



# Data Glacier

Your Deep Learning Partner

## Week 10 Deliverables

- Name Manoj Nagaraja
- Email [manojn.7270@gmail.com](mailto:manojn.7270@gmail.com)
- Country United Kingdom
- College University of Liverpool
- Specialization Data Science
- Github Repo Link [https://github.com/ManojN7270/Final-Project-week\\_7-to-week\\_13.git](https://github.com/ManojN7270/Final-Project-week_7-to-week_13.git)

### *Problem description*

Requires to implement various clustering algorithms using the Python programming language and apply them to cluster a given dataset. The purpose of this project is to assess the understanding of various clustering algorithms by implementing the algorithms and applying them to text clustering.

### *EDA performed on the data*

Implementation of k-means and k-means++ clustering algorithms on a dataset of word embeddings. The code first reads the dataset containing the word embeddings, creates a numpy array to store the embeddings and then performs k-means clustering on the embeddings. The number of clusters (k) is set to 5. The code also computes the silhouette coefficient for each value of k ranging from 2 to 9 and plots the silhouette coefficients against the number of clusters. The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a value of 1 indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.

After performing k-means clustering, the code implements the k-means++ algorithm for clustering. It defines a class K-means++ that contains the fit and predict methods to perform k-means++ clustering on the dataset. The fit method initializes the centroids using the k-means++ algorithm and then updates the centroids for a maximum of 100 iterations. The predict method assigns each data point to its closest centroid. Finally, the code reads the file containing the word embeddings, creates a numpy array to store the embeddings, and performs k-means++ clustering on the embeddings. The number of clusters is set to 3, and the maximum number of iterations is set to 100.