Adam Vaccaro - INF 553

Description for Homework 2

Implementation of the SON algorithm:

1. Load in text as RDD
2. Create RDD for baskets
3. Convert baskets RDD of strings to RDD of integers and create integer-to-string lookup table
4. SON map phase 1: implement the a-priori algorithm to find possible frequent itemsets
5. SON reduce phase 1: collect results from map phase 1 to get set of candidate frequent itemsets
6. SON map phase 2: count the occurence of candidate frequent itemsets in the original dataset
7. SON reduce phase 2: aggregate results of map phase 2 to find true frequent itemsets
8. Translate the integer keys back into strings using lookup table
9. Sort results
10. Print to output file

Spark version: 2.2.1

Main class name: RunSON

To run using spark-submit:
spark-submit --driver-memory 4G --class RunSON Adam_Vaccaro_SON.jar <caseNumber> <path2input.csv> <support>

For example:
spark-submit --driver-memory 4G --class RunSON Adam_Vaccaro_SON.jar 1 Data/books.csv 1200

Runtime table for Problem 2:

File Name	Case Number	Support	Runtime (sec)
beauty.csv	1	50	553
beauty.csv	2	40	31
books.csv	1	1200	439
books.csv	2	1500	92