COMP 527 - 2016 - 1 CA Re-sit Assignment
Data Clustering
Implementing hierarchical clustering algorithms

Assessment Information

Assignment Number	1 (of 2 Re-sit)
Weighting	12%
Assignment Circulated	Wednesday, 20th July 2016
Deadline	Wednesday, 3rd August 2016, 15:00 UK Time (UTC)
Submission Mode	by email to danushka.bollegala@liverpool.ac.uk
Learning outcome assessed	(1) A critical awareness of current problems and research issues in data mining. (3) The ability to consistently apply knowledge concerning current data mining research issues in an original manner and produce work which is at the forefront of current developments in the sub-discipline of data mining.
Purpose of assessment Marking criteria	This assignment assesses the understanding of hierarchical clustering algorithm. Marks for each question are indicated under the corresponding question.
Submission necessary in order to satisfy Module requirements?	No
Late Submission Penalty	Standard UoL Policy.

1 Objectives

This assignment requires you to implement hierarchical clustering algorithms using the Python programming language.

> **Note that *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 and scipy libraries for accessing data structures such as numpy.array or scipy.sparse. But it is not a requirement of the assignment to use numpy or scipy. You must provide a README file describing how to run your code to produce the results. Programs that do not run will result in a mark of zero!***

2 Document Clustering using Hierarchical Clustering Algorithms

In this assignment, you are required to cluster Amazon product reviews that belong to four product categories: *books*, *electronic appliances*, *dvds*, and *kitchen appliances*. Moreover, each category is further divided into positive-valued sentiment reviews and negative-valued sentiment reviews. In total, you will find reviews that belong to $4 \times 2 = 8$ categories in the data file provided for the assignment from `http://cgi.csc.liv.ac.uk/~danushka/lect/dm/data.txt`.

The format of the data file is as follows. Each line of the data file corresponds to one review. The first element in the line represents the label (eg. *kitchen-positive* indicates that the review is a positive sentiment review about some kitchen appliance) of the instance. The next elements (separated by spaces) in the line represent the unigram and bigram features extracted from the review. Note that the two words in a bigram feature are connected by two underscores. Reviews are represented using binary-valued features (ie. each feature appears exactly once in a given line).

Questions

(1) Write a program to load the data instances to memory from the provided file *data.txt*. (**10 marks**)

(2) Implement Group Average Hierarchical Clustering (GAAC) using Euclidean distance as the distance measure. (**40 marks**)

(3) Implement the Complete Linkage clustering algorithm using Euclidean distance as the distance measure. (**30 marks**)

(4) Compare the run time of the Group Average Hierarchical Clustering and Complete Linkage Clustering algorithms that you implemented above (**20 marks**)

3 Deadline and Submission Instructions

- Deadline for submitting this assignment is Wednesday **3rd August 2015, 15:00 UK time (UTC)**
- Submit

(a) the source code for all your programs,

(b) a README file (plain text) describing how to compile/run your code to produce the various results required by the assignment, and

(c) a PDF file providing the answers to the question (4).

Compress all of the above files into a single tar ball (tgz) file and specify the filename as *studentid.tgz*. It is extremely important that you provide all the files described above and not just the source code! (If you are unable to create a tgz file then create a zip file)

- Submission is via email to danushka.bollegala@liverpool.ac.uk.