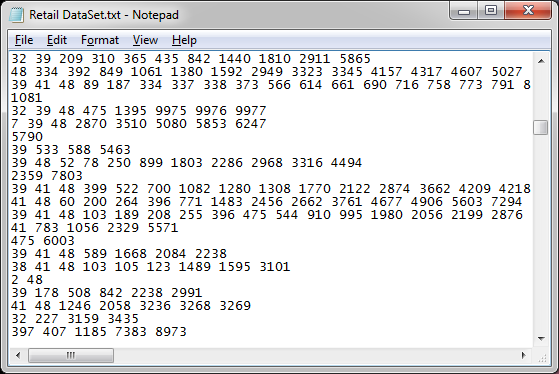
**Abstract.** This project aims to mining Sequential pattern data by using Genetic algorithm based on Hadoop Mapreduce. The basic idea is to extract the fittest rules from a huge sequential datasets.

* ***The Framework should be as following:***

1. The job should be done with many rounds until a **maximum number of generations** is reached.
2. **First round :** 
   1. **Map Class**: Each Mapper has its own sub-population → each value consider as a chromosome→ Represent each chromosomes as a Vertical Bitmap Representation (0s and 1s)→Generate **N** possible rules from each record (each line); Here we defined the **maximum number of generations** to check if we have to proceed or stop.
   2. **Combiner** : Should calculate the fitness value for each rule within its corresponding sub-population. The fitness is the result of (Support \* Confidence) , where “Conf(sp|s) divide number of s and sp occurrence, in the database, over number of s occurrence; while Support(s) divide number of s occurrence, in the sub-population, over number of all sequences available in the database”. In other word, Support who many times the rule repeated /**N, and the Confidence = Support/ Support of Antecedent of the rule.**
   3. **Reducer**: Select the best rules (highest finesses) which are greater than or equal   
      **minimum\_fitness**; Hint: the reducer must take the maximum fitness (values) for each rule (key) not the summation.
3. **Second round:**
   1. **Map Class**: Do the crossover and mutation on the best rules and generate all new possible rules.
   2. **Combiner:** recalculates the fitness for those new rules.
   3. **Reducer**: Selection operation is conducted between the parents and the children, and the best rules chosen between the parents and the children; The better replaced the other one.
4. **These two around are repeated until the maximum number of generations is reached, In which case the Output file should contains all the best rules and their corresponding fitness values e.g. Chromosome 1: [3] 🡪 [39, 48]; Fitness = 0.66…..etc**

***Sample of the Input file:*** We will work on a real Retail Dataset.



***Example:***

1. **Take sample form retail market: Suppose the N= 4, and the minimum\_Fitness=0.25;**

|  |  |
| --- | --- |
|  | **Data Sequence** |
| **1** | **38 39 47 48** |
| **2** | **38 39 48 49 50** |
| **3** | **32 41 60 61** |
| **4** | **3 39 48** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **32** | **38** | **39** | **41** | **47** | **48** | **49** | **50** | **51** | **60** | **61** |
| **0** | **1** | **1** | **0** | **1** | **1** | **0** | **0** | **0** | **0** | **0** |
| **0** | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **0** | **0** |
| **1** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **1** |
| **0** | **0** | **1** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** |

**Table 1.** Vertical Bitmap Representation

