STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Analysis of skills affected due to Automation and AI

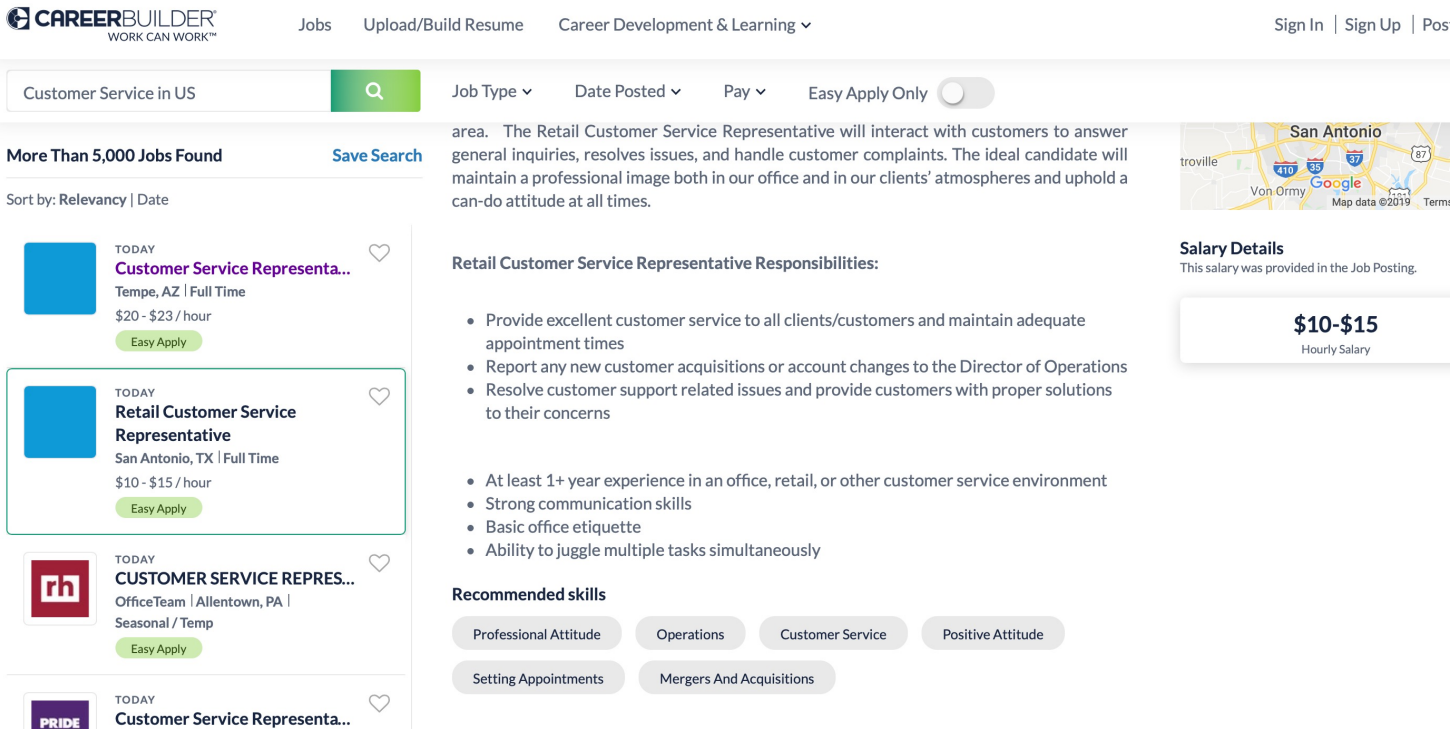
Instructor: Rong Liu
Course: BIA-660-C

Team Members: Group 5
Rohan Shah
Parth Parab
Purva Khopkar
Bhagyashree Shende

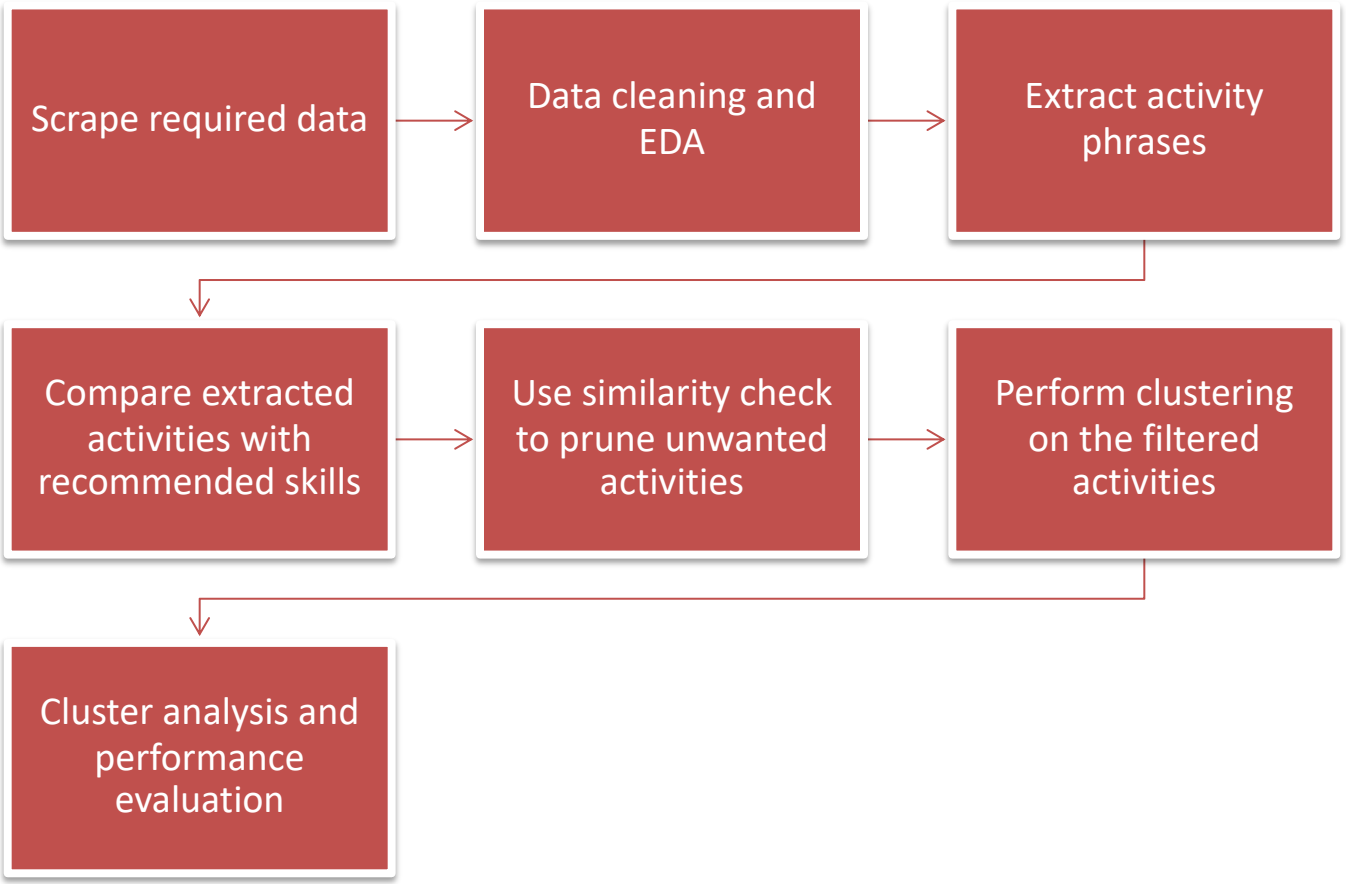
Introduction

Goal of the project: To cluster skills and activities from various job postings scraped from “careerbuilder.com”

