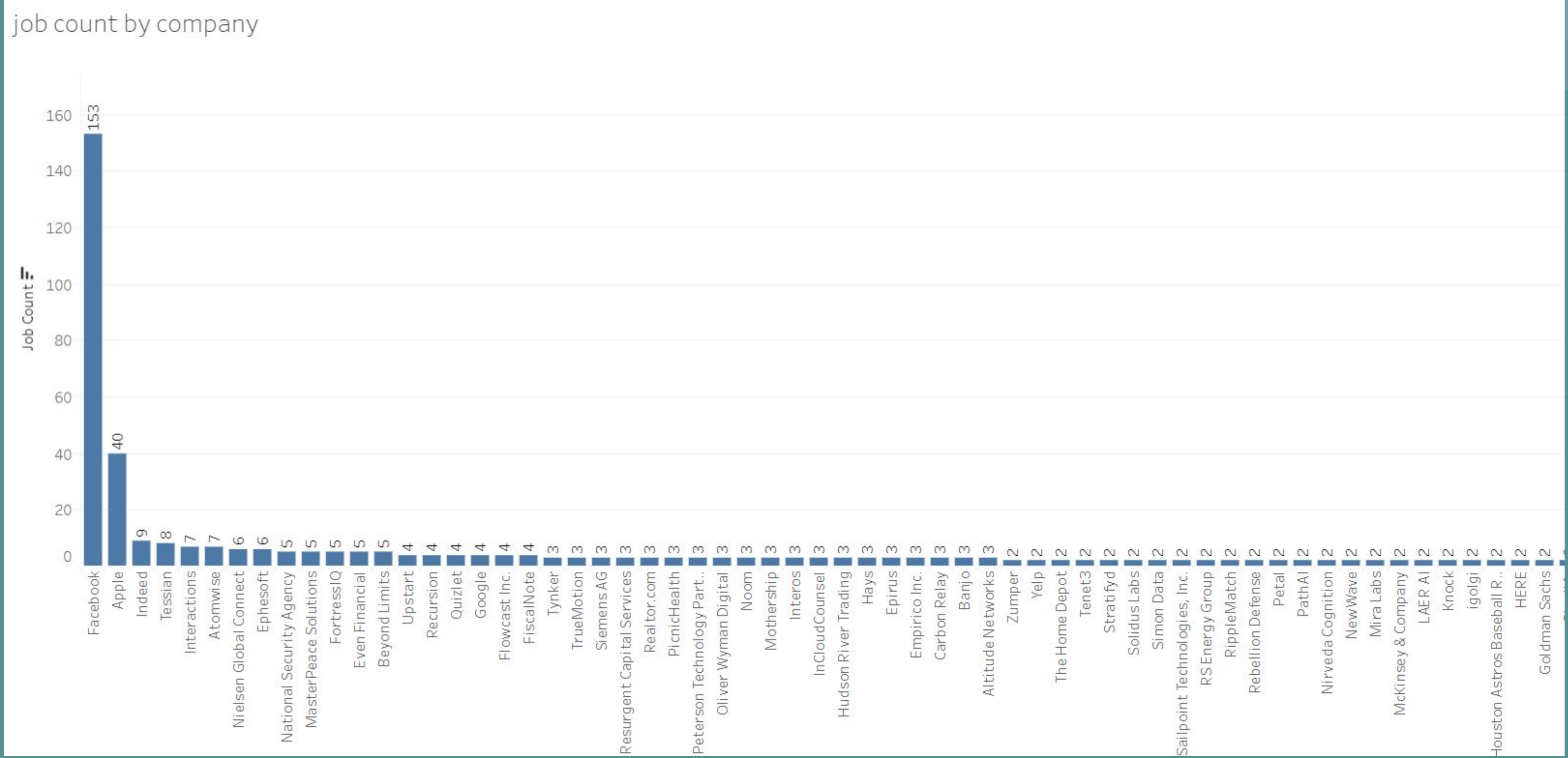
Stack Overlow Tools and Skills list

https://public.tableau.com/profile/vijaybabu.gangaprasad#!/vizhome/StackOverflowToolsandSkills/Sheet1

