LẬP TRÌNH CƠ BẢN VỚI JAVA

Bài tập tuần 14

Bài 1. Bạn có một blog trong đó có một số bài viết hay. Với hy vọng kiếm được lợi nhuận, bạn thêm vào trên đầu mỗi bài viết các đoạn quảng cáo. Sau một thời gian, một số bài viết đã giúp bạn thu được đáng kể lợi nhuận, nhưng cũng có một số bài viết thì không thu được lợi nhuận nào. Giả sử việc quyết định một bài viết có khả năng giúp thu lợi nhận hay không phụ thuộc vào số bức hình và số đoạn văn trong bài. Vậy với một bài viết mới, bạn rất muốn biết liệu bài viết này có thể giúp bạn thu được lợi nhuận hay không?

Cho dữ liệu về n bài viết được lưu trong tệp *ads.txt* gồm thông tin về số bức ảnh, số đoạn văn có trong bài viết và nhãn của bài viết. Trong đó bài viết được gán nhãn "Y" là bài viết cho lợi nhuận, và gán nhãn "N" là bài viết không sinh lợi nhuận. Bảng (1) minh hoạ cấu trúc của tệp *ads.txt*.

Để biết được một bài viết mới có thể sinh lợi nhuận hay không, cần:

1. Tính khoảng cách (Euclid) giữa bài viết mới và n bài viết
2. Sắp xếp khoảng cách theo thứ tự giảm dần
3. Xét tập gồm k bài viết có khoảng cách gần nhất đến bài viết mới

số bức ảnh	số đoạn văn	nhãn
10	2	Y
12	3	Y
9	2	Y
0	10	N
1	9	N
3	11	N

Bảng 1: Ví dụ về dữ liệu trong tệp *ads.txt*

4. Nhãn của bài viết mới sẽ là nhãn xuất hiện nhiều nhất trong tập k bài viết xác định ở bước 3.

Yêu cầu:

- Tạo lớp **Classifier** chứa các phương thức cho phép xác định nhãn của một bài viết
- Viết lớp **ClassifierTest** để lấy thông tin về các bài viết đã được gán nhãn (lưu trong tệp *ads.txt*) và các bài viết mới chưa được gán nhãn (lưu trong tệp *newarticles*), thực hiện gán nhãn cho những bài viết mới và lưu kết quả gán nhãn vào tệp *labelarticles* (có định dạng giống tệp *ads.txt*).

Chú ý: Dữ liệu về số hàng xóm gần nhất (giá trị k) được nhập là đối dòng lệnh.