

# **JOB BOARD SCRAPER**

**Version 1.0**

**10/05/2020**

**Amal Das**

## Table of Contents

<b>Assignment Scope and Requirements</b>	<b>3</b>
<b>INTRODUCTION</b>	<b>3</b>
<b>Problem/Challenge 1 :</b>	<b>4</b>
<b>Problem/Challenge 2 :</b>	<b>4</b>
<b>CODE / WEB SCRAPER</b>	<b>5</b>
<b>CODE / ANALYSIS</b>	<b>6</b>
<b>ALGORITHM</b>	<b>7</b>
<b>CHALLENGES</b>	<b>7</b>
<b>RESULT</b>	<b>8</b>
<b>Skillsets for an Analyst job within the AR/VR industry.</b>	<b>10</b>

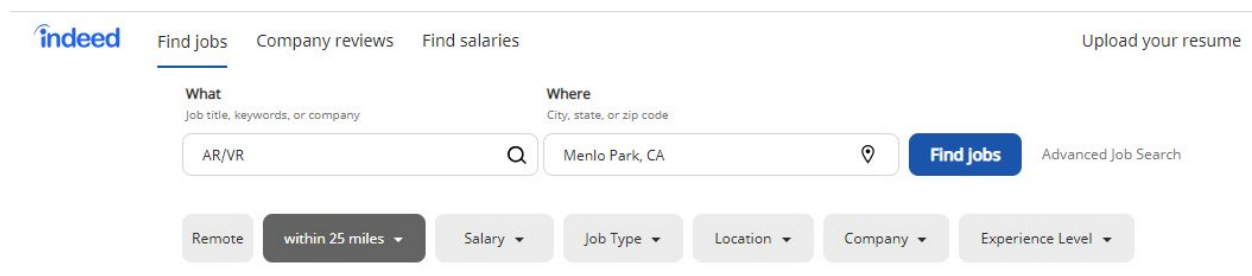
## Assignment Scope and Requirements

Using whichever software that you feel comfortable with and have access to, please complete the following assignment:

- The assignment is to determine the skills required for AR/VR jobs in Menlo Park, California.
- Please complete a web scraping of at least 50 “AR/VR” job posts. Postings may come from LinkedIn, Indeed, and/or Monster. Please save your results and store them in a .csv file.
- Please clean and process the job posts and use text mining to identify what you believe to be “key” or “critical” skills for this occupation. You can use text, tables, and/or visualizations to communicate your findings.
- Please briefly explain which classification or clustering algorithms you feel are best used in order to extract “key” or “critical” skills from those job posts.
- Please write a paragraph on the specific challenges we all face when scraping job boards for employment data.

## INTRODUCTION

In this assignment, I will be using Job postings from Indeed.com to find the skillsets for the listed jobs in the area of AR / VR within the city Menlo Park.