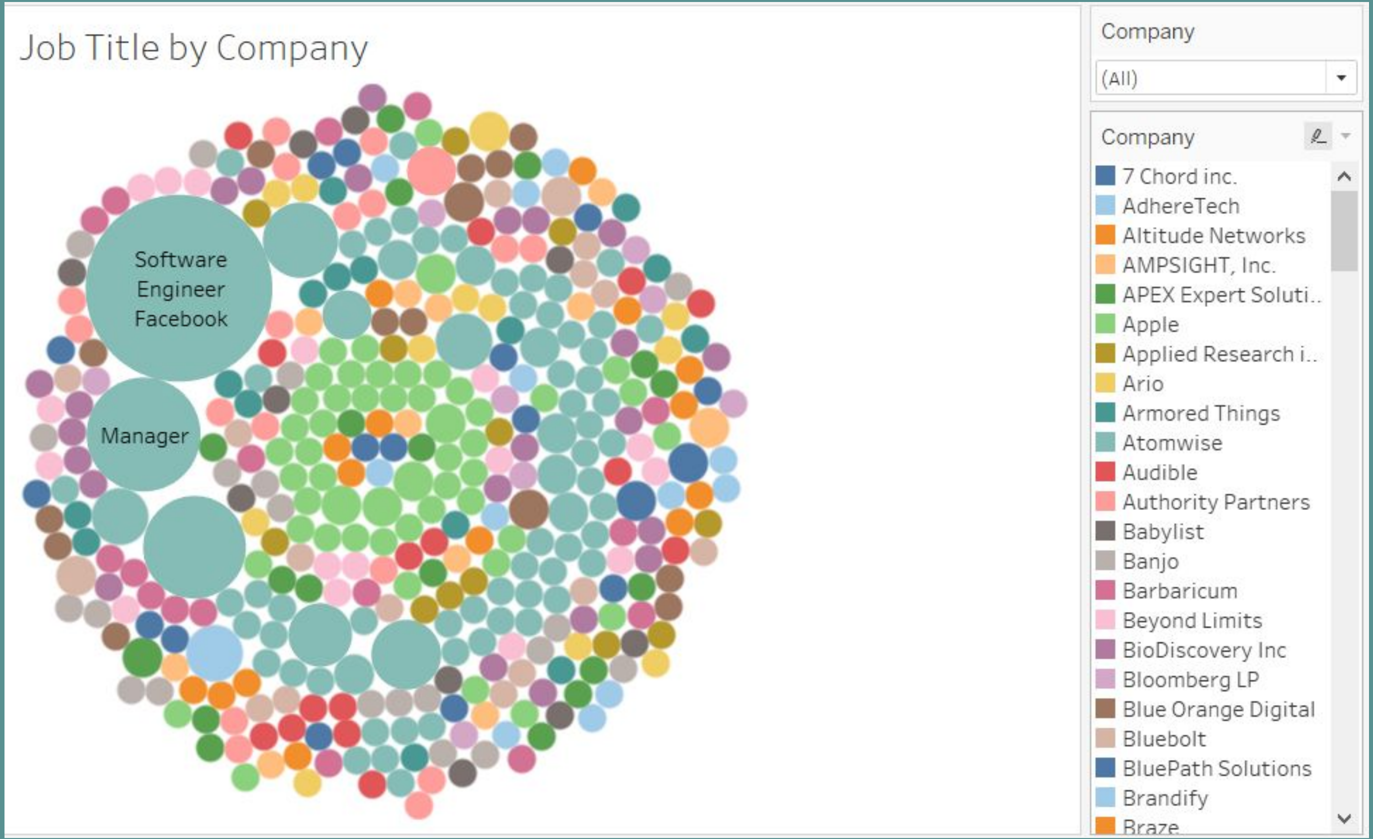


https://public.tableau.com/profile/vijaybabu.gangaprasad#!/vizhome/IndeedSkillsandtools/Indeed

