



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO

BÀI TẬP THỰC HÀNH TUẦN 8

Khai thác dữ liệu Web



Giảng viên hướng dẫn : Khoa Phó Ngọc Đăng

Sinh viên thực hiện : Vũ Mạnh Hùng

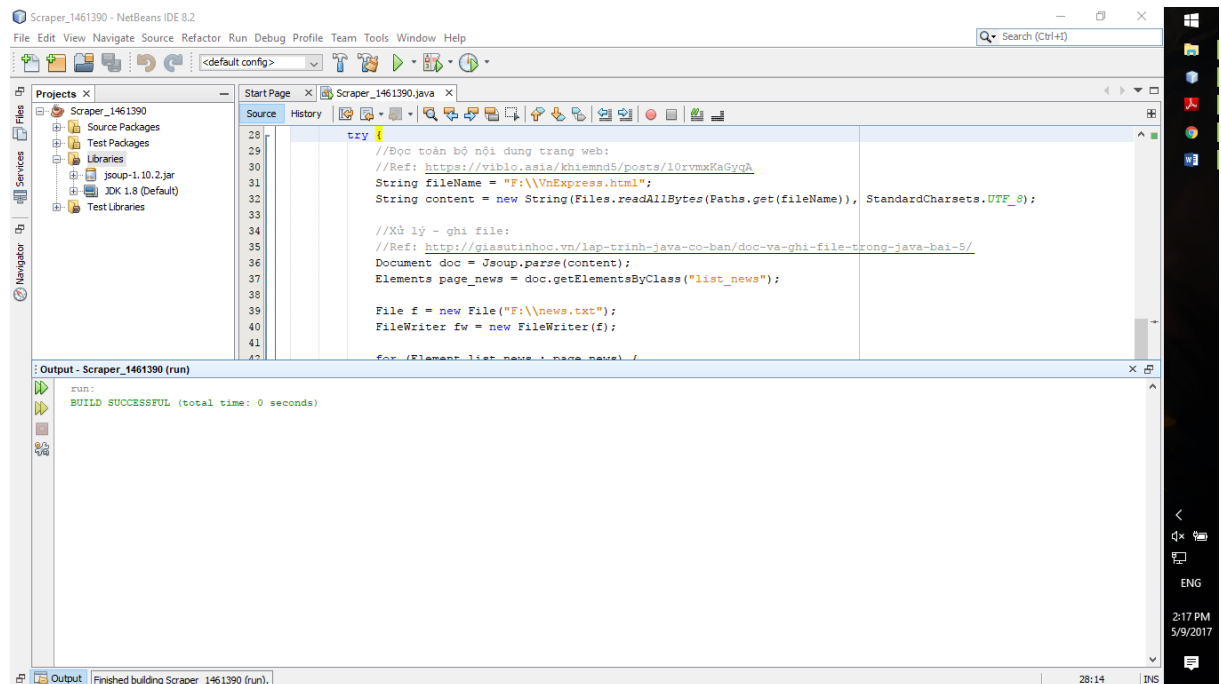
Mã số sinh viên : 1461390

Lớp : 15CK3

Ca : 1 – C6

TP HCM, tháng 5 năm 2017

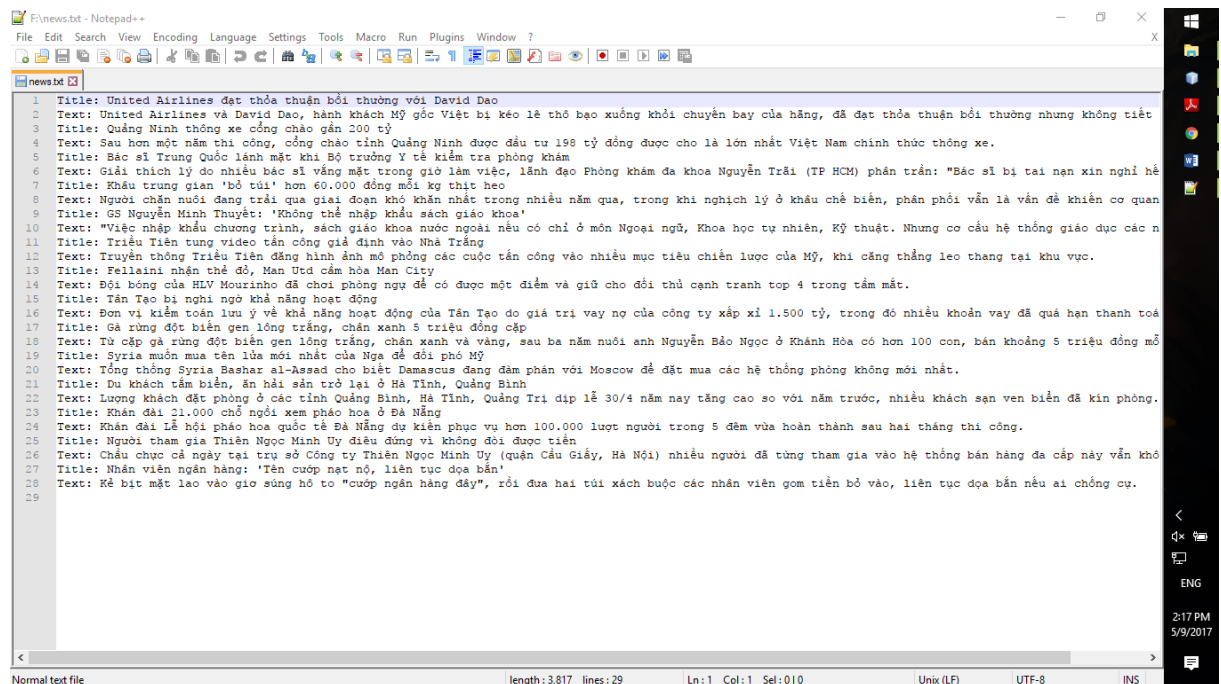
Hoàn thành bài tập!!!!



```
28      try {
29          //Đọc toàn bộ nội dung trang web:
30          //Ref: https://viblo.asia/khiemnd5/posts/10rvmxKaGyqA
31          String fileName = "F:\\VnExpress.html";
32          String content = new String(Files.readAllBytes(Paths.get(fileName)), StandardCharsets.UTF_8);
33
34          //Xử lý - ghi file:
35          //Ref: http://giasutinhoc.vn/lap-trinh-java-co-ban/doc-va-ghi-file-trong-java-bai-5/
36          Document doc = Jsoup.parse(content);
37          Elements page_news = doc.getElementsByClass("list_news");
38
39          File f = new File("F:\\news.txt");
