The task is to identify clusters of records in a given dataset. Similarity is calculated based on the provided list of fields, type of comparison and weight. Comparisons: **Exact** is completely equal values (case insensitive), **Fuzzy** is similar values.

Input:

1. Records in CSV format with arbitrary columns, example

|  |  |  |  |
| --- | --- | --- | --- |
| Record Id | Name | Email | Address |
| 00011 | John Smith | john@apple.com | 123 Main St, London, UK |
| 00022 | Alice Smith | [alice@company.com](mailto:alice@company.com) | 123 Main St, London, UK |
| 00033 | Alice Smith | [alice@company.com](mailto:alice@company.com) | 345 Main St, London, UK |

1. A list of fields, example below, types: Exact or Fuzzy, weight: 0-100

[

{field: “name”, type: “Exact”, weight: 50},

{field: “email”, type: “Exact”, weight: 20},

{field: “address”, type: “Fuzzy”, weight: 30}

]

1. Match threshold for clustering 1-100.

Output:

A CSV list of clustered records with cluster info, which similarity is higher or equal to the “match threshold”. Example

|  |  |  |  |
| --- | --- | --- | --- |
| Record Id | Cluster Id | Cluster size | Score |
| 00011 |  |  |  |
| 00022 | 00022 | 2 | 100 |
| 00033 | 00022 | 2 | 100 |
| 00044 | 00044 | 2 | 70 |
| 00055 | 00044 | 2 | 70 |

The workload will be executed in AWS.

Requirements: The algorithm should be able to process datasets with 10M records on a 16 CPU/32Gb RAM within minutes, not hours.