- Cluster skills and activities from various job postings
- Analyze the activities in the cluster to predict which clusters may be impacted due to Automation and AI



Project Workflow



STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Web Scraping and Data Preprocessing

Web Scraping

- Scrapped first 100 pages of Careerbuilder having 25 jobs each page to get details 10000 Jobs for 4 types of keywords namely Software Engineering, Manufacturing, Fabrication, Customer Service and store the data into file using BeautifulSoup

Editor - C:\Users\rohan\Downloads\test.py

test.py*

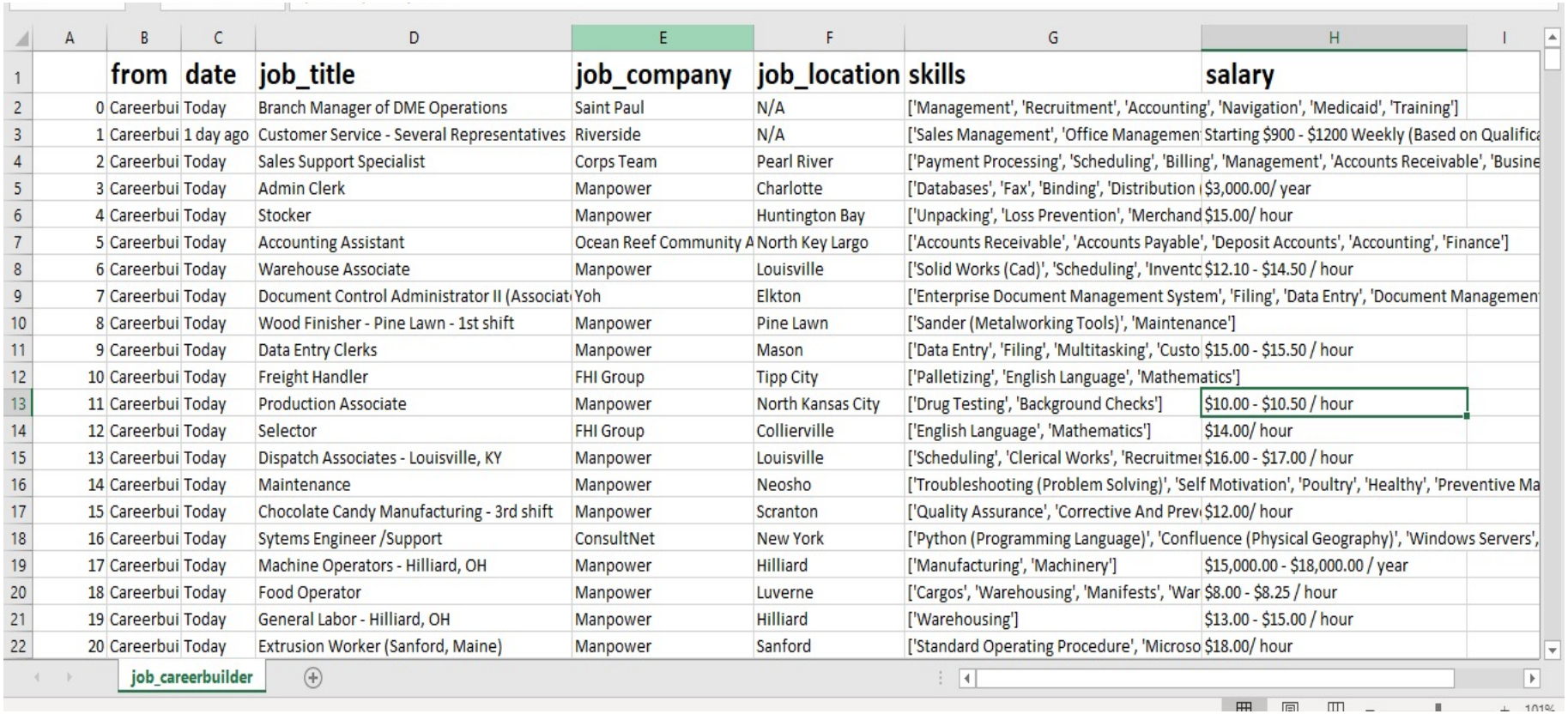
```
import pandas
import requests
import bs4

BASE_URL_careerbuilder = 'https://www.careerbuilder.com/jobs'

rawcode = requests.get(BASE_URL_careerbuilder)
soup = bs4.BeautifulSoup(rawcode.content, "html.parser")

num_pages_careerbuilder=100;
job_df_careerbuilder = pandas.DataFrame()

for i in range(1, num_pages_careerbuilder+1):

    url = ''.join([BASE_URL_careerbuilder,'?page_number=', str(i)])
    rawcode = requests.get(url)
    soup = bs4.BeautifulSoup(rawcode.content, "html.parser")
    job_divs = soup.select("div#jobs_collection div.data-results-content-parent")
    temp_URL_careerbuilder = 'https://www.careerbuilder.com/'
    for each in job_divs:

        job_location_span = each.select("div.data-details span")
        if(len(job_location_span) ==3):
            job_location= job_location_span[1].get_text()

        else:
            job_location= "N/A"

        job_company = job_location_span[0].get_text()

        posted_date = each.select("div.data-results-publish-time")[0].get_text()
```

Editor - C:\Users\rohan\Downloads\test.py

test.py*

```
        job_title = soup.select("div.data-display-header_info-content h1")[0].get

        job_description_span= soup.select("div.col-2 div.col.big.col-mobile-full

        if(job_description_span != None and len(job_description_span) > 0):
            job_description = job_description_span[0].get_text()
            if(job_description==""):
                job_description = "N/A"
        else:
            job_description = "N/A"

        salary_div=soup.select("div.data-snapshot div.block")
        if(salary_div != None and len(salary_div) > 0 ):
            salary = salary_div[0].get_text()
        else:
            salary = "Not Disclosed"
        job_df_careerbuilder = job_df_careerbuilder.append({'job_title': job_titl

                                'job_company': job_company,
                                'date': posted_date,
                                'from':'Careerbuilder',
                                'job_location':job_location,
                                'skills': list_of_skills,
                                'salary' : salary
                                },ignore_index=True)
cols=['from','date','job_title','job_company','job_location','skills','salary']
job_df_careerbuilder = job_df_careerbuilder[cols]

path = 'C:\\Users\\pkhopkar\\Downloads\\BIA660\\BIA660\\' + 'job_careerbuilder'
job_df_careerbuilder.to_csv(path)
```

