This assignment analyses text data using k-Means Algorithm an Unsupervised Machine Learning Algorithm.

K-Means Python Application

K-Means Python Application Executive Summary

Name: Mubanga Nsofu

Learner ID: 149050

Module: BAN6440 (Applied Machine Learning for Analytics)

Lecturer: Professor Bucciarelli

Date: 4th December 2024

Module Four Assignment

**1. Executive Summary: Phrase Clustering Application**

This application has been developed to cluster text phrases into distinct groups using machine learning. The application identifies similar phrases in the dataset and groups them into clusters using distance measurements. The dataset is sourced from the Registry of Open Data on AWS (Phrase Clustering Dataset, n.d.).

**2. Step by Step Process Followed:**

**Data Acquisition & Preprocessing:** The AWS CLI was used to download the dataset as discussed by (Zuckerberg, n.d.) and (AWS, 2024), and after that, feature extraction and standardisation. The entire process is automated using Python 3, and the user does not need to worry about manually downloading the data set.

**Clustering & Visualisation:** The K-means algorithm was used to cluster the pre-processed dataset. The results were then visualised using scatter plots and PCA (Principal Component Analysis), providing clear insights into cluster distributions. The visualisation is included in the submission.

**3. Results:**

The results show four clusters with different phrase patterns. PCA does a good job of creating the clusters, with a few overlaps. Clustering is a technique that helps uncover hidden patterns; in this case, cluster one had a shorter word count than cluster three.

**4. Conclusions**

This project shows the potential to automate the analysis of text data and help businesses mine their text data. Text data is typically dirty, and organisations need ways of tapping into actionable insights from text data.

**5. References:**

Phrase Clustering Dataset. (n.d.). *Phrase Clustering Dataset (PCD)*. Retrieved December 4, 2024, from <https://registry.opendata.aws/pcd>

Zuckerberg, C. (n.d.). *Download Data - CryoET Data Portal Documentation*. Retrieved December 4, 2024, from <https://chanzuckerberg.github.io/cryoet-data-portal/cryoet_data_portal_docsite_aws.html>

AWS. (2024). *What is the AWS Command Line Interface? - AWS Command Line Interface*. <https://docs.aws.amazon.com/cli/latest/userguide/cli-chap-welcome.html>