JOB RECOMMENDATION SYSTEM

Objective: To understand the trends in data science jobs and build a job recommendation system using text analysis.

Introduction:

We have a dataset of data science job listings in Australia from 2019- 2021 that contains information such as job titles, Classification, and sector, location, required skills (programming language), company names, work type (Full/Part/Intern), etc. We initially perform data visualization to understand the dataset and the trends it reflects. Then we build a system that gets input from users, such as – desired job, its field, location, work type, and skills, and from this info, give a list of jobs that are suitable for the user. We use text analysis for this task.

Dataset: The dataset used in the project is from "Deep Exploration of Data Science Jobs". The dataset contains 52 columns and thus is not shown here.

Data Preprocessing:

Data cleaning – To clean the dataframe of unnecessary columns that do not affect the model and/or have a significant null content , we drop those- company Id, companyRating, area, suburb, salary string, companyProfileUrl, desktopAdTemplate, companyName, first_seen, last_seen, recruiter, isRightToWorkRequired, nation, jobId, advertiserId, mobileAdTemplate, listingDate, expiryDate, seekJobListingUrl. The remaining data is free of null values and thus is ready to be used.

Dataset modifications - The programming languages required for a particular job are in binary encodings, and we want them in text format (as we perform text analysis later). Thus we make a column containing strings of required languages and drop the columns with binary values.

Feature selection – For performing the text analysis, we choose a set of features that affect the job recommendations, which are – jobTitle, job classification, state, city, work type, and language.

User Input:

This step involves getting input from the user on his job search criteria. The input is in string format and contains criteria that the user desires in his job hunt, such as - desired job, its field, location, work type, and programming skills. Example input – 'full time Sydney data scientist information technology python R'.

Text Analysis:

To perform analysis on the dataset, we first combine all the selected features into a single column.

Text cleaning- Then we perform a text cleaning to remove 'stopwords,' punctuations, and symbols and tokenize the string of data that we name as 'description' as it efficiently describes the job. The 'nltk' and 'string' packages enable us to do this.

Text Analytics-

1. Now, we construct TF-IDF matrices for the description data we have. The TF-IDF matrix is a numerical representation of text data that reflects the importance of each word in each document. It is a way to represent text data in a vector space model, where each document is a vector and each word is a dimension. This is done using 'tfidfvectorizer' from sklearn.

2. Then, we find cosine similarities of obtained matrices. Cosine similarity is a metric used to measure the similarity between two vectors of an inner product space. In the context of text analysis, it is often used to determine the similarity between two documents based on their word frequency vectors. This is done using 'cosine_similarity'.

3. Finally, we obtain the similarity scores of input data with the entire dataset and sort them in order of decreasing similarity values.

We also perform text cleaning on the input data to ensure it does not affect the analysis.

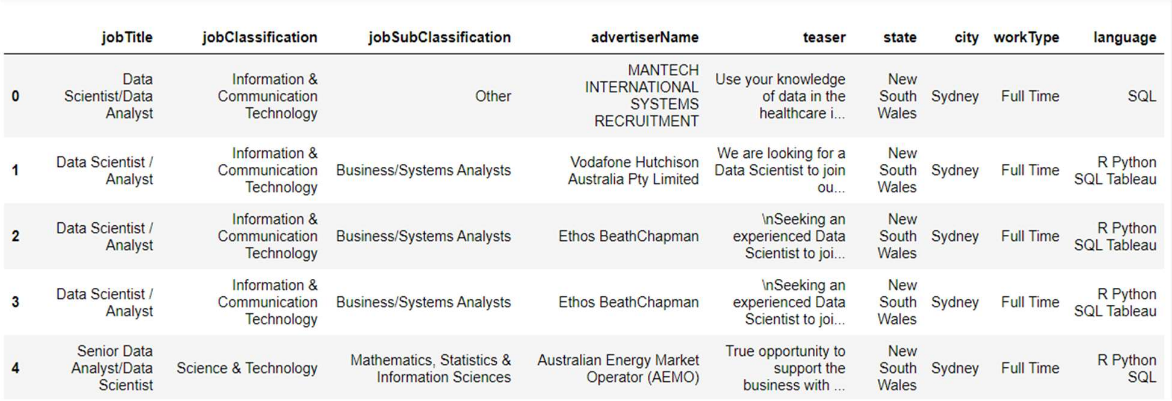
Output:

After sorting similarity scores, we take the first 50 similarity score indexes. Then the job postings corresponding to those indexes are given as output of jobs that the user might find relevant to his search. The jobs dataset contains 50 relevant jobs.

Below is an example and output dataset of 5 jobs.

User search – "full-time data analyst information technology Sydney Python R "

Jobs recommended (first 5) -



	jobTitle	jobClassification	jobSubClassification	advertiserName	teaser	state	city	workType	language
0	Data Scientist/Data Analyst	Information & Communication Technology	Other	MANTECH INTERNATIONAL SYSTEMS RECRUITMENT	Use your knowledge of data in the healthcare i...	New South Wales	Sydney	Full Time	SQL
1	Data Scientist / Analyst	Information & Communication Technology	Business/Systems Analysts	Vodafone Hutchison Australia Pty Limited	We are looking for a Data Scientist to join ou...	New South Wales	Sydney	Full Time	R Python SQL Tableau
2	Data Scientist / Analyst	Information & Communication Technology	Business/Systems Analysts	Ethos BeathChapman	\nSeeking an experienced Data Scientist to joi...	New South Wales	Sydney	Full Time	R Python SQL Tableau
3	Data Scientist / Analyst	Information & Communication Technology	Business/Systems Analysts	Ethos BeathChapman	\nSeeking an experienced Data Scientist to joi...	New South Wales	Sydney	Full Time	R Python SQL Tableau
4	Senior Data Analyst/Data Scientist	Science & Technology	Mathematics, Statistics & Information Sciences	Australian Energy Market Operator (AEMO)	True opportunity to support the business with ...	New South Wales	Sydney	Full Time	R Python SQL