Resume Matching and Job Matching

Zeya Ahmad

Goal

Resume matching is pretty important in job searching. Our goal is to design an engine that could scan your resume and match the information in it to the jobs in the job boards, and recommend the most suitable job link for you to apply based on the similarity of the resume and job description.

This project consists of 3 parts:

Three sub-components of the Project:

1. Parsing Resume-

Libraries used -PyPDF2,pdfplumber

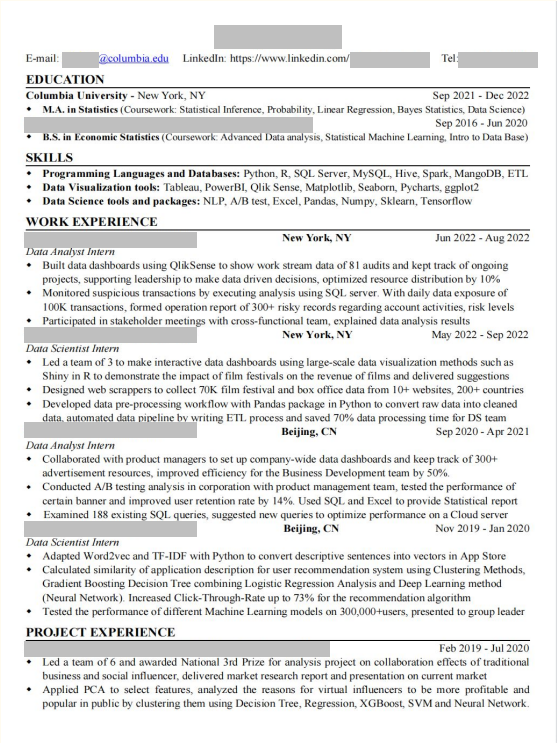
2. Web Scraping LINKEDIN Job Search Page

Libraries used- BeautifulSoup,requests

3.Comparing Resume Text Data to Job Description

Libraries used-CountVectorizer,cosine_similarity

1.Resume Information Extracting



E-mail: @columbia.edu LinkedIn: https://www.linkedin.com/ Tel:

EDUCATION

Columbia University - New York, NY — Sep 2021 - Dec 2022
- **M.A. in Statistics** (Coursework: Statistical Inference, Probability, Linear Regression, Bayes Statistics, Data Science)

Sep 2016 - Jun 2020
- **B.S. in Economic Statistics** (Coursework: Advanced Data analysis, Statistical Machine Learning, Intro to Data Base)

SKILLS
- **Programming Languages and Databases:** Python, R, SQL Server, MySQL, Hive, Spark, MangoDB, ETL
- **Data Visualization tools:** Tableau, PowerBI, Qlik Sense, Matplotlib, Seaborn, Pycharts, ggplot2
- **Data Science tools and packages:** NLP, A/B test, Excel, Pandas, Numpy, Sklearn, Tensorflow

WORK EXPERIENCE

New York, NY — Jun 2022 - Aug 2022

Data Analyst Intern
- Built data dashboards using QlikSense to show work stream data of 81 audits and kept track of ongoing projects, supporting leadership to make data driven decisions, optimized resource distribution by 10%
- Monitored suspicious transactions by executing analysis using SQL server. With daily data exposure of 100K transactions, formed operation report of 300+ risky records regarding account activities, risk levels
- Participated in stakeholder meetings with cross-functional team, explained data analysis results

New York, NY — May 2022 - Sep 2022

Data Scientist Intern
- Led a team of 3 to make interactive data dashboards using large-scale data visualization methods such as Shiny in R to demonstrate the impact of film festivals on the revenue of films and delivered suggestions
- Designed web scrappers to collect 70K film festival and box office data from 10+ websites, 200+ countries
- Developed data pre-processing workflow with Pandas package in Python to convert raw data into cleaned data, automated data pipeline by writing ETL process and saved 70% data processing time for DS team

Beijing, CN — Sep 2020 - Apr 2021

Data Analyst Intern
- Collaborated with product managers to set up company-wide data dashboards and keep track of 300+ advertisement resources, improved efficiency for the Business Development team by 50%.
- Conducted A/B testing analysis in corporation with product management team, tested the performance of certain banner and improved user retention rate by 14%. Used SQL and Excel to provide Statistical report
- Examined 188 existing SQL queries, suggested new queries to optimize performance on a Cloud server

