# Machine Learning (CS60050)

Assignment 3: Clustering
Due date: **31st March 2019 (EOD)**

---

**AAAI needs your help !!!**

The Association for the Advancement of Artificial Intelligence (AAAI) organizes a conference of Artificial Intelligence, which is one of the most prestigious conferences relevant to the field of AI. Every year many researchers across the world submit to the conference and after a rigorous process of review and evaluation, only a selected set of papers get accepted. 150 such papers got 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..

You can find a dataset containing all the accepted submissions **here**. 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. For example, if the set of topics for a paper is H = {Clustering, Unsupervised, Machine Learning} and if the set of topics for another paper is S = {Classification, Supervised, Machine Learning}, then the Jaccard Coefficient between the two papers is $JC_{HS} = JC_{SH} = \frac{|H \cap S|}{|H \cup S|} = \frac{1}{5} = 0.2$.

1. 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.

2. With the same notion of similarity, design a graph where each node is a research paper, and an edge is to be drawn between two nodes if their topics are very similar (you can decide some threshold for this purpose, based on the quality of the clusters - see later). Now apply the **Girvan-Newman clustering algorithm** on this graph to find **9 (nine)**

clusters. State the clusters identified in your report. **For this part, you can use any graph library (e.g., the Networkx package in Python, or igraph library in C) for creating / updating the graph, and evaluating the centralities. However, direct use of any implementation of Girvan-Newman clustering algorithm will be penalised.**

3. Consider the 'gold standard' clustering to be based on the high-level domains associated with the papers, i.e., all papers of a particular high-level domain (according to the dataset) constitute one cluster.
   Now you have three sets of clusters (having 9 clusters each) of the given set of research papers, one identified by the hierarchical clustering method with complete linkage, the second identified by the hierarchical clustering method with single linkage, and the third identified by the graph-based method.
   The final step is to evaluate the quality of these three clusterings. To this end, implement the **Normalized Mutual Information (NMI)** metric mentioned in this document to evaluate the quality of the clusters that you have got. Report the NMI values of the three clusterings.

Note that the clustering algorithms that you will implement in Parts 1 and 2 should use only the Topics (low-level keywords) in the data; they should NOT use the High-level domains. The High-level domains should be used only in Part 3, for evaluation of the clusterings.

## Submission instructions

---

For each part, you should submit the source code and all result files. Write a separate source code file for each part. **You should include a README file describing how to execute each of your codes**, so that the evaluators can test your code.

**You can use C / C++ / Java / Python for writing the codes; no other programming language is allowed. You cannot use any library/module meant for clustering or Machine Learning. You can use libraries for other purposes, such as generation and formatting of data, and a graph library for Part 2 (as specified above), but NOT for the ML part. Also you should not use any code available on the Web. Submissions found to be plagiarised or having used ML libraries will be awarded zero marks for all the students concerned.**

Along with the source codes and the results, you should submit a report (pdf) including the following:

- For each of the three clustering methods -- final learned clusters and the number of entities in each of them.
- NMI values of the three clusterings you obtained
- What are the thresholds that you are considering while doing the assignments have to be clearly mentioned. Note that you can decide upon what sort of thresholding you want to do based on the goodness of your clusters.

All source codes, data and result files, and the final report **must be uploaded via the course Moodle page, as a single compressed file (.tar.gz or .zip)**. The compressed file should be named as: **{ROLL_NUMBER}_ML_A3.zip or {ROLL_NUMBER}_ML_A3.tar.gz**
Example: If your roll number is 16CS60R00, then your submission file should be named as 16CS60R00_ML_A3.tar.gz or 16CS60R00_ML_A3.zip

*****Note that the evaluators can deduct marks if the deliverables are not found in the way that has been asked for the assignment.**

**Submission deadline: March 31, 2019, 23:59 IST [hard deadline]**

For any questions about the assignment, contact the following TAs:

1. Abhisek Dash (assignmentad @ gmail . com)
2. Paheli Bhattacharya (paheli.cse.iitkgp @ gmail . com)