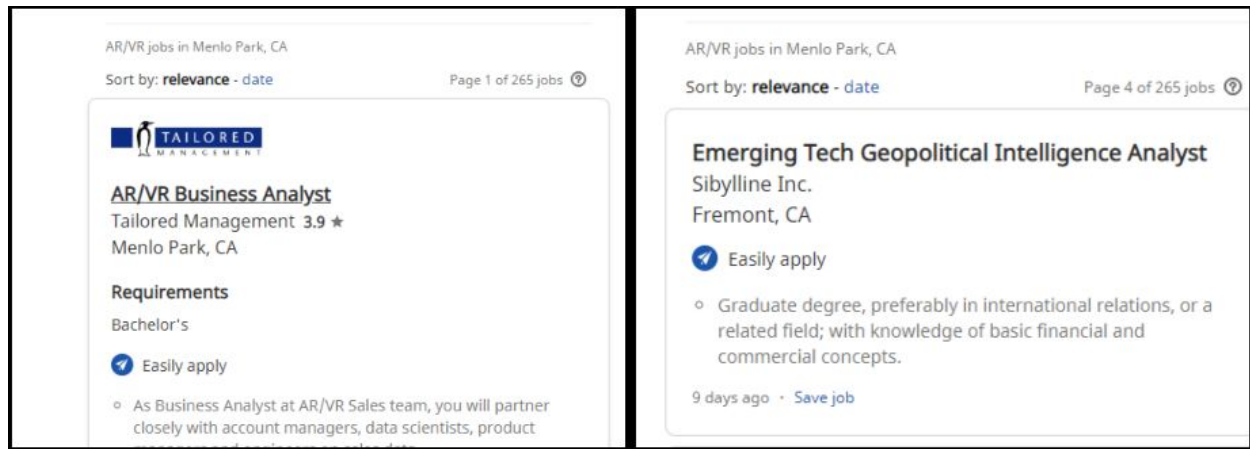


The screenshot shows the Indeed job search interface. At the top, there are links for 'Find jobs', 'Company reviews', and 'Find salaries', along with a link to 'Upload your resume'. The main search area has two input fields: 'What' (Job title, keywords, or company) and 'Where' (City, state, or zip code). The 'What' field contains 'AR/VR' and the 'Where' field contains 'Menlo Park, CA'. A blue 'Find Jobs' button is to the right of the 'Where' field. Below the search bar are several filter buttons: 'Remote', 'within 25 miles', 'Salary', 'Job Type', 'Location', 'Company', and 'Experience Level'.

In order to collect the data / create a dataset for the purpose of finding the skillsets for this particular job title, we will have to write a web scraping script to generate the data we want. Before we get started on writing the scripts, let's explore the website and see some results.

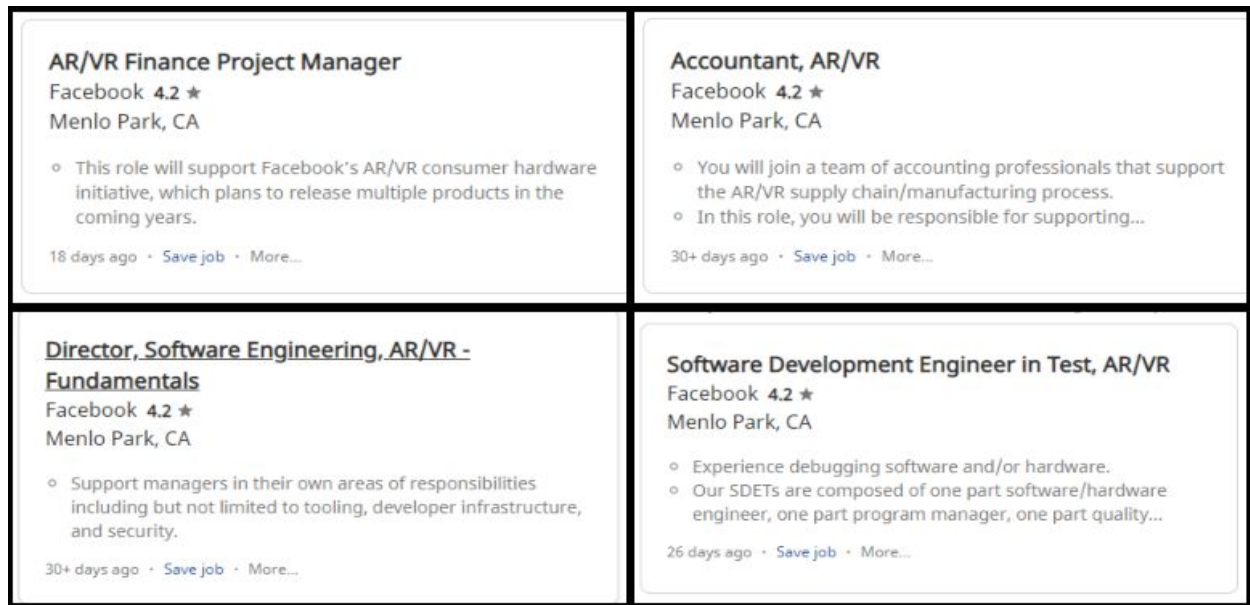
### Problem/Challenge 1 :

Sometimes Indeed itself gives irrelevant Job listings on our search query. The below image is an example of the search result from indeed for AR/VR jobs in Menlo Park, CA. The left side image shows one good result and the right side image is a bad result. So this issue should be considered while creating our web scraper.