Visualizations: Job Count by Company



Visualizations: Job Title by Company



Job Title by Company

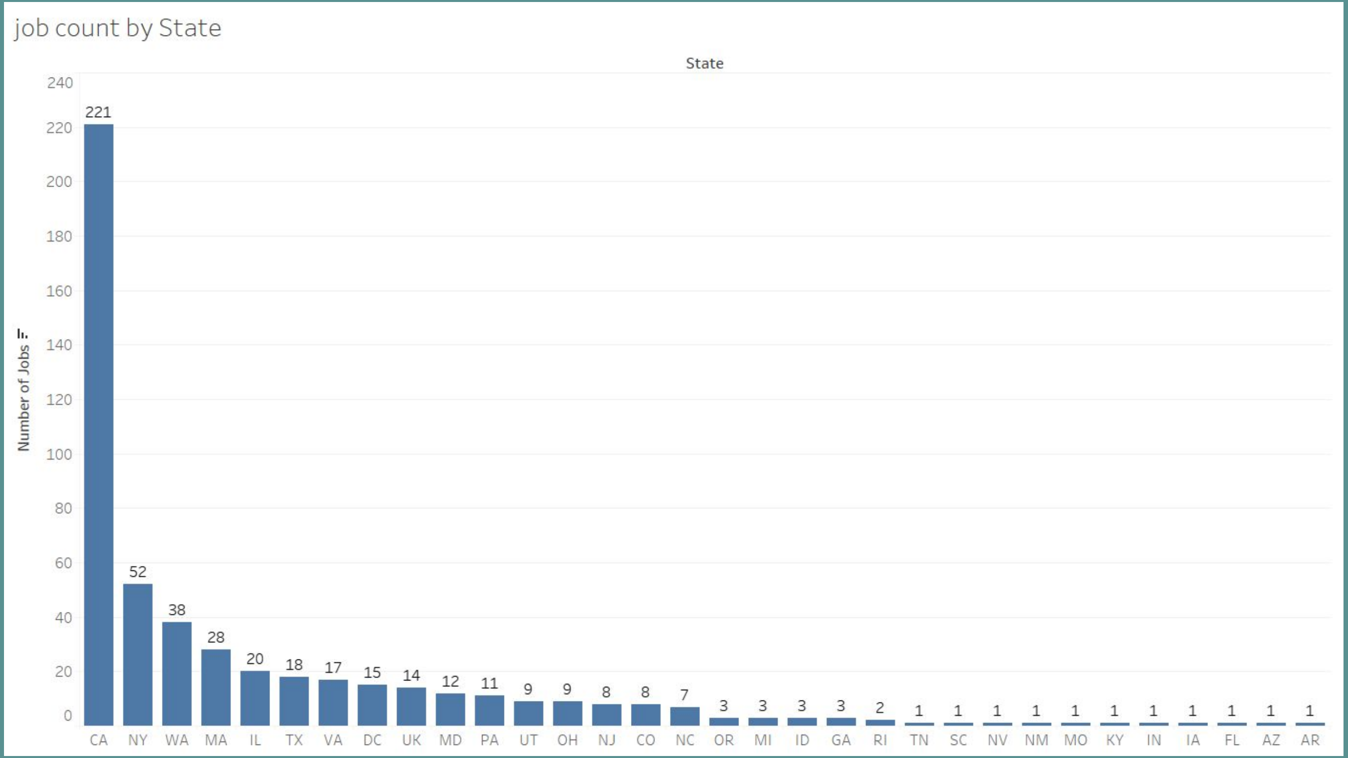
Company

(All)

Company

- 🟦 7 Chord inc.
- 🟦 AdhereTech
- 🟧 Altitude Networks
- 🟧 AMPSIGHT, Inc.
- 🟩 APEX Expert Soluti..
- 🟩 Apple
- 🟨 Applied Research i..
- 🟨 Ario
- 🟩 Armored Things
- 🟩 Atomwise
- 🟥 Audible
- 🟥 Authority Partners
- 🟫 Babylist
- ⬜ Banjo
- 🟥 Barbaricum
- 🟪 Beyond Limits
- 🟪 BioDiscovery Inc
- 🟪 Bloomberg LP
- 🟫 Blue Orange Digital
- 🟫 Bluebolt
- 🟦 BluePath Solutions
- 🟦 Brandify
- 🟧 Braze

