SMA Lab 5: Clustering the job titles of LinkedIn Connections using Greedy Heuristic Algorithm

NAME : ANNAPOORNIMA S (225229101)

In [1]:
```
pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in c:\users\arulk\anaconda3
\lib\site-packages (1.2.2)
Requirement already satisfied: numpy>=1.17.3 in c:\users\arulk\anaconda3
\lib\site-packages (from scikit-learn) (1.24.3)
Requirement already satisfied: scipy>=1.3.2 in c:\users\arulk\anaconda3
\lib\site-packages (from scikit-learn) (1.10.1)
Requirement already satisfied: joblib>=1.1.1 in c:\users\arulk\anaconda3
\lib\site-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\arulk\an
aconda3\lib\site-packages (from scikit-learn) (2.2.0)
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
import pandas as pd
```

In [4]:
```
df = pd.read_csv('Coonections.csv')
df
```

Out[4]:

	First Name	Last Name	URL	Unnamed: 3	
0	Bennet	Samuel	https://www.linkedin.com/in/bennet-samuel-2361...	NaN	
1	Arockia	Rexy	https://www.linkedin.com/in/arockia-rexy-b2031...	NaN	
2	Princy	A	https://www.linkedin.com/in/princy-a-71b31a248	NaN	
3	quini	inisha	https://www.linkedin.com/in/quini-inisha-98156...	NaN	
4	Muhammad Ismaeel	Shareef S S	https://www.linkedin.com/in/sec-sha23	NaN	I
5	Sridhar	S	https://www.linkedin.com/in/sridhar-s-66a08224a	NaN	
6	Joshua	E	https://www.linkedin.com/in/joshua-e-0448b41b1	NaN	
7	Rethinagiri	G	https://www.linkedin.com/in/rethinagiri-g-0542...	NaN	
8	Pragadeesh	M	https://www.linkedin.com/in/kumarpragadeesh	NaN	SYNC
9	VIMAL	S E	https://www.linkedin.com/in/vimal-s-e-0a0186221	NaN	
10	Hariharan	S	https://www.linkedin.com/in/hariharan-s-12a016224	NaN	
11	Saranya	Santhanam	https://www.linkedin.com/in/saranya-santhanam-...	NaN	
12	ASHRAFALI	M	https://www.linkedin.com/in/ashrafali-m-769b25246	NaN	G
13	Santhana Pandi	P	https://www.linkedin.com/in/santhana-pandi-p-3...	NaN	
14	Allwín	Réx	https://www.linkedin.com/in/allw%C3%ADn-r%C3%A...	NaN	
15	Shree Krishna Kanth	S	https://www.linkedin.com/in/shree-krishna-kant...	NaN	
16	Hari Prasath	Senthil	https://www.linkedin.com/in/hari-prasath-senth...	NaN	
17	Hariharasudhan	D	https://www.linkedin.com/in/hariharasudhan-d-6...	NaN	M TECH
18	Harish	Mitha	https://www.linkedin.com/in/hareeshmitha	NaN	
19	Ezhilarasan	C	https://www.linkedin.com/in/ezhilarasan-c-3474...	NaN	

In [16]:
```
import pandas as pd
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [17]:
```
linkedin_connections = [
    "Software Engineer",
    "Data Analyst",
    "Product Manager",
    "Software Developer",
    "Data Scientist",
    "Software Engineer",
    "Data Engineer",
    "Product Manager",
    "Data Analyst",
    "Data Scientist",
    "Product Manager",
    "Software Engineer",
    "Data Engineer",
    "Data Scientist"
]
```

In [18]:

```
def greedy_clustering(titles, threshold=0.5):
    clusters = []
    similarity_matrix = calculate_cosine_similarity(titles)
    title_counts = pd.Series(titles).value_counts()
    sorted_titles = title_counts.index.tolist()

    for title in sorted_titles:
        added_to_cluster = False
        for cluster in clusters:
            cluster_similarity = np.mean(similarity_matrix[[titles.index(t

            if cluster_similarity >= threshold:
                cluster.append(title)
                added_to_cluster = True
                break

        if not added_to_cluster:
            clusters.append([title])

    return clusters

# Example usage
linkedin_connections = [
    "Software Engineer",
    "Data Scientist",
    "Product Manager",
    "Software Developer",
    "Data Analyst",
    "Product Designer",
    # Add more job titles as needed
]

clusters = greedy_clustering(linkedin_connections, threshold=0.4)

for i, cluster in enumerate(clusters):
    print(f"Cluster {i + 1}: {cluster}")
```

```
Cluster 1: ['Software Engineer', 'Software Developer']
Cluster 2: ['Data Scientist', 'Data Analyst']
Cluster 3: ['Product Manager', 'Product Designer']
```