Scrapped data from Careerbuilder



	from	date	job_title	job_company	job_location	skills	salary
0	Careerbui	Today	Branch Manager of DME Operations	Saint Paul	N/A	['Management', 'Recruitment', 'Accounting',	'Navigation', 'Medicaid', 'Training']
1	Careerbui	1 day ago	Customer Service - Several Representatives	Riverside	N/A	['Sales Management', 'Office Managemen	Starting $900 - $1200 Weekly (Based on Qualifica
2	Careerbui	Today	Sales Support Specialist	Corps Team	Pearl River	['Payment Processing', 'Scheduling', 'Billing', 'Management', 'Accounts Receivable', 'Busine
3	Careerbui	Today	Admin Clerk	Manpower	Charlotte	['Databases', 'Fax', 'Binding', 'Distribution	$3,000.00/ year
4	Careerbui	Today	Stocker	Manpower	Huntington Bay	['Unpacking', 'Loss Prevention', 'Merchand	$15.00/ hour
5	Careerbui	Today	Accounting Assistant	Ocean Reef Community A	North Key Largo	['Accounts Receivable', 'Accounts Payable', 'Deposit Accounts', 'Accounting', 'Finance']
6	Careerbui	Today	Warehouse Associate	Manpower	Louisville	['Solid Works (Cad)', 'Scheduling', 'Inventc	$12.10 - $14.50 / hour
7	Careerbui	Today	Document Control Administrator II (Associat	Yoh	Elkton	['Enterprise Document Management System', 'Filing', 'Data Entry', 'Document Managemen
8	Careerbui	Today	Wood Finisher - Pine Lawn - 1st shift	Manpower	Pine Lawn	['Sander (Metalworking Tools)', 'Maintenance']
9	Careerbui	Today	Data Entry Clerks	Manpower	Mason	['Data Entry', 'Filing', 'Multitasking', 'Custo	$15.00 - $15.50 / hour
10	Careerbui	Today	Freight Handler	FHI Group	Tipp City	['Palletizing', 'English Language', 'Mathematics']
11	Careerbui	Today	Production Associate	Manpower	North Kansas City	['Drug Testing', 'Background Checks']	$10.00 - $10.50 / hour
12	Careerbui	Today	Selector	FHI Group	Collierville	['English Language', 'Mathematics']	$14.00/ hour
13	Careerbui	Today	Dispatch Associates - Louisville, KY	Manpower	Louisville	['Scheduling', 'Clerical Works', 'Recruitmer	$16.00 - $17.00 / hour
14	Careerbui	Today	Maintenance	Manpower	Neosho	['Troubleshooting (Problem Solving)', 'Self Motivation', 'Poultry', 'Healthy', 'Preventive Ma
15	Careerbui	Today	Chocolate Candy Manufacturing - 3rd shift	Manpower	Scranton	['Quality Assurance', 'Corrective And Prev	$12.00/ hour
16	Careerbui	Today	Sytems Engineer /Support	ConsultNet	New York	['Python (Programming Language)', 'Confluence (Physical Geography)', 'Windows Servers',
17	Careerbui	Today	Machine Operators - Hilliard, OH	Manpower	Hilliard	['Manufacturing', 'Machinery']	$15,000.00 - $18,000.00 / year
18	Careerbui	Today	Food Operator	Manpower	Luverne	['Cargos', 'Warehousing', 'Manifests', 'War	$8.00 - $8.25 / hour
19	Careerbui	Today	General Labor - Hilliard, OH	Manpower	Hilliard	['Warehousing']	$13.00 - $15.00 / hour
20	Careerbui	Today	Extrusion Worker (Sanford, Maine)	Manpower	Sanford	['Standard Operating Procedure', 'Microso	$18.00/ hour

job_careerbuilder

Data Pre-processing

In [25]:
```
#removing columns Unnamed and From based on relevance
data = data.drop(columns=['Unnamed: 0'])
data = data.drop(columns=['from'])
data.shape
```

Out[25]:
```
(6583, 7)
```

In [26]:
```
#replacing 'day' with 'days' in date for better processing
data1 = data["date"].replace(to_replace ="1 day ago",
value = "1 days ago")
data["date"] = data1
data.date[124]
```

Out[26]:
```
'1 days ago'
```

In [27]:
```
#replacing date with numerical date values
for i in range(len(data)):
    if(data.date[i] == "Today"):
        data1 = data["date"].replace(to_replace ="Today",
        value = datetime.datetime.now().strftime("%x"))
        data["date"] = data1

    elif((data.date[i][-1]) == 'o' and data.date[i] != "30+ days ago"):
        today = datetime.datetime.now()
        daysAgo = int((data.date[i]).split(' days ago')[0])
        DD = datetime.timedelta(days= daysAgo)
        earlier = today - DD
        earlier_str = earlier.strftime("%x")
        data1 = data["date"].replace(to_replace = data.date[i],
        value = earlier_str)
        data["date"] = data1

data.head()
```

Removing irrelevant columns and standardizing date format

Data Pre-processing

In [43]:
```
#removing rows if column 'job_description' value is Null
data = data[pd.notnull(data['job_description'])]
#data.head()
```

In [29]:
```
#seperating list of skills into seperate rows for simplicity
dataTest = data.copy()
x = dataTest["skills"].apply(lambda s: list(ast.literal_eval(s)))
dataTest["skills"] = x
lst_col = 'skills'
dataTest.skills.apply(pd.Series)
dataTest = pd.DataFrame({
        col:np.repeat(dataTest[col].values, dataTest[lst_col].str.len())
       for col in dataTest.columns.difference([lst_col])
   }).assign(**{lst_col:np.concatenate(dataTest[lst_col].values)})[dataTest.columns.tolist()]
#data1 = dataTest[['date','job_title','job_company','skills','job_location']]
```

In [36]:
```
dataTest.head()
```

Out[36]:

	date	job_title	job_company	job_location	skills	salary	job_description
0	12/05/19	Senior Software Developer	Kforce Technology	Houston, TX	Advanced Audio Coding (Aac)	\n115,000.00–120,000.00 / year\n	Attending the agile ceremonies, performing so...
1	12/05/19	Senior Software Developer	Kforce Technology	Houston, TX	C Sharp (Programming Language)	\n115,000.00–120,000.00 / year\n	Attending the agile ceremonies, performing so...
2	12/05/19	Senior Software Developer	Kforce Technology	Houston, TX	C++ (Programming Language)	\n115,000.00–120,000.00 / year\n	Attending the agile ceremonies, performing so...
3	12/05/19	Senior Software Developer	Kforce Technology	Houston, TX	Software Development Life Cycle	\n115,000.00–120,000.00 / year\n	Attending the agile ceremonies, performing so...
4	12/05/19	Senior Software Developer	Kforce Technology	Houston, TX	.Net Framework	\n115,000.00–120,000.00 / year\n	Attending the agile ceremonies, performing so...

