**Project 4 – Clustering**

**CS548 / BCB503 / CS583 Knowledge Discovery and Data Mining - Fall 2019**

**Prof. Carolina Ruiz**

**Students:** <replace this with your names **in alphabetical order by last name**>

|  |  |
| --- | --- |
| **Dataset :*** Dataset Description (not needed in this project)
* Data Exploration
* Initial Data Preprocessing (if any)
 | /05/05 |
| **Code Description:**  | /40 |
| **Experiments:** Guiding Questions | /10 |
| K-means - Sufficient & coherent set of experiments | /10 |
| * Objectives, Parameters, Additional Pre/Post-processing
 | /10 |
| * Presentation of results
 | /10 |
| * Analysis of individual experiments’ results
 | /10 |
| Hierarchical - Sufficient & coherent set of experiments | /10 |
| * Objectives, Parameters, Additional Pre/Post-processing
 | /10 |
| * Presentation of results
 | /10 |
| * Analysis of individual experiments’ results
 | /10 |
| DBSCAN - Sufficient & coherent set of experiments | /10 |
| * Objectives, Parameters, Additional Pre/Post-processing
 | /10 |
| * Presentation of results
 | /10 |
| * Analysis of individual experiments’ results
 | /10 |
| Quantitative Analysis of Results and Discussion  | /45 |
| Qualitative Analysis of Results, Discussion, and Visualizations  | /45 |
| Advanced Topic | /30 |
| Total Written Report Project 4 | /300 = /100 |

**Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)**

**[0 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)**

*(no needed in this project as this is the same dataset that you have used in previous two projects)*

**[05 points] Data Exploration: (e.g., comments on aspects of the dataset THAT ARE RELEVANT FOR CLUSTERING . This could include visualizations, issues with the data, and so on.)**

**[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, necessary dimensionality reduction, converting attributes to numeric, scaling attributes if needed, and so on.)**

**Code Description: Python Libraries and Functions you used and what parameters you experimented with. (At most 1.75 page.)**

**[05 points] Preprocessing Techniques for Clustering:**

**[05 points] K-means Clustering:**

**[05 points] Hierarchical Clustering:**

**[05 points] DBSCAN:**

**[10 points] Quantitative Clustering Evaluation: including metrics listed on the project description and possibly others you used**

**[10 points] Qualitative Clustering Evaluation: using Visualization, including MDS and at least one more visualization technique (e.g., heatmap of the correlation between proximity matrix and incidence matrix) you used**

**[10 points] Three Guiding Questions about the dataset domain that can be answered by Clustering methods (at most 1/4 page):**

1. **…**
2. **…**
3. **…**

|  |
| --- |
| **[40 points] Summary of Experiments with K-means.** *At most 1 page.* |
|  | **Pre-process** | **# clusters** | **Distance****function****(if not Euclidean)** | **#** **iterations** | **SSE** | **% of instances** **per cluster** | **Observations about experiment** **(e.g., observations from visualization,****interpretation of centroids,****analysis of similarity among instances in the same cluster)** | **You can add** **other columns****or remove this one** |
| P1 |  |  |  |  |  |  |  |  |
| P2 |  |  |  |  |  |  |  |  |
| P3 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |

|  |
| --- |
| **[40 points] Summary of Experiments with Hierarchical Clustering (single link, complete link, average, Ward).** *At most 1 page.* |
|  | **Pre-process** | **# clusters** | **Link****type** | **Time****taken** | **% of instances** **per cluster** | **Observations about experiment****(e.g., observations from visualization,****analysis of nested clusters,****analysis of similarity among instances in the same cluster)** | **You can add** **other columns** **or remove this one** |
| H1 |  |  |  |  |  |  |  |
| H2 |  |  |  |  |  |  |  |
| H3 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |

|  |
| --- |
| **[20 points] Summary of Experiments with DBSCAN** *At most 1 page.* |
|  | **Pre-process** | **Epsilon** | **minPts** | **#** **clusters** | **Time taken** | **% of instances** **per cluster** | **Observations about experiment****(e.g., observations from visualization,****analysis of core, border and noise points,****analysis of similarity among instances in the same cluster)** | **You can add** **columns** |
| D1 |  |  |  |  |  |  |  |  |
| D2 |  |  |  |  |  |  |  |  |
| D3 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |

**[45 points] Quantitative Analysis of Results and Discussion (at most 2 pages).**

Include here: (1) Calculations or quantitative analysis you did to obtain good initial parameter values for K-means and DBSCAN; evaluation of clusterings using (2) internal indices, (3) relative indices and (4) external indices; and (5) other quantitative results across experiments and clustering methods. Explain your work.

**[45 points] Qualitative Analysis of Weka and Python Results on and Visualizations (at most 2 pages)**

Include here (1) visualizations of the clustering using MDS and other visualization methods; (2) inspection of the actual clusters' members to find similarities among data instances in a cluster and dissimilarities with data instances in different clusters; (3) additional analysis of the results from the point of view of the dataset domain; and (4) answers that the experiments provided to your guiding questions.

**Advanced Topic: <include name of the topic here> (at most 1 page)**

**[7 points] List of sources/books/papers used for this topic (include URLs if available):**

* …
* …
* …

...

**[20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.**

**[3 points] How does this topic relate to clustering?**

**Authorship:** Although each student on the team is expected to be involved in every aspect of the project, describe in detail here the main contributions that each of the team members made to this project. This authorship description must accurately reflect the work done by each team member, and must be approved by all of the members of the team (at most 1/3 page)