Given data set has 150 papers which were accepted in AAAI this year. The submissions were found to span over several different domains of computer science such as: Machine Learning, Optimization, Knowledge-Based systems, Robotics, Natural Language Processing, etc.

In this dataset you will find the following relevant attributes of each paper.

● Title: Free Text; Title of the paper

● Keywords: Free Text: author-generated keywords

● Topics: Categorical; author-selected, low-level keywords from conference-provided list

● High-level Domains: Categorical; author-selected, high-level keywords from conference-provided list. There are 9 distinct high-level domains in the dataset.

● Abstract: Free Text: abstract of the paper

AAAI wants an automated unsupervised script to group these documents into different clusters, so that all papers having similar high-level domains will be grouped together. For example: Let's assume there is a paper by Harish on a novel Clustering Algorithms, and there is another paper by Surya on a novel Classification Algorithm. Topics of the two papers might be different such as: {Clustering, Unsupervised, Machine Learning} and {Classification, Supervised, Machine Learning}, however both come under the same high-level domain, i.e., Machine Learning.

To complete this task there has to be a notion of similarity among different papers. For this assignment, the simple jaccard coefficient of two sets of topics is considered as the notion of similarity between the two papers

Implement a bottom-up hierarchical clustering algorithm considering the aforementioned notion of similarity, to find 9 (nine) clusters using both the (i) complete linkage and (ii) single linkage strategies.

State the clusters identified in your report.