Dropping rows with null values based on column ‘job_description’ and separating list of skills to separate rows

Data Pre-processing

In [32]:

```
#Pre-processing for Graph for top 20 skills with maximum and minimum average salary
#getting salary values to float type
for i in range(len(eda_sal)):
    if("(Careerbuilder est.)" in eda_sal["salary"].iloc[i] and "$" in eda_sal["salary"].iloc[i]):
        val = eda_sal["salary"].iloc[i].split('(Careerbuilder est.)')
        strip = val[0].strip().replace(',', '')
        eda_sal["salary"].iloc[i] = float(strip[1::])
        data = [{'skills':eda_sal["skills"].iloc[i],'salary':eda_sal["salary"].iloc[i]}]
        df.loc[len(df.index)]=list(data[0].values())

    elif(" / year" in eda_sal["salary"].iloc[i] and "$" in eda_sal["salary"].iloc[i]):
        x = eda_sal["salary"].iloc[i].strip().split(' / year')
        val = x[0].replace('$', '').split(' - ')
        eda_sal["salary"].iloc[i] = float(val[1].replace(',',''))
        data = [{'skills':eda_sal["skills"].iloc[i],'salary':eda_sal["salary"].iloc[i]}]
        df.loc[len(df.index)]=list(data[0].values())

    elif(" / hour" in eda_sal["salary"].iloc[i] and "$" in eda_sal["salary"].iloc[i]):
        x = eda_sal["salary"].iloc[i].strip().split(' / hour')
        val = x[0].replace('$', '').split(' - ')
        eda_sal["salary"].iloc[i] = (float(val[1].replace(',',''))*40*52)
        data = [{'skills':eda_sal["skills"].iloc[i],'salary':eda_sal["salary"].iloc[i]}]
        df.loc[len(df.index)]=list(data[0].values())
```

In [33]:

```
df.head()
```

Out[33]:

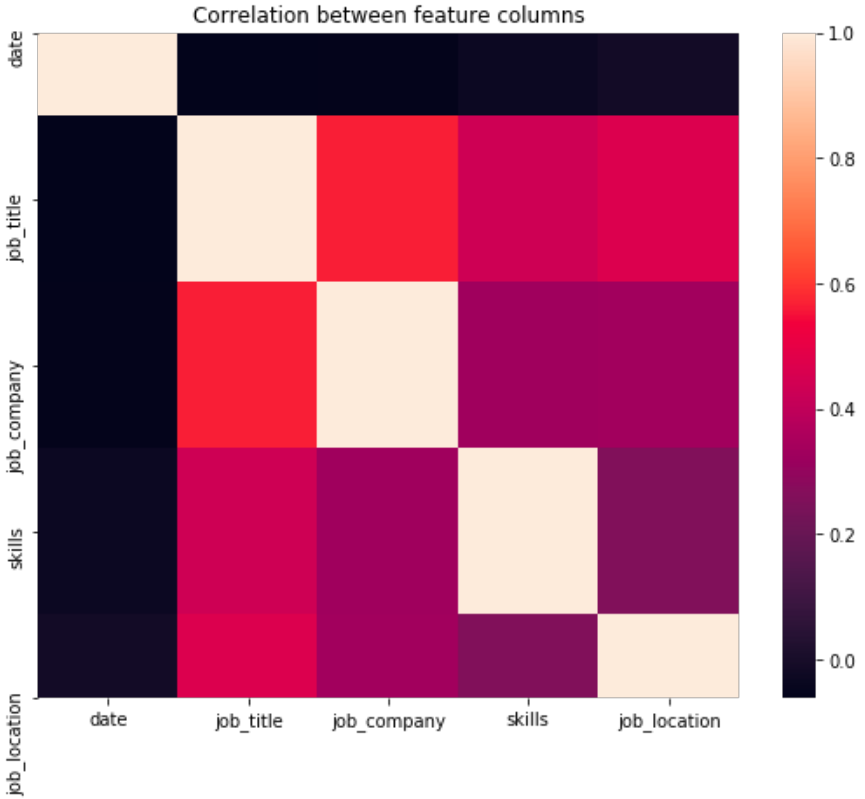
	skills	salary
0	Advanced Audio Coding (Aac)	120000.0
1	C Sharp (Programming Language)	120000.0
2	C++ (Programming Language)	120000.0
3	Software Development Life Cycle	120000.0
4	.Net Framework	120000.0

Getting salary values to float type

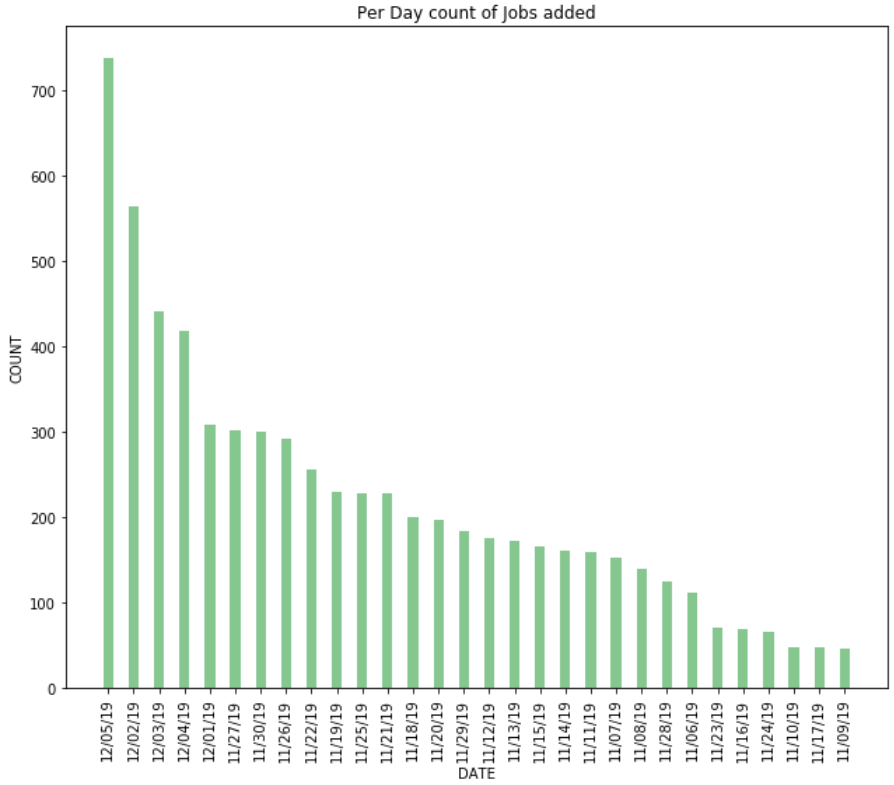
STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

