Project – Predicting Skillsets

# Group Members:

1. Yogitha Mekala – Project Lead - <https://github.com/yogithamekala>
2. John Mahith Pagi - Data Scientist - <https://github.com/Johnmahith>
3. Maheshwar Rao Bandi - Research Analyst - <https://github.com/Mahehwarraobandi>

# 1. Introduction

This report summarizes the development of predictive models to analyze and forecast in-demand skillsets from job descriptions and tweets collected from Twitter. The objective is to extract meaningful insights into the required skills for various job roles, using Natural Language Processing (NLP) and Machine Learning models.

# 2. Data Collection

The dataset consists of job-related tweets collected using the Twitter API v2 through the tweepy library. A query was constructed to search for recent tweets containing keywords like 'skills required', 'we are hiring', 'job opening', and 'looking for developers with skills'. The retrieved tweets included details such as the timestamp, author ID, and the tweet text, which were stored in a pandas DataFrame for analysis.

# 3. Data Preprocessing

The raw tweet text was processed using various NLP techniques to clean and transform the text into meaningful data. The following steps were performed:  
1. \*\*Lowercasing\*\*: All text was converted to lowercase to maintain uniformity.  
2. \*\*Special Character Removal\*\*: Non-alphabetical characters and numbers were removed.  
3. \*\*Tokenization\*\*: The text was split into individual words using NLTK's word\_tokenize function.  
4. \*\*Stopword Removal\*\*: Common English stopwords were removed to retain only significant words.

# 4. Feature Extraction

The cleaned text was converted into a numerical representation using CountVectorizer, which transforms the text into a matrix of token counts. The feature names (i.e., individual words) were extracted to understand the vocabulary of the dataset. This step helps the machine learning model identify patterns in the text data.

# 5. Predictive Modeling

A KMeans clustering algorithm was used to group similar job descriptions into clusters based on the extracted features. The number of clusters was set to 4. Each job description was assigned to a cluster, and the top terms for each cluster were identified. These clusters help categorize the skillsets into meaningful groups, such as data analysis, machine learning, visualization skills, etc.

# 6. Results & Analysis

The KMeans clustering revealed the following insights:  
1. \*\*Cluster 0\*\*: Focused on SQL, data modeling, and machine learning.  
2. \*\*Cluster 1\*\*: Highlighted visualization tools like Tableau and experience requirements.  
3. \*\*Cluster 2\*\*: Centered on data analysis skills with Python expertise.  
4. \*\*Cluster 3\*\*: Focused on data science skills with tools such as pandas and numpy.  
These insights help identify the key skills required across different job roles.

# 7. Visualization

Bar plots were created to visualize the top terms within each cluster, providing a clear representation of the skillsets grouped by the clustering model. This visualization helps better understand the skill demands for various job roles.

# 8. Model Saving

The trained KMeans model and the CountVectorizer were saved using the joblib library for future use. These saved models can be reloaded to perform similar clustering on new datasets without retraining the model.

# 9. Conclusion

This project demonstrates how social media data and job descriptions can be analyzed using NLP and machine learning techniques to extract meaningful insights into skill demand. The clustering approach helps categorize job roles and the skills required for each. This information can assist employers and job seekers in understanding current trends in the job market.

**GitHub Link - h**ttps://github.com/Mahehwarraobandi/Assignment\_3\_datasciencecapstone