Visualizations: Job Count by State



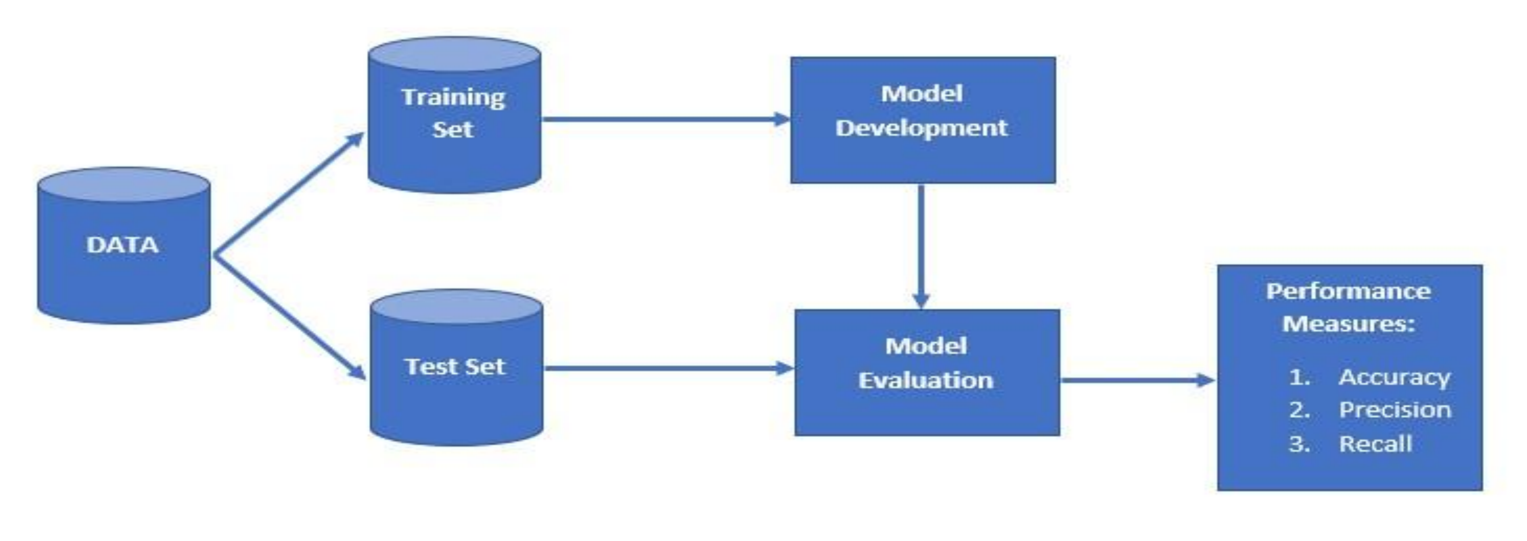
Naive Bayes Classifier for Text Classification

Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction.Naive Bayes classifiers are linear classifiers that are known for being simple yet very efficient.
In our case,the algorithm is applied to classify the job listing as either 'Machine Learning' or not.

Steps:

- Data collection :StackOverflow job listings,data exploration, cleaning
- Create all the features to the data set:Labeling,Tokenization,Stemming,stopword removal,Hashing TF,IDF
- Naive Bayes model and fit training data
- Transform the model with the testing data
- Evaluation
- Prediction

Naive Bayes Classification Workflow



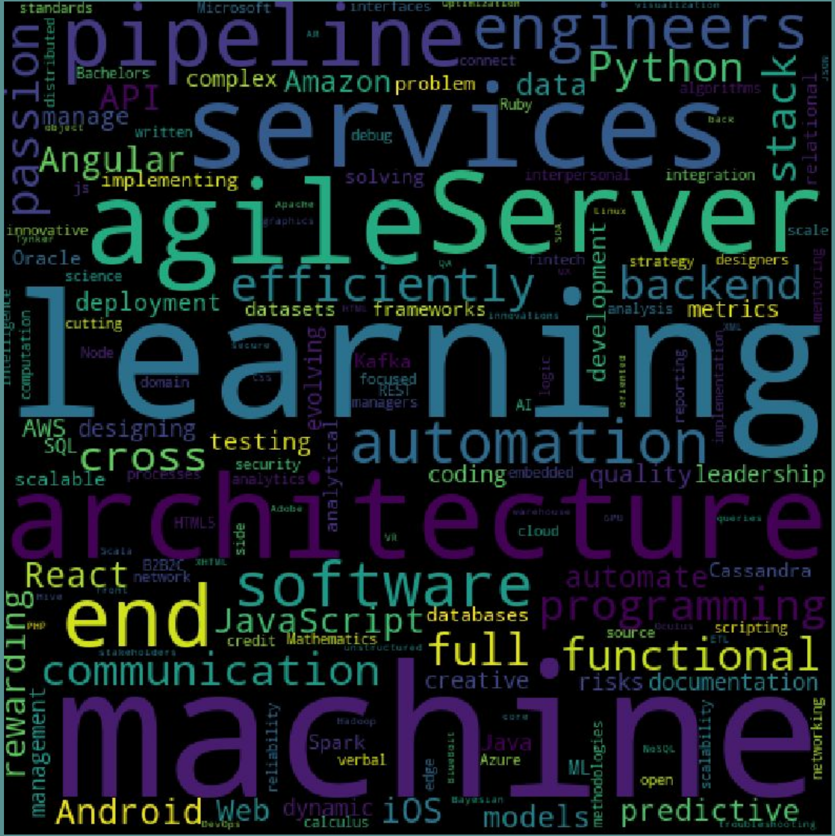
Terms:

