Programming Assignment: Frequent Itemset Mining Using Apriori

Description

In this programming assignment, you are required to implement the Apriori algorithm and apply it to mine frequent itemsets from a real-life dataset.

Input

The provided input file (`categories.txt`) consists of the category lists of 77,185 places in the US. Each line corresponds to the category list of one place, where the list consists of a number of category instances (e.g., hotels, restaurants, etc.) that are separated by semicolons.

An example line is provided below:

```
Local Services;IT Services & Computer Repair
```

In the example above, the corresponding place has two category instances: "Local Services" and "IT Services & Computer Repair".

Output

You need to implement the Apriori algorithm and use it to mine category sets that are frequent in the input data. When implementing the Apriori algorithm, you may use any programming language you like. We only need your result pattern file, not your source code file. After implementing the Apriori algorithm, please set the relative minimum support to 0.01 and run it on the 77,185 category lists. In other words, you need to extract all the category sets with absolute support larger than (non-inclusive) 771.

Part 1

Please output all the length-1 frequent categories with their absolute supports into a text file named `part1.txt`. Every line corresponds to exactly one frequent category and should be in the following format:

`support:category`

For example, suppose a category (Fast Food) has an absolute support of 3,000, then the line corresponding to this frequent category set in `part1.txt` should be:

`3000:Fast Food`

Part 2

Please write all the frequent category sets along with their absolute supports into a text file named `part2.txt`. Every line corresponds to exactly one frequent category set and should be in the following format:

`support:category_1;category_2;category_3;...`

For example, suppose a category set (Fast Food; Restaurants) has an absolute support of 2,851, then the line corresponding to this frequent category set in `part2.txt` should be:

`2851:Fast Food;Restaurants`

Important Tips

Make sure that you format each line correctly in the output file. For instance, use a semicolon instead of another character to separate the categories for each frequent category set.

In the result pattern file, the order of the categories does not matter. For example, the following two cases will be considered equivalent by the grader:

Case 1: `2851:Fast Food;Restaurants`

Case 2: `2851:Restaurants;Fast Food`