COMP527 - JAN21 - CA Assignment 2
Data Clustering
Implementing the k-means and k-medians clustering algorithm

Assessment Information

Assignment Number	2 (of 2)
Weighting	15%
Assignment Circulated	08th June 2021
Deadline	28th July 2021, 17:00 UK Time (UTC)
Submission Mode	Electronic via Canvas
Learning outcome assessed	(1) A critical awareness of current problems and research issues in data mining.
Purpose of assessment	This assignment assess the understanding of k-means clustering algorithm by implementing k-means for text clustering.
Marking criteria	Marks for each question are indicated under the corresponding question.
Late Submission Penalty	Standard UoL Policy applies.

1 Submission Instructions

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

1. the source code for all your programs0 **(do not provide ipython/jupyter/colab notebooks, instead submit standalone code in a single .py file)**
2. a README file (plain text) describing how to compile/run your code to produce the various results required by the assignment, and
3. a PDF file providing the answer to the questions.

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

2 Objectives

This assignment requires you to implement the k-means and k-medians clustering algorithm using the Python programming language.

> **No credit will be given for implementing any other types of clustering algorithms or using an existing library for clustering instead of implementing it by yourself.However, you are allowed to use numpy library for accessing data structures such as numpy array. But it is not a requirement of the assignment to use numpy. You can use matplotlib for plotting but it is not compulsory to use matplotlib. You must provide a README file describing how to run your code to re-produce your results. Programs that do not run will result in a mark of zero!**

3 Assignment Description

In the assignment, you are required to cluster words belonging to four categories: *animals*, *countries*, *fruits* and *veggies*. The words are arranged into four different files 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

(1) **(25 marks)** Implement the k-means clustering algorithm to cluster the instances into k clusters.

(2) **(25 marks)** Implement the k-medians clustering algorithm to cluster the instances into k clusters.

(3) **(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 B-CUBED precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and the B-CUBED precision, recall and F-score in the vertical axis in the same plot.

(4) **(10 marks)** Now re-run the k-means clustering algorithm you implemented in part (1) but normalise each object (vector) to unit ℓ_2 length before clustering. Vary the value of k from 1 to 9 and compute the B-CUBED precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and the B-CUBED precision, recall and F-score in the vertical axis in the same plot.

(5) **(10 marks)** Run the k-medians clustering algorithm you implemented in part (2) over the unnormalised objects. Vary the value of k from 1 to 9 and compute the B-CUBED precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and the B-CUBED precision, recall and F-score in the vertical axis in the same plot.

(6) **(10 marks)** Now re-run the k-medians clustering algorithm you implemented in part (2) but normalise each object (vector) to unit ℓ_2 length before clustering. Vary the value of k from 1 to 9 and compute the B-CUBED precision, recall, and F-score for each set of clusters. Plot k

in the horizontal axis and the B-CUBED precision, recall and F-score in the vertical axis in the same plot.

(7) **(10 marks)** Comparing the different clusterings you obtained in (3)-(6), discuss in which setting you obtained best clustering for this dataset.