CA Assignment 2
Data Clustering
Implementing clustering algorithms

Assessment Information

Assignment Number	2 (of 2)
Weighting	15%
Assignment Circulated	3 March
Deadline	24 March 17:00
Submission Mode	Electronic via Canvas
Purpose of assessment	The purpose of this assignment is to assess the understanding of various clustering algorithms by implementing the algorithms and applying them to text clustering.
Learning outcome assessed	(1) A critical awareness of current problems and research issues in data mining.

Objectives

This assignment requires you to implement various clustering algorithms using the Python programming language and apply them to cluster a given dataset.

Assignment description

In the assignment, you are required to cluster words. The words are stored in a file that you will find in the archive *CA2data.zip*. The first entry in each line is a word followed by 300 features (word embedding) describing the meaning of that word.

Questions/Tasks

1. **(20 marks)** Explain the k-means clustering algorithm. Provide pseudo code of the algorithm. It should be the version of the k-means clustering algorithm discussed in the lectures. Implement the k-means clustering algorithm following your explanation and the pseudo code. In the implementation, select initial cluster representatives randomly.
2. **(20 marks)** Explain the k-means++ clustering algorithm. Provide pseudo code of the algorithm. It should be the version of the k-means++ clustering algorithm discussed in the lectures. Implement the k-means++ clustering algorithm following your explanation and the pseudo code.
3. **(20 marks)** Explain the Bisecting k-Means hierarchical clustering algorithm. Provide pseudo code of the algorithm. It should be the version of the Bisecting k-Means clustering algorithm discussed in the lectures. Implement the Bisecting k-Means algorithm following your explanation and the pseudo code.
4. **(10 marks)** Run the k-means clustering algorithm you implemented in part (1) to cluster the given instances. Vary the value of k from 1 to 9 and compute the Silhouette coefficient for each set of clusters. Plot k in the horizontal axis and the Silhouette coefficient in the vertical axis in the same plot.
5. **(10 marks)** Run the k-means++ clustering algorithm you implemented in part (2) to cluster the given instances. Vary the value of k from 1 to 9 and compute the Silhouette coefficient for each set of clusters. Plot k in the horizontal axis and the Silhouette coefficient in the vertical axis in the same plot.
6. **(10 marks)** Run the Bisecting k-Means algorithm you implemented in part (3) to compute a hierarchy of clusterings that refines the initial single cluster to 9 clusters. For each s from 1 to 9, extract from the hierarchy of clusterings the clustering with s clusters and compute the Silhouette coefficient for this clustering. Plot s in the horizontal axis and the Silhouette coefficient in the vertical axis in the same plot.
7. **(10 marks)** Comparing the different clusterings you obtained in (4)-(6), discuss in which setting you obtained best clustering for this dataset.

Submission Instructions

Submit via Canvas the following **two** files (**please do NOT zip files into an archive**)

1. the source code for all your programs (**do not provide ipython/jupyter/colab notebooks, instead submit standalone code in a single .py file**), and
2. a PDF file (report) of **no more than 5 pages** providing the answers to the questions.

It is extremely important that you provide the two files described above and not just the source code!

Important notes

(read carefully and double check compliance before submission)

1. No credit will be given for implementing any other type of clustering algorithms or using an existing library for clustering instead of implementing it by yourself. However, you are allowed to use
   - numpy library (any function);
   - random module;
   - matplotlib for plotting; and
   - pandas.read_csv, csv.reader, or similar modules **only** for reading data from the files.

   However, it is not a requirement of the assignment to use any of those modules.
2. Your program
   - should run and produce **all results for Questions 4, 5, and 6** in one click without requiring any changes to the code;
   - should output only the required data in a clearly structured way; it should NOT output any intermediate steps;
   - should assume that the input file is named 'dataset' and is located in the same folder as the program; in particular, it should NOT use absolute paths.
3. Programs that do not run will result in a mark of zero!
4. Your code should be as clear as possible and should contain only the functionality needed to answer the questions. Provide as much comments as needed to make sure that the logic of the code is clear enough to a marker. Marks may be deducted if the code is obscure, implements unnecessary functionality, or is overly complicated.
5. If you use module random to make some random actions, use a fixed seed value so that your program always produces the same output. This output should be exactly the one that you provide in the PDF report.
6. Your answers in the PDF report should be succinct, but complete and clear. The clarity and presentation of the report will be assessed.
7. The report should contain the explanations of the algortihms and their pseudo codes (for Questions 1,2,3) and the answer to Question 7. The python code of the implementation of the algorithms should be included in the .py file, and not in the report.
8. You are allowed to (re)use any part of the function that computes Silhouette coefficient from the solution to the lab tasks for Week 7.
9. Your submission should be **your own** work. Do not copy or share! Make sure that you clearly understand the severity of penalties for academic misconduct (https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/appendix_L_cop_assess.pdf).