### Problem/Challenge 2 :

Another problem that I encountered while browsing the search results was, with the keyword AR/VR there are many kinds of job listings, so in order to identify the skillsets for this job role, we will have to identify the title so that our algorithm can perform better.



Each job role has its own key skills, so we will have to identify that while doing our analysis.

## CODE / WEB SCRAPER

Libraries used :

- BeautifulSoup

We will be using the library BeautifulSoup for web scraping and parsing the HTML file.

- Pandas

Pandas library will be used for data manipulation and analysis.

Elements to find from indeed website HTML code by using the inspect element feature of chrome/firefox.

- Links to the job posts: a href tags and its classes
- How indeed is handling page numbers: it's a multiple of 10 for each page, like 10 for page number 2, 20 for page number 3, etc.
- The job description of individual job posts: which is usually a LI tag within the UL. we will be taking this part along with company name and location from inner pages



```
aboveFullJobDescription"></div>
<script type="text/javascript">...</script>
<div id="jobDescriptionText" class="jobsearch-
jobDescriptionText">
  <div>
    <div>...</div>
    <p></p>
    <div>
      <div>
        AR SILICON ARCHITECT RESPONSIBILITIES</div>
      <div></div>
    <div>
      <div>
        <ul>
          <li>
            <div>
              <div>
                Architecture definition and specification
                of a chip subsystem or entire SoC.</div>
              </div>
            </li>
          </ul>
        </div>
      </div>
    </div>
  </div>
</div>
```

Link to the script:

<https://github.com/amald94/indeed-scraper/blob/master/IndeedScraper.ipynb>

## CODE / ANALYSIS

Once we generated our dataset and considering the problems listed above, let's do some text mining and NLP on the raw text data and get the keywords out of it.

Libraries Used :

- Pandas: for data manipulation and analysis
- WordCloud: for visualizing the text data
- Matplotlib: for visualization
- Rake: it's an NLP library for processing raw unstructured data to find the keywords
- Stop-words: it's a library that has common stop words in many languages, that won't add any special meaning to a text or don't represent a keyword.

RAKE short for Rapid Automatic Keyword Extraction algorithm is a domain-independent keyword extraction algorithm that tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text.

Steps Performed on raw unstructured data:

1. Clean the JD
2. Remove stop words from the JD
3. Create a new processed JD feature
4. Pass the processed data as an input to the NLP library
5. Identify the keywords

Code / explanation Link :

<https://github.com/amald94/indeed-scraper/blob/master/Analysis.ipynb>

## ALGORITHM

When working with raw unstructured data to find the insights from it, we can use either classification algorithms or clustering algorithms but it depends on the task and the problem we are trying to solve.

Clustering algorithms are useful when we wanted to group some data points based on some features, hence these algorithms are less preferred for text extraction.

In this specific problem of finding keywords/skills from a JD, classification algorithms are the best to use. RAKE is a widely used approach in text mining to extract keywords from text data, and that is built on top of the natural language processing library NLTK.

How RAKE algorithm works behind the scene:

- Candidates ( words and phrases ) are extracted from a given string of texts by eliminating stopwords.
- A co-occurrence graph is built to identify the frequency of the candidates.
- A score is calculated for each candidate that is the sum of the individual candidate's scores from the co-occurrence graph.
- Adjoining keywords are also included if they occur more than twice in the given text.
- Then top keywords are extracted and will give back in sorted order of their score.

Combining both clustering and classification would significantly increase the performance of our model. For instance, there can be multiple job titles for some job roles, so identifying those posts and grouping them together will help to identify and extract keywords from the JDs.

