Ton Duc Thang University
Faculty of Information Technology

TIỂU LUẬN GIỮA KỲ

Môn: Xử lý dữ liệu lớn

Thời gian làm bài: 03 tuần

I. Hình thức

- Đồ án giữa kỳ được thực hiện theo nhóm **03 – 05** sinh viên.
- Nhóm sinh viên thực hiện các yêu cầu và nộp bài theo hướng dẫn chi tiết bên dưới.

II. Yêu cầu

Cho tập tin **data.csv** chứa dữ liệu mua hàng của người dùng. Trong đó, dòng đầu tiên chứa tiêu đề (header), các dòng còn lại là dữ liệu tương ứng.

- **Member_number**: mã số khách hàng
- **Date**: ngày mua hàng dạng dd/mm/yyyy
- **itemDescription**: tên của một món hàng
- **year**: năm mua
- **month**: tháng mua
- **day**: ngày mua
- **day_of_week**: thứ trong tuần

Ví dụ

Member_number	Date	itemDescription	year	month	day	day_of_week
1249	01/01/2014	citrus fruit	2014	1	1	2
1249	01/01/2014	coffee	2014	1	1	2
1381	01/01/2014	curd	2014	1	1	2
1381	01/01/2014	soda	2014	1	1	2
1440	01/01/2014	other vegetables	2014	1	1	2
1440	01/01/2014	yogurt	2014	1	1	2
1659	01/01/2014	specialty chocolate	2014	1	1	2
1659	01/01/2014	frozen vegetables	2014	1	1	2
1789	01/01/2014	hamburger meat	2014	1	1	2
1789	01/01/2014	candles	2014	1	1	2

Dữ liệu trong data.csv (hiển thị trên Google Colab)

a) Câu 1 (1.0 điểm): Giỏ hàng

- Sử dụng **DataFrame** trong thư viện PySpark để đọc tập tin **data.csv** và tìm ra *danh sách món hàng mỗi người khách mua trong một ngày (ngày/tháng/năm)*.
- Kết quả xử lý ghi xuống tệp **baskets.csv**.
- Nội dung kết quả gồm: *Mã khách hàng, Ngày mua, Danh sách món hàng*. Lưu ý các cột cách nhau bằng dấu "**;**" và các phần tử món hàng trong cột *Danh sách món hàng* cách nhau bằng dấu "**,**". Các dòng dữ liệu xếp tăng dần theo *Mã khách hàng, Ngày mua*.
- Sinh viên đảm bảo mỗi món hàng trong một giỏ hàng là duy nhất (không trùng lắp).
- Sau khi hoàn thành, hiển thị DataFrame kết quả ra màn hình để minh chứng.

Ví dụ tập tin kết quả

part-00000 ×

```
1 Member_number;Date;itemDescription
2 1249;01/01/2014;citrus fruit,coffee
3 1381;01/01/2014;curd,soda
4 1440;01/01/2014;other vegetables,yogurt
5 1659;01/01/2014;specialty chocolate,frozen vegetables
6 1789;01/01/2014;hamburger meat,candles
7 1922;01/01/2014;tropical fruit,other vegetables
8 2226;01/01/2014;sausage,bottled water
9 2237;01/01/2014;bottled water,Instant food products
10 2351;01/01/2014;cleaner,shopping bags
```

Kết quả tìm danh sách món hàng mỗi người mua trong một ngày.

Dòng đầu tiên phát sinh từ header ban đầu. Các phần tử cách nhau bởi dấu ';'

Danh sách món hàng là dạng chuỗi, ngăn cách bởi dấu ','.

Member_number	Date	itemDescription
1249	01/01/2014	citrus fruit,coffee
1381	01/01/2014	curd,soda
1440	01/01/2014	other vegetables,...
1659	01/01/2014	specialty chocola...
1789	01/01/2014	hamburger meat,ca...

DataFrame kết quả

b) Câu 2 (4.0 điểm): A-Priori

Cài đặt lớp đối tượng APriori để thực hiện thuật toán cùng tên.

- Phương thức khởi tạo: nhận vào đường dẫn đến tập tin chứa danh sách giỏ hàng tương tự như **baskets.csv** ở câu trên; hằng số s là ngưỡng support (ví dụ s = 0.3); hằng số c là ngưỡng confidence (ví dụ c = 0.5).
- Phương thức run(): chạy thuật toán. Sau đó, lưu DataFrame kết quả chứa các cặp phổ biến xuống tệp **apriori_frequent_pairs.csv** và lưu DataFrame chứa danh sách các association rules xuống **apriori_association_rules.csv**. Cấu trúc các DataFrame dựa theo nhưng không bắt buộc giống hoàn toàn ví dụ từ **FPGrowth**.
- Sinh viên có thể thêm các hàm và thuộc tính hỗ trợ cần thiết nhưng phải đảm bảo mã nguồn tinh gọn, tối ưu, không rờm rà.

c) Câu 2 (4.0 điểm): PCY

Cài đặt lớp đối tượng PCY để thực hiện thuật toán cùng tên.

- Phương thức khởi tạo: nhận vào đường dẫn đến tập tin chứa danh sách giỏ hàng tương tự như **baskets.csv** ở câu trên; hằng số s là ngưỡng support (ví dụ s = 0.3); hằng số c là ngưỡng confidence (ví dụ c = 0.5).
- Phương thức run(): chạy thuật toán. Sau đó, lưu DataFrame kết quả chứa các cặp phổ biến xuống tệp **pcy_frequent_pairs.csv** và lưu DataFrame chứa danh sách các association rules xuống **pcy_association_rules.csv**. Cấu trúc các DataFrame dựa theo nhưng không bắt buộc giống hoàn toàn ví dụ từ **FPGrowth**.

- Sinh viên có thể thêm các hàm và thuộc tính hỗ trợ cần thiết nhưng phải đảm bảo mã nguồn tinh gọn, tối ưu, không rờm rà.

d) Câu 4 (1.0 điểm): FPGrowth

- Sử dụng lớp FPGrowth để thực hiện tìm tập phổ biến và các association rules.
- So sánh kết quả của FPGrowth và kết quả hai thuật toán tự cài đặt ở trên.

III. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp **midterm_<Mã nhóm>**:
  - **source.ipynb:** chứa mã nguồn đồ án
  - **source.pdf:** kết xuất pdf mã nguồn đồ án từ Google Colab.
  - **report.pdf:** báo cáo đồ án.
    - Tóm tắt cách giải quyết các yêu cầu.
    - Danh sách thành viên, phân công công việc và mức độ hoàn thành.
- Lưu ý giữ lại kết quả thực thi của các ô trong cả hai tập tin .ipynb và .pdf.
- Nén thư mục thành tệp zip và nộp theo deadline.

IV. Quy định

- **Nộp bài trễ thì cả nhóm nhận 0.0 điểm.**
- **Mọi hành vi sao chép code trên mạng, chép bài bạn hoặc cho bạn chép bài nếu bị phát hiện đều sẽ bị điểm 0.0.**
- **Nếu bài làm của sinh viên có dấu hiệu sao chép trên mạng hoặc sao chép nhau, sinh viên sẽ được gọi lên phỏng vấn code để chứng minh bài làm là của mình.**

-- HẾT --