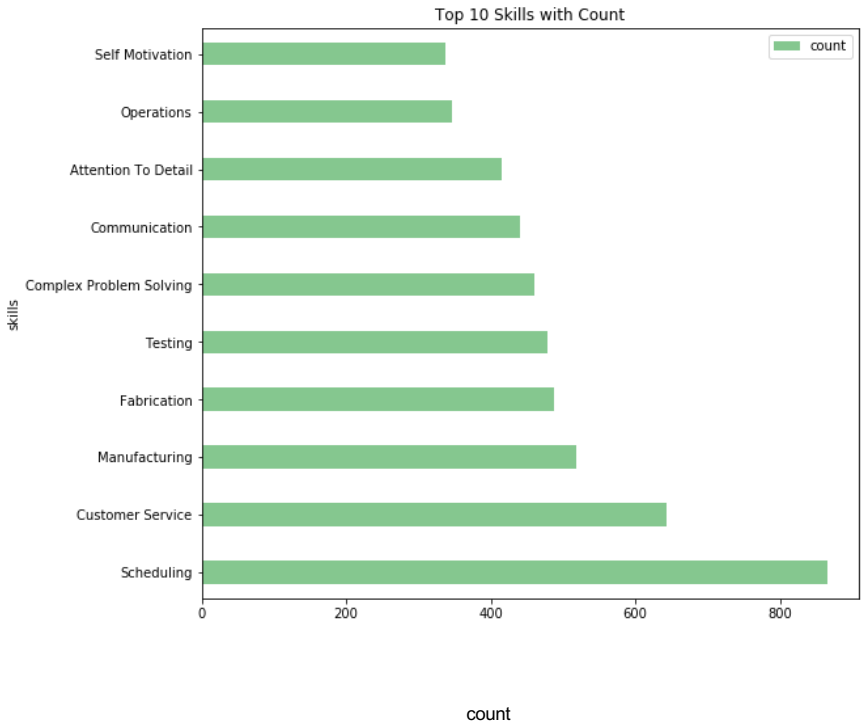


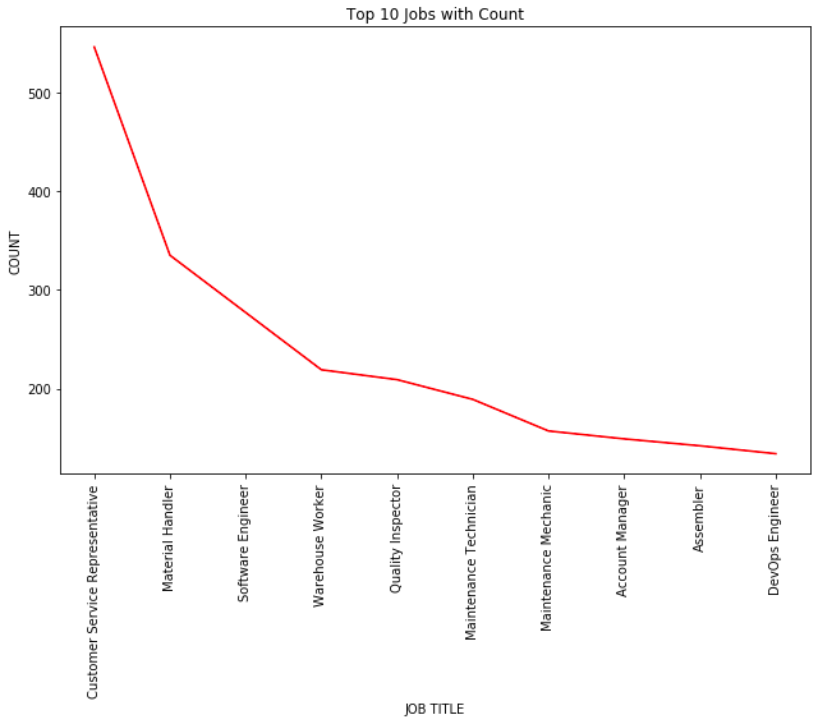
Exploratory Data Analysis (EDA)



