BÀI TẬP SPARK

Viết chương trình trên nền Spark thực hiện "Gợi ý người quen" trên mạng xã hội. Ý tưởng chính là: nếu hai người có nhiều bạn chung thì hệ thống gợi ý họ nên làm bạn với nhau (ngoài đời thực rất có thể họ là bạn của nhau nhưng chưa kết bạn trên mạng xã hội).

Dữ liệu:

- File soc-LiveJournal1Adj.txt trong thư mục q1/data.
- File có nhiều dòng mỗi dòng theo dạng: **<User><TAB><Friends>**

Trong đó <User> là mã user dạng số nguyên, mỗi user có 1 mã duy nhất; <Friends> là danh sách các mã không trùng nhau của các user là bạn với mã user bên trái ký tự TAB. Lưu ý quan hệ bạn bè là quan hệ hai chiều, tức là nếu A là bạn của B thì B cũng là bạn của A. Dữ liệu trên được cung cấp với qui ước ngầm định này.

Thuật toán: Sử dụng thuật toán đơn giản sau: Với mỗi user U thuật toán sẽ kiến nghị N=10 người không là bạn của U nhưng có số lượng bạn chung với U nhiều nhất.

Output:

- Kết quả xuất ra mỗi dòng cho một user theo dạng: `<User><TAB><Recommendations>` Trong đó <User> là mã user dạng số nguyên, <Recommendations> là danh sách các mã user không trùng lặp tương ứng với kết quả kiến nghị của thuật toán được sắp xếp giảm dần theo số bạn chung với user bên trái ký tự TAB. Nếu có nhiều user cùng số lượng bạn chung thì sắp xếp tăng dần theo mã user.
- Trong trường hợp kết quả có ít hơn 10 người thì xuất hết theo thứ tự giảm dần số bạn chung. Nếu user không có bạn thì xuất ra danh sách rỗng. Nếu danh sauser

Phác thảo cách làm: Hãy mô tả ngắn ngọn cách bạn dùng Spark dể giải quyết bài toán này. Bản mô tả trong khoảng 3 đến 4 câu mô tả một cách khái quát nhất chiến lược của bạn.

Gợi ý:

- Dùng Google Colab giúp chạy trên dữ liệu lớn mà không phải lo về máy của mình. Có thể copy colab tại đây sau đó dùng các phần cài đặt Spark cho bài toán này.
- Trước khi bấm nút chạy ứng dụng trên Spark bạn nên chạy thử trước một phần nhỏ dữ liệu để đảm bảo rằng code mình viết là đúng, tránh trường hợp chờ quá lâu để thấy kết quả… sai. Lệnh Command .take(X) sẽ rất có ích nếu bạn muốn lấy X phần tử đầu trong the RDD.
- Để kiểm tra thuật toán của bạn đúng bạn có thể so sánh kết quả của bạn với danh sách gợi ý cho user ID 11 là (top 10): 27552, 7785, 27573, 27574, 27589, 27590, 27600, 27617, 27620, 27667.

Tài liệu cần nộp

(1) File colab cuối cùng.
(2) Bản mô tả tổng quát dạng word (docx).
(3) Danh sách gợi ý cho các mã user sau (ghi trong file docx trên): 924, 8941, 8942, 9019, 9020, 9021, 9022, 9990, 9992, 9993.