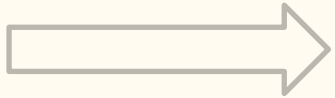
Beijing, CN — Nov 2019 - Jan 2020

Data Scientist Intern
- Adapted Word2vec and TF-IDF with Python to convert descriptive sentences into vectors in App Store
- Calculated similarity of application description for user recommendation system using Clustering Methods, Gradient Boosting Decision Tree combining Logistic Regression Analysis and Deep Learning method (Neural Network). Increased Click-Through-Rate up to 73% for the recommendation algorithm
- Tested the performance of different Machine Learning models on 300,000+users, presented to group leader

PROJECT EXPERIENCE

Feb 2019 - Jul 2020
- Led a team of 6 and awarded National 3rd Prize for analysis project on collaboration effects of traditional business and social influencer, delivered market research report and presentation on current market
- Applied PCA to select features, analyzed the reasons for virtual influencers to be more profitable and popular in public by clustering them using Decision Tree, Regression, XGBoost, SVM and Neural Network.

PDF Resume



Text

se)SKILLS\uf077 Programming Languages andDatabases: Python, R, SQL Server, MySQL, Hive, Spark, MangoDB, ETL\uf077 Data Visualization tools:Tableau, PowerBI, QlikSense, Matplotlib, Seaborn, Pycharts, ggplot2\uf077 Data Sciencetools andpackages: NLP, A/B test, Excel, Pandas, Numpy, Sklearn, TensorflowWORK EXPERIENCEBankofChina, NewYork Branch NewYork, NY Jun 2022-Aug 2022DataAnalyst Intern\uf077 Built datadashboards using QlikSensetoshowwork stream data of81audits and kept track of ongoingprojects, supporting leadership to make datadriven decisions, optimized resource distribution by10%\uf077 Monitored suspicious transactionsbyexecuting analysis using SQL server. Withdaily data exposureof100Ktransactions, formed operation report of300+ risky records regarding account activities, risk levels\uf077 Participated instakeholder meetings with cross-functional team, explained data analysis results72Dragons FilmProductions,LLC NewYork, NY May2022- Sep 2022Data Scientist Intern\uf077 Led ateam of3to make interactive datadashboards usinglarge-scale datavisualization methods such asShiny inR to demonstratetheimpact offilm festivals ontherevenue offilmsand delivered suggestions\uf077 Designed web scrappers to collect 70K filmfestival and boxoffice data from 10+websites,200+countries\uf077 Developed data pre-processing workflowwith Pandas package in Python to convert rawdata intocleaneddata,automated data pipelinebywriting ETL process and saved 70%data processing timefor DS teamSo-Young International Inc. Beijing, CN Sep 2020-Apr 2021DataAnalyst Intern\uf077 Collaborated with product managers to set upcompany-wide data dashboardsand keep track of300+advertisement resources, improved efficiency for theBusiness Development team by50%.\uf077 Conducted A/B testinganalysis incorporation with product management team, tested theperformance ofcertain banner and improved userretention rate by14%.Used SQL and Excel toprovideStatistical report\uf077 Examined 188existing SQL queries, suggested new queries to optimizeperformance onaCloud serverLenovoGroupLimited Beijing, CN Nov 2019 -Jan 2020Data Scientist Intern\uf077 Adapted Word2vecand TF-IDF with Python to convert des

2. Web Scraping LINKEDIN Job Search Page

Libraries used- BeautifulSoup,requests

Steps for Web Scraping

1. Extract the HTML content using the requests library.
2. Analyze the HTML structure and identify the tags which have our content.
3. Extract the tags using BeautifulSoup and put the data in a Python list.

Requests (HTTP for Humans) Library for Web Scraping – It is used for making various types of HTTP requests like GET, POST, etc. It is the most basic yet the most essential of all libraries.

BeautifulSoup-extracts content from an HTML page. After extraction, we'll convert it to a Python list or dictionary using Beautiful Soup.

Posted Job Information

Enter your desired position: Data Scientist

Enter your desired location:

