Lab 2: k-Means Clustering (5 points total)

In this lab you will have to implement k-Means Clustering clusterization algorithm using pure Python and NumPy. Your task is to perform clusterization on a dataset of texts, which includes reviews on Netflix.

What you will have to do:

Part 1: Choose open-source LLM to generate embeddings [link].

Part 2: Download dataset “Netflix Reviews” [link]. Select at least 2000 random objects of data. Consider the “score” column, so the data will be uniformly distributed alongside this column. You are interested in columns “content” and “score”.

Part 3: Use LLM of your choice to create embeddings of your data. Put them on the table.

Part 4: Perform clusterization with different types of distance metrics [‘euclidean’, ‘minkowski’, ‘mahattan’]. Find and evaluate intersection betwenn column “content” and results of clusterization.

Part 5: Visualize results and write a conclusion.

Requirements:

- Explain and justify each step of your solution.

Submission:

- Deadline: Friday (June 21), 23:59
- Submit exactly one .ipynb file