## CHALLENGES

Some challenges I came across when scraping the Indeed website are as follows

- Sometimes the search results its self is giving non-relevant results. So that has to be identified when you are specifically looking for skillsets of a job role. For instance, the AR/VR analyst has a different set of required skills and AR/VR mobile app developer has another set of skills. So when identifying key skills for AR/VR jobs we have to consider the different types of jobs within the AR/VR industry.
- Sometimes there will be duplicate job postings for the same role. So it has to considered when building the script/model.

- For a same job role, there can be multiple job titles with almost similar JD and requirements. This should also be considered. For instance AR/VR developer, AR/VR software engineer, AR/VR software developer, etc.
- Other challenges include changes in HTML tags and classes.

## RESULT

This is how our Dataset looks like after scraping the website.

read the scraped data using pandas

```
In [2]: raw_data = pd.read_csv("raw_data.csv")
```

```
In [3]: raw_data.head(2)
```

```
Out[3]:
```

	Title	Company	Location	JD	URL
0	AR/VR Business Analyst	Tailored Management	Menlo Park, CA	skills: Consolidate third-party retailer sale...	https://www.indeed.com/pagead/clk?mo=r&ad=-6N...
1	Producer, Augmented Reality (Menlo Park)	The Mom Project	Menlo Park, CA	skills: Work closely with creative leads to s...	https://www.indeed.com/pagead/clk?mo=r&ad=-6N...

Features we have in our dataset

- Job Title
- Company Name
- Location
- Job Description
- URL of the Job

Since I scraped only around 50 job listings it's easy to assign job roles for each job postings based on the Job Title.

A function to assign Job roles based on the job title

```
In [7]: def assign_roles(row):
row = row.lower()
if 'accountant' in row:
    return 'accountant'
elif 'marketing' in row:
    return 'marketing'
elif 'security' in row and 'manager' in row:
    return 'security'
elif 'manager' in row and 'engineer' in row:
    return 'Engineering manager'
elif 'manager' in row:
    return 'manager'
elif 'director' in row:
    return 'director'
elif 'software' in row or 'engineer' in row:
    return 'engineer'
elif 'finance' in row:
    return 'financial analyst'
elif 'analyst' in row:
    return 'analyst'
else:
    return 'other'
```



So our processed dataset will look like as below,

```
In [8]: # add job role feature to the dataset
# categorize job listing
raw_data['Role'] = raw_data['Title'].apply(lambda row: assign_roles(row))
raw_data.head()
```

```
Out[8]:
```

	Title	Company	Location	JD	URL	JD_processed	Role
0	AR/VR Business Analyst	Tailored Management	Menlo Park, CA	skills: Consolidate third-party retailer sale...	https://www.indeed.com/pagead/clk?mo=r&ad=-6N...	skills: consolidate third-party retailer sale...	analyst
1	Producer, Augmented Reality (Menlo Park)	The Mom Project	Menlo Park, CA	skills: Work closely with creative leads to s...	https://www.indeed.com/pagead/clk?mo=r&ad=-6N...	skills: work closely with creative leads to s...	other
2	Content Manager, AR/VR (Japanese Market)	An Innovative VR Company	Menlo Park, CA	skills: Curate and manage the App Store and a...	https://www.indeed.com/pagead/clk?mo=r&ad=-6N...	skills: curate and manage the app store and a...	manager

So by grouping based on the Job roles, we assigned we can get the result as follows,

```
In [9]: # check how many job rols in each category we have
raw_data.Role.value_counts()
```

```
Out[9]: manager      17
other      17
engineer    15
marketing   11
analyst      9
director     7
Engineering manager  2
accountant   1
financial analyst  1
Name: Role, dtype: int64
```

Now we have more than 1 job listings for some category which definitely will help to reduce noises and generating non-relevant skill sets for a particular job role. Since manager, developer, an accountant has their own skill sets and which are totally independent.