*Tokenization:*general process of breaking down a text corpus into individual elements
*Stemming:*process of transforming a word into its root form
*Stop Words:*words that are particularly common in a text but rather un-informative (e.g., words such as so, and, or, the, …”)
*Bag of Words:*a collection of words from a text with word count and disregarding the order in which they appear
*TF(Term Frequency):*number of times a word appears in a document divided by total number of words in the document
*IDF(inverse document frequency):*measures how important a term is.
IDF(t) = log_e(Total number of documents / Number of documents with term t in it)
*Data Pipeline:*data processing elements connected in series, where the output of one element is the input of the next one

Word Cloud:Words Before Cleanup

around world
Qualifications strong
working together
connects billions
solution
gender identity
ul li
li ul
Facebook mission
color national
state local
helps bring
expand reach
together empower
li years
years experience
gender expression
religion color
family apps
br Facebook
different kind
ways share
reasonable accommodation
give people
world gives
br strong
team
together
closer
people
around
bring world
help people
people closer
national origin
qualified applicants
federal state
world closer
based upon
li
work
sex including
including pregnancy
platform
orientation gender
protected veteran
community bring
solving problems
strong br
share matter
will
Action employer
local law
br
system
Responsibilities
new product
race religion
people power
strong ul
billions people
engineer
veteran status
Affirmative Action
getting started
Opportunity
Equal
apps services
gives ways
power build
bring people
Responsibilities strong
Computer Science
sexual orientation
Experience
build community
builders heart
product
ul Facebook
small business
employer discriminate
medical condition
together family
individual disability

Word Cloud: Machine Learning Focus



Process for App

Design --->

Used NLP techniques to design application using Python to match the keywords from user input resume to match with the keywords in the corpus of job postings. The application was based on the ranking results of TF-IDF.

Flask App --->

Built Flask App to serve our application and created a user interface with HTML to allow the user to enter their resume as text. The application is then able to run the resume through the NLP model to create the match.

Hosting

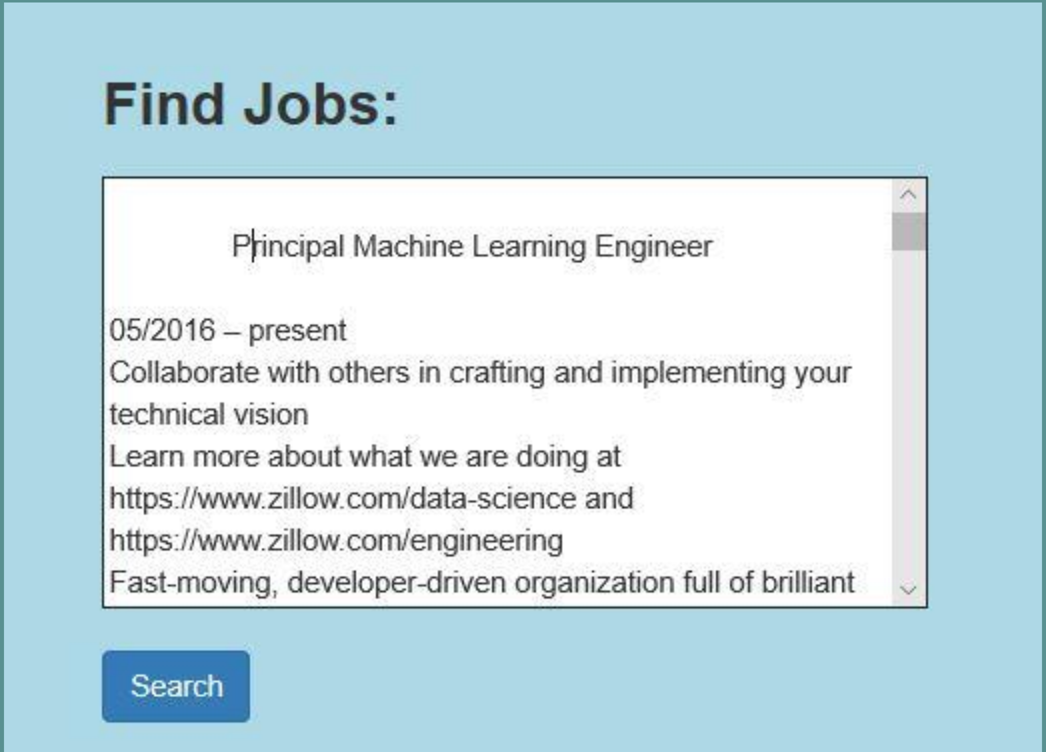
Used AWS to host the server and datasets in S3 buckets. We are able to launch the application in the cloud through AWS.