Enter keywords to filter

	Job_title	Company	Job_posted_date	Link
0	Data Scientist Full Time	Bardess Group Ltd	2022-12-06	https://www.linkedin.com/jobs/view/data-scient...
1	Data Scientist Full Time	Bardess Group Ltd	2022-12-06	https://www.linkedin.com/jobs/view/data-scient...
2	Machine Learning Engineer	CareerWellness	2022-12-06	https://www.linkedin.com/jobs/view/machine-lea...
3	Data Scientist Full Time	Bardess Group Ltd	2022-12-06	https://www.linkedin.com/jobs/view/data-scient...
4	Data Scientist	Jobot	2022-12-06	https://www.linkedin.com/jobs/view/data-scient...
...	...	...	...	...
120	Data Scientist	FinTech LLC	2022-09-28	https://www.linkedin.com/jobs/view/data-scient...
121	Machine Learning Engineer (NLP)	BLACKBIRD.AI	2022-09-11	https://www.linkedin.com/jobs/view/machine-lea...
122	Machine Learning Engineer (NLP)	BLACKBIRD.AI	2022-09-11	https://www.linkedin.com/jobs/view/machine-lea...
123	Machine Learning Engineer (NLP)	BLACKBIRD.AI	2022-09-11	https://www.linkedin.com/jobs/view/machine-lea...
124	Machine Learning Engineer (NLP)	BLACKBIRD.AI	2022-09-11	https://www.linkedin.com/jobs/view/machine-lea...

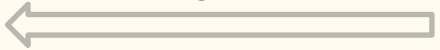
Satisfying jobs

Information Matching

Resulting Data Frame is based on the filter- DESIRED JOB TITLE, LOCATION,NUMBER OF PAGES SCRAPED in descending of Job posted date .

	Job_title	Company	Job_posted_date	Link	Matching_percentage
89	Data Scientist	Atlantic Partners Corporation	2022-10-23	https://www.linkedin.com/jobs/view/data-scient...	55.478849
24	Data Scientist I	The Walt Disney Company	2022-12-01	https://www.linkedin.com/jobs/view/data-scient...	55.237787
28	Data Scientist I	The Walt Disney Company	2022-12-01	https://www.linkedin.com/jobs/view/data-scient...	55.237787
52	Data Scientist (REMOTE)	Foot Locker	2022-11-20	https://www.linkedin.com/jobs/view/data-scient...	54.462898
81	Data Scientist	Afficiency	2022-11-03	https://www.linkedin.com/jobs/view/data-scient...	54.452135
...	...	...	...	...	...
67	Junior Machine Learning	Diverse Lynx	2022-11-16	https://www.linkedin.com/jobs/view/junior-mach...	27.789568
62	Machine Learning Engineer	Diverse Lynx	2022-11-16	https://www.linkedin.com/jobs/view/machine-lea...	26.710361
64	Junior Machine Learning Developer	Diverse Lynx	2022-11-16	https://www.linkedin.com/jobs/view/junior-mach...	26.108326
61	Junior Machine Learning	Diverse Lynx	2022-11-16	https://www.linkedin.com/jobs/view/junior-mach...	24.939839
68	Machine Learning Engineer	ACHIEVA Group Limited	2022-11-15	https://www.linkedin.com/jobs/view/machine-lea...	17.197384

Recommended jobs based on information matching



3.Comparing Resume Text Data to Job Description

1.Extract Job Description from 2nd page after the search i.e the landing page for each Job post
2.Create a list containing the two documents i.e,Resume and the Job Description and then run **CountVectorizer()** on them

CountVectorizer-

CountVectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample. This can be visualized as follows –

	hello	is	james	my	name	notebook	python	this
0	1	1	1	1	1	0	0	0
1	0	1	0	1	0	1	1	1

3.**Cosine Similarity**-Cosine similarity is a metric used to measure how similar the documents are irrespective of their size

Incomplete/Failed Implementations

- After filtering our results based on DESIRED POSITION,DESIRED LOCATION and NUMBER OF PAGES to scrape ,we also wanted to filter results based on SKILLS .
- We tried to extract skills from the Job Description and Resume by using NER(Named Entity Recognition) functionality of NLTK.
- Used the database of skills from Skills API and Lightcast API to extract only skills from the Resume .
- Unsuccessful in extracting skills from the Job Descriptions.

- Future Plans

1. Match jobs from other websites, including Indeed, Glassdoor, etc.

2. Match more specifically. Skill, experience level match, etc.

Reference:

Matching CV/resume To Job Description using Python :
https://www.kaggle.com/code/nezarabdilahprakasa/matching-cv-to-job-description-using-python/notebook
https://randerson112358.medium.com/resume-scanner-2c30f5baf92c

Automating job search in Indeed with Python :
https://www.chrislovejoy.me/job-scraper
https://www.youtube.com/watch?v=070e7nMYt6c
https://github.com/MrFuguDataScience/Webscraping/blob/master/IndeedJune2022_href_jobDescription.ipynb
https://www.youtube.com/watch?v=-SjrfjKJqqI

LinkedIn: https://medium.com/@kurumert/web-scraping-linkedin-job-page-with-selenium-python-e0b6183a5954

Thank you!