Let's initialize our RAKE instance with a set of stopwords and other parameters like the max length of a candidate ( that is once we split the words from a given string, each will be considered as a candidate ), minimum length, etc.

```
In [13]: # import rake library
from rake_nltk import Rake
# Uses stopwords for english from NLTK, and all punctuation characters.
r = Rake(
    stopwords=en_stop,
    max_length=3,
    min_length=1,)
```

If we have a lot of Job listing in each category we can further include parameters like frequency of a word. Which will help to assign more scores on those words based on the no.of times it appeared.

### Skillsets for an Analyst job within the AR/VR industry.

The top-scored words extracted from the JD for an analyst job.

```
In [17]: df_analyst.head()
```

Out[17]:

	score	keywords
0	9.0	similar scripting language
1	9.0	retail industry preferred
2	9.0	presto spark hive
3	9.0	nps csat etc
4	9.0	manipulating raw datasets

The least scored words for the same job role

```
In [18]: df_analyst.tail()
```

Out[18]:

	score	keywords
98	1.0	internal
99	1.0	customers
100	1.0	collaborate
101	1.0	area
102	1.0	8

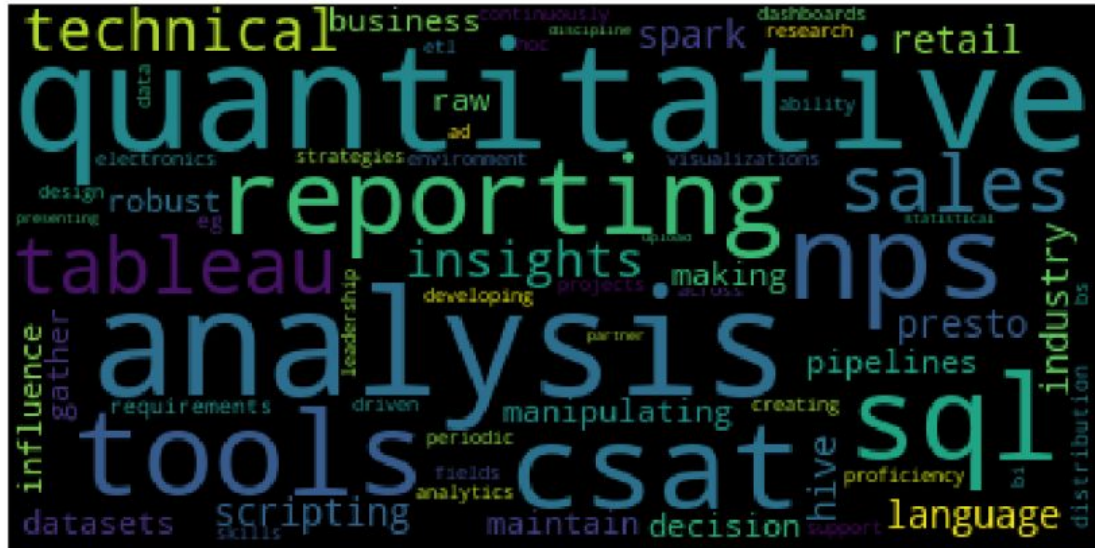
As we can see here, when the scores are low it doesn't show its a relevant word and hence we can ignore that.

So I decided to draw a word cloud from the result and as we can see from the below image and from the table above most required skillsets for an analyst role are as follows,

- SQL
- SPARK
- HIVE
- PRESTO
- Manipulating raw datasets
- Scripting language experience
- Tableau
- Etc

Word cloud for an Analyst job role within the AR/VR industry.

```
In [22]: draw_word_cloud(text, "analyst", st_wrds)
```



Analysis of other job roles can be found here :

<https://github.com/amald94/indeed-scraper/blob/master/Analysis.ipynb>

## References :

<https://www.amazon.com/Text-Mining-Applications-Michael-Berry/dp/0470749822/>

<https://monkeylearn.com/keyword-extraction/>