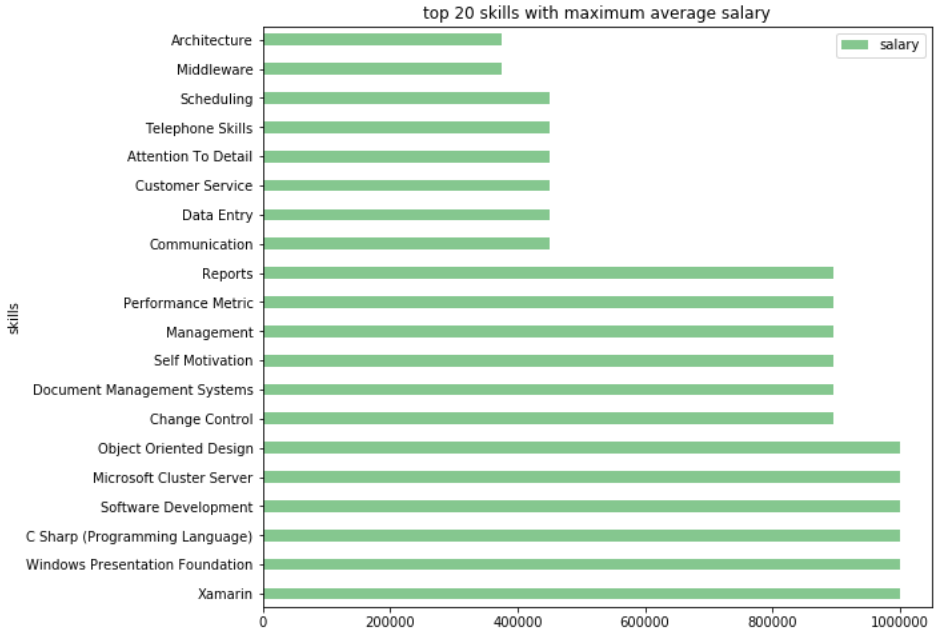
Dates from 9th Nov - 5th Dec

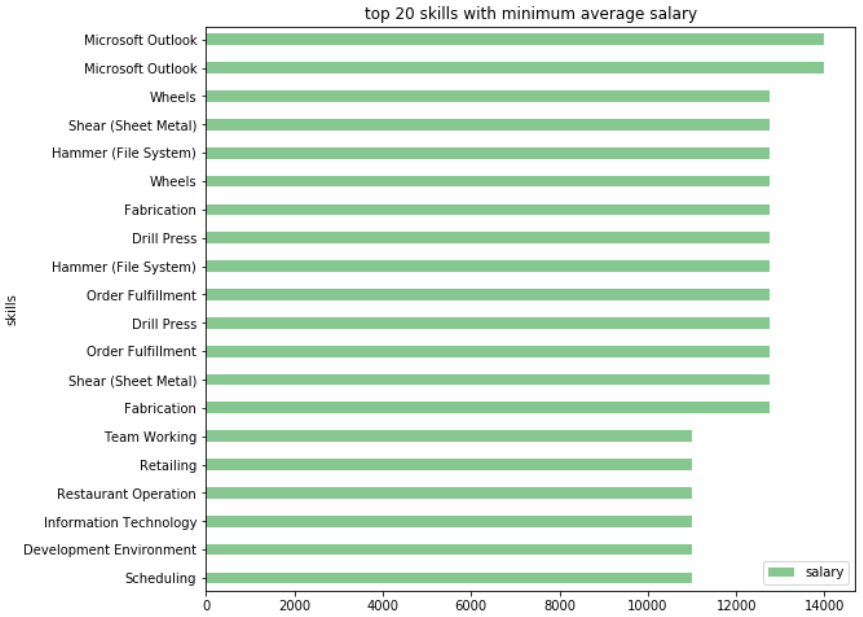
Exploratory Data Analysis (EDA)



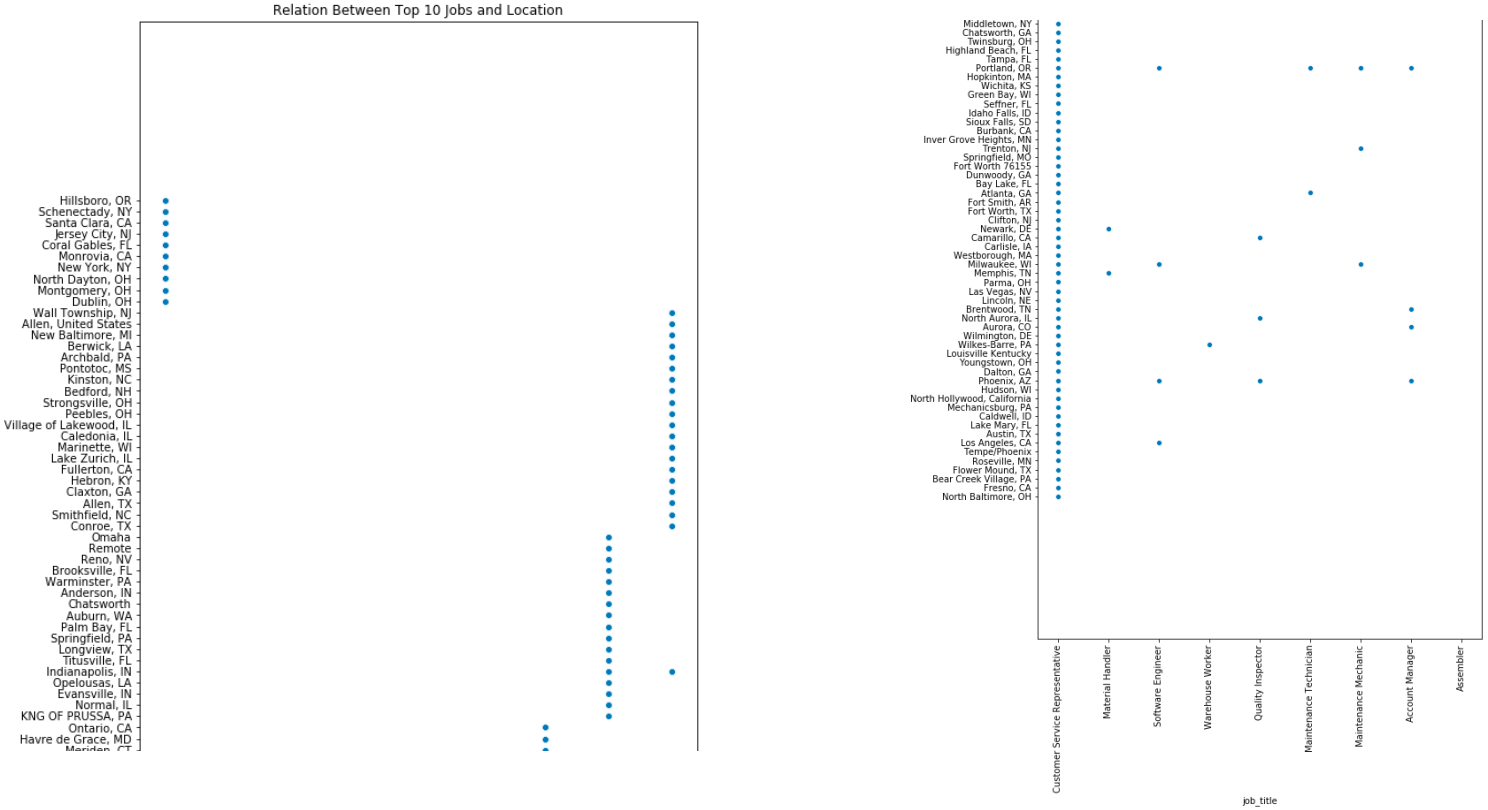


Exploratory Data Analysis (EDA)





Exploratory Data Analysis (EDA)



Extract Activity Phrases

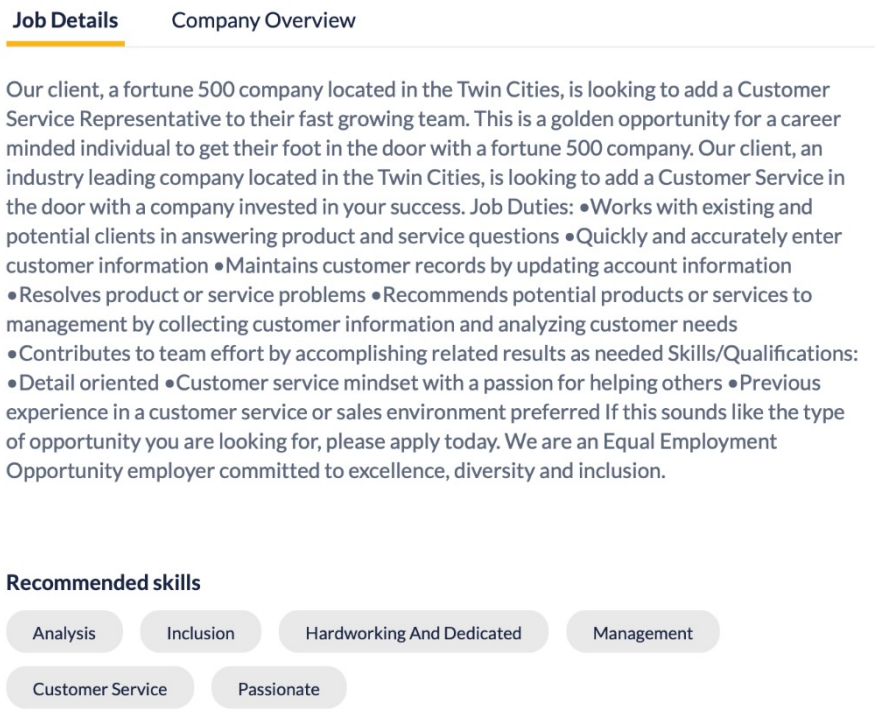
Why extract activity phrases??