Programs/Libraries/Languages/Services Used

- Python
- Flask App
- HTML
- Pandas
- NLTK
- Gensim

- AWS
- Tableau
- Scikit-Learn
- Matplotlib
- WordCloud
- Numpy
- Math

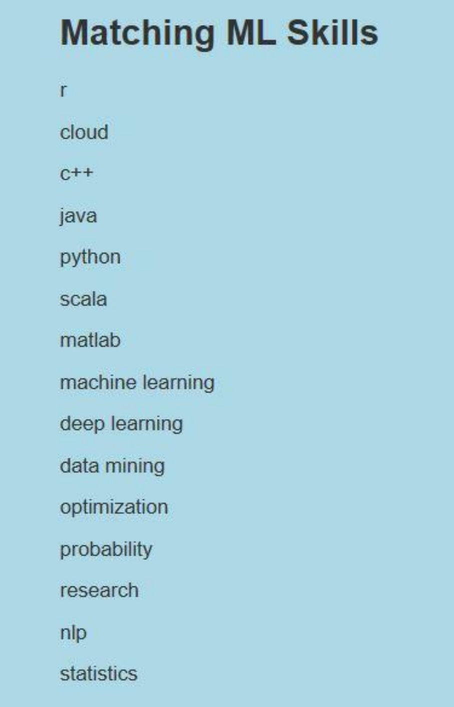
Demo of Resume and Job Posting Application



Find Jobs:

Principal Machine Learning Engineer

05/2016 – present
Collaborate with others in crafting and implementing your technical vision
Learn more about what we are doing at https://www.zillow.com/data-science and https://www.zillow.com/engineering
Fast-moving, developer-driven organization full of brilliant



Matching ML Skills

r

cloud

c++

java

python

scala

matlab

machine learning

deep learning

data mining

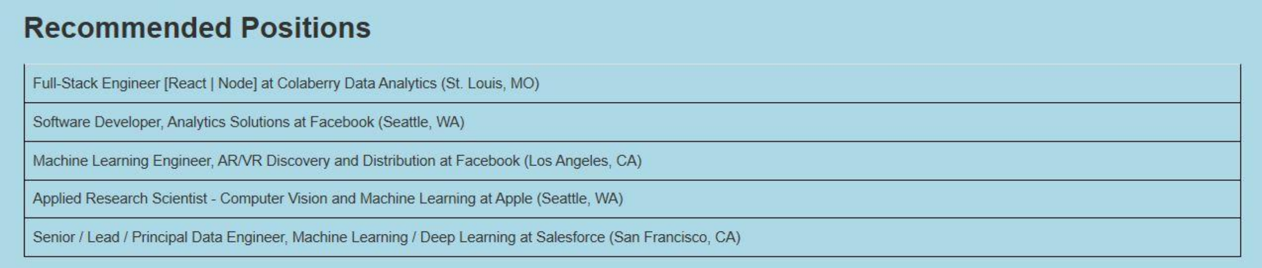
optimization

probability

research

nlp

statistics



Recommended Positions

Full-Stack Engineer [React | Node] at Colaberry Data Analytics (St. Louis, MO)

Software Developer, Analytics Solutions at Facebook (Seattle, WA)

Machine Learning Engineer, AR/VR Discovery and Distribution at Facebook (Los Angeles, CA)

Applied Research Scientist - Computer Vision and Machine Learning at Apple (Seattle, WA)

Senior / Lead / Principal Data Engineer, Machine Learning / Deep Learning at Salesforce (San Francisco, CA)

Github

Please visit our Github repository to see code and other resources:

https://github.com/vgangaprasad/ML_Skills_Match

Questions?