[docs passing

Apache Flink Tutorial

Frequent items mining (FIM)

1. Introduction

Basket analysis, Apriori algorithm and opportunities for parallelisation

2. Hands-on

Data used:

- Tutorial/src/main/resources/
    - simple_data.dat: 9 transactions, 5 items
    - T10I4D100K.dat: 100K transactions, 870 items
    - generated_data.csv (retail data): 2246 lines, one item per line, one transaction represented by several lines with unique transaction id. 162 transactions, 300 items
- Tutorial/OnlineRetail/
    - OnlineRetail.csv: 541910 lines, 795 transactions, 61 items
    - OnlineRetail-short.csv: a short version of the previous set

In this tutorial you will learn to:

- Read and pre-process data (transactions data)
- Make iterations
- Broadcast data in each iteration
- Set up a convergence criterion

3. Import in your IDE the Tutorial project

- 3.1 Try the Apriori algorithm, run it with simple data

EXAMPLE 1, parameters: data=simple_data.dat, minSuppor=2, numIterations=3, numParallelisation=4

```
./src/main/resources/simple_data.dat 2 3 4
```

Expected output:

(1 2 3): 2

(1 2 5): 2

EXAMPLE 2, parameters: data=T10I4D100K.dat, minSuppor=150, numIterations=10, numParallelisation=4

```
./src/main/resources/T10I4D100K.dat 150 10 4
```

Expected output:

(75 205 207 285 403 461 529 829 896 950): 190

(8 71 75 108 242 438 486 684 766 958): 187

EXAMPLE 3 parameters: data=generated_data.csv, minSuppor=2, numIterations=5

```
./src/main/resources/generated_data.csv 2 5
```

Expected output:

[6753405175621, 915540909914, 2750765935791, 8505809512760, 9625609114634]

[1256496891938, 241140567124, 4945840519593, 6032308399817, 3122398523832]

[7709209653122, 1081941090150, 8579865195899, 2337333124395, 5341302202256]

[7361987825133, 155147135649, 3119115207602, 7992951519390, 1983031219791]

[8359975670874, 4540282343591, 4682196863917, 5035816984106, 2726698887444]

- 3.2 Try yourself to analyse the data in OnlineRetail.csv
    - start first with the small set: OnlineRetail-short.csv
    - here time stamps should be taken into account
    - the data is not that clean in OnlineRetail.csv

      EXERCISE: find out the reason for the problem, propose a solution

EXAMPLE 4, parameters: data=OnlineRetail-short, minSuppor=3, numIterations=3

```
./OnlineRetail/OnlineRetail-short.csv 3 3
```

Expected output:

[21730, 71053, 22752]

[21730, 85123A, 84029E]

[21730, 85123A, 84029G]

[85123A, 84029E, 84029G]

[21730, 84029E, 84029G]

[21730, 85123A, 22752]

Steps for processing OnlineRetail data:

- Load data
- Join items per transaction, sort by time

- Map transaction items into integers
- Mine frequent items
- Map back items