- Recommended skills listed are abstract
- Give no idea about the context

Example:

Recommended Skill is “Analysis”

Doesn’t give clear picture of what is actually expected



Extract Activity Phrases

Implementation Idea

Part of Speech Tagging

- Using Bag of words approach doesn't capture structure of sentence
- It is also important to understand context
- Assigns part of speech to each word according to input sentence

Chunking

- Chunking works on top of POS tagging, it uses pos-tags as input and provides chunks as output.

Phrase Generator

Concept Used – Part of Speech Tagging

1. Part of Speech Tagging – Reads text and assign Part of Speech to each word
2. Examples of POS Tags – (NN, Noun) (VB, Verb) (JJ, Adjective) (RB, Adverb) etc.

Python library used: NLTK

Steps Involved:

- Input : Job Description as text
- Tokenize text using word_tokenize()
- Remove stop words
- apply pos_tag to above step that is nltk.pos_tag(tokenize_text)

Phrase Generator – Contd.

Concept Used - Chunking

1. Chunking - add more structure to the sentence by following parts of speech (POS) tagging
2. primary usage of chunking is to make a group of "noun phrases."

Rules for Chunking:

- No pre-defined rules, we can combine them according to need and requirement.
- Our rule – {<JJ.?>+<NN|NNS|NNP.?>+<NN|NNS|NNP.?>*}}
- Examples of chunks returned:

[‘strong experience JAVA’, ‘verbal communication skills’, ‘punctual delivery’, ‘trade-off analysis’ , ‘operational support’]

Similarity Check

Concept Used –Word Vectors, Sum and Mean of Word Vectors

- Train word2vec model for skill phrases obtained from Phrase Generator
- Word2vec model builds vectors of individual words
- Find mean vector for skill phrases

1. Use word vector of each word in skill phrase from the vocabulary
2. Add all word vectors
3. Divide it by the number of word in the phrase

- Use the mean word vector to compare skill phrases to recommended skills

- Examples of word vectors:

Construction Management [-2.53128121e-03 -1.77700538e-04 -2.65417900e-03 -3.87543434e-04
-1.15432322e-03 1.62815582e-03 1.00338110e-03 -1.52967067e-03.....-3.40627413e-03 -3.35135451e-03
-2.22289260e-03 3.08321090e-04 -3.39231896e-03 -2.66436837e-03
2.76020076e-03 4.11090790e-04 2.04549381e-03 9.39245452e-04
-1.60909526e-03 -1.57659827e-03 -1.14465761e-03 -3.12960986e-03
2.36541731e-03 -1.92798441e-03 -6.19727012e-04 -3.52221658e-03
-1.10748247e-03 -3.67690041e-03 3.10475053e-03 -1.39216939e-03]

internal company business systems [2.3599188e-03 -1.2206250e-03 7.9751259e-04 7.0600610e-05
-1.8862416e-03 -1.2368266e-04 -9.1530092e-06 8.7684812e-
.....
1.4695667e-03 -1.0500317e-03 -8.4408821e-04
-9.4342022e-04 5.0632062e-04 -3.1003775e-04 -1.4311107e-04
-4.1986880e-04 -1.4218048e-04 4.7879951e-04 -1.1398435e-03]

Similarity Check

Concept Used – Word Vectors, Sum and Mean of Word Vectors

Similarity Calculator:

Measure used -> Cosine Similarity

Python library used -> scipy import spatial

Finds the angle between the two vectors and calculates cosine value of that angle

Similarity score of 0.25 used as threshold

All the phrases greater than threshold value are selected

2	new features	Integration	1	2	0.35856548
20	strong knowledge software engineering discipline	Java (Programming Language)	3	0	0.273530692
20	strong knowledge software engineering discipline	C++ (Programming Language)	3	1	0.297596008
20	strong knowledge software engineering discipline	C (Programming Language)	3	2	0.279707402
20	workflow management tools Azkaban	Open AI	5	3	0.252860934
7	equivalent expertise	Aws Best Practices	6	0	0.254608423
23	new automation technologies	Systems Design	7	5	0.291011453
0	wireless analytics space IEEE	Systems Architecture	10	4	0.259406716
15	new processes	Data Integration	11	5	0.273268759
12	negotiate solutions Hardware/Software	Software Development Life Cycle	12	4	0.276912421
14	system-level computational architectures	Software Development Life Cycle	12	4	0.273831517
6	new projects	Integration	14	2	0.341670662
18	continuous improvement system stability performance	Architecture	14	1	0.335764796
1	multi-threaded software development experience Applic	Databases	15	1	0.273800403
6	continuous integration Software development experienc	Software Engineering	15	2	0.291767538
12	different problems	Software Engineering	20	5	0.282554746
15	reusable components	Kubernetes	20	0	0.252689928
11	related experience Master	Spring Framework	27	5	0.277940065
19	WORKING CONDITIONS Typical office environment	Hibernate (Java)	27	4	0.26278162
26	technical staff Establishes	Enterprise Java Beans	27	0	0.256945223
28	interactive Web User Interfaces	Hibernate (Java)	27	4	0.325454235
7	common Java Design	Java (Programming Language)	29	2	0.344111264
6	steady upward career progression art Software Engineer	Software Engineering	30	4	0.486126363
6	steady upward career progression art Software Engineer	Engineering	30	5	0.402050108

Word2Vec Model

- Pretrained Model with GloVe - Global Vectors for Word Representation
- GloVe is an unsupervised learning algorithm for obtaining vector representations for words
- Gigaword 5th Edition corpora (6B tokens, 400K vocab) with 100 dimensions

Clustering

- No labels for the matched activities- Unsupervised learning

- Used two clustering methods
  - K means clustering – Centroid model
  - Hierarchical clustering – Connectivity model

- Clustering performance evaluation (Based on cohesion and separation)
  - Silhouette Score (-1 to 1)

    S is calculated as : $s=b-a/\max(a,b)$ where,

    b - the mean distance between a sample and all other points in the next nearest cluster

    a- the mean distance between a sample and all other points in the same cluster
  - Calinski-Harabaz Index (Higher the better)

    S is calculated as : $s=b/a$ where

    b - mean between cluster separation

    a - mean within cluster separation

K-Means clustering

- It is a partitioning algorithm - partitions dataset into as number of clusters to minimize intra-partition distances
- Starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster
- Performs iterative (repetitive) calculations to optimize the positions of the centroids

Library – sklearn.cluster import Kmeans

Parameters:

- K – Number of clusters (7-11)
- Max_iterations (20 ,25)

Hierarchical Agglomerative Clustering

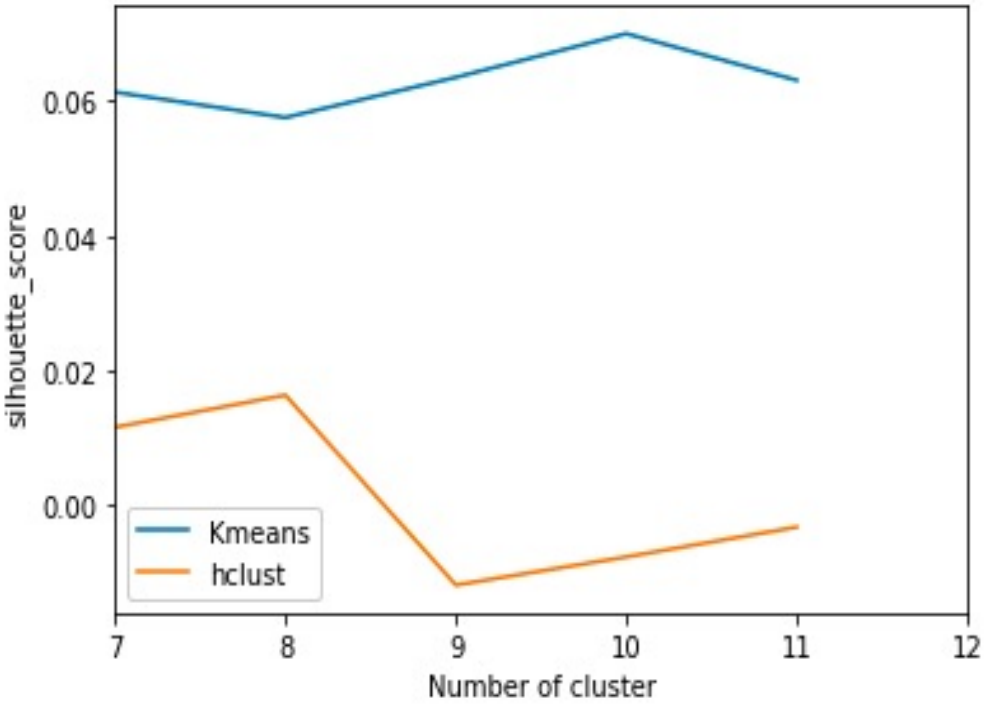
- In this technique, each data point is considered as an individual cluster.
- At each iteration, the similar clusters merge with other clusters until K clusters are formed.

- **Library - from sklearn.cluster import AgglomerativeClustering**

Parameters:

- Linkage = ward
- Number of clusters k =range(7,11)

Performance Analysis of Clustering Algorithms



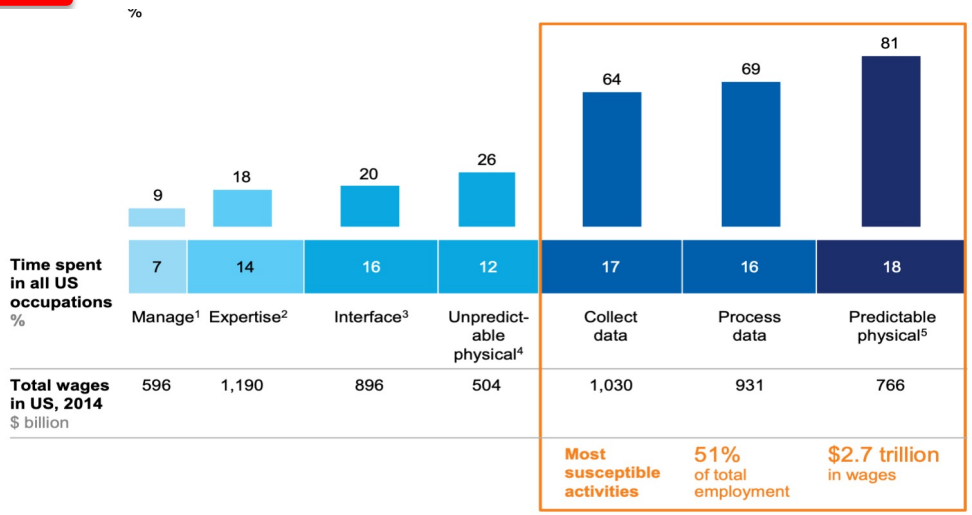
For clusters: 10
The average silhouette_score is : 0.06998661
The value of Calinski_value is : 233.82560805202831

Dimensions of activities and impact of Automation

- According to studies, all the activities in any occupation can be categorized into the following segments:

1. Managing and leading people
2. Expertise to decision making, planning and creative tasks
3. Interfacing with stakeholders
4. Collecting and Processing data
5. Physical work

MOST IMPACTED BY AUTOMATION



Source: US Bureau of Labor Statistics ; McKinsey Global Institute Analysis

Cluster Analysis

Physical work and Manual intervention

Cluster 0:

[‘able work’, ‘able lift’, ‘assigned work area’, ‘heavy work’, ’assist help callers’, ’regular basis help’, ’preferred ability lift’]

Cluster 1:

[‘on-site fabrication’, ‘hydraulic system’, ’disassemble mechanical equipment’, ’de-dusting equipment’]

Cluster 2:

[‘professional customer service’, Customer service experience Persuasive, ‘extraordinary customer service experience customers’]

Cluster 6:

[‘assist customer questions needs’,’ assist customer questions needs’,’ effective resolution issues’,’ protect quality condition items’]

Outlier Cluster:

[‘degree Information Systems’,’ degree Marketing’,’ degree Electrical Engineering’]

Expertise to decision making, planning and creative tasks

Cluster 3:

[‘service-oriented architecture’, ‘object oriented programming’, ‘third party auditors Judgment Decision Making’, ‘Complete video closeout audits’]

Cluster 4:

[‘critical thinking problem-solving skills’,’ deep consulting skills’,’ technical leadership’,’ Technical writing skills’]

Cluster 5:

[‘big data analysis application Data integration’,’ complex data models’,’ in-depth knowledge products services clients’]

Cluster 7:

[‘equivalent Docker Experience OpenShift Kubernetes’,’ Solid Experience Azure DevOps Able create Jenkins pipelines’,’ AWS Certified ’]

Cluster 8:

[‘implement optimize Data Engineering ETL’,’ stable Java applications’, ‘continuous integration server Fortran code’,’ popular Big Data frameworks Hadoop’]

Conclusion

- Converting sentences to word vectors and using the word vectors for K-Means clustering gives the best results for this project
- Value of k=10 is the most optimal value of k as the cohesion and separation scores are maximized at k=10

Future Work

- Phrase chunking using more patterns to improve the overall performance
- Check the clusters by using more clustering models(DBScan and BTM)
- Compare the activities based on the historical data of job postings and compare which all activities are automated
- Implementation of Doc2vec for activities and check cluster performance

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

THANK YOU!!

Questions???