40          FileWriter fw = new FileWriter(f);
41
42          for (Element list_news : page_news) {
```

Output - Scraper_1461390 (run)

```
run
BUILD SUCCESSFUL (total time: 0 seconds)
```



```
1 Title: United Airlines đạt thỏa thuận bồi thường với David Dao
2 Text: United Airlines và David Dao, hành khách Mỹ gốc Việt bị kéo lê thô bạo xuống khỏi chuyến bay của hãng, đã đạt thỏa thuận bồi thường nhưng không tiết
3 Title: Quảng Ninh thông xe công cảng gần 200 tỷ
4 Text: Sau hơn một năm thi công, công cảng tỉnh Quảng Ninh được đầu tư 198 tỷ đồng được cho là lớn nhất Việt Nam chính thức thông xe.
5 Title: Bác sĩ Trung Quốc lãnh mất khi Bộ trưởng Y tế kiểm tra phòng khám
6 Text: Giải thích lý do nhiều bác sĩ vắng mặt trong giờ làm việc, lãnh đạo Phòng khám đa khoa Nguyễn Trãi (TP HCM) phản trả: "Bác sĩ bị tai nạn xin nghỉ hê
7 Title: Khẩu trung gian 'bò tui' hơn 60.000 đồng mỗi kg thất bại
8 Text: Người chăn nuôi đang trải qua giai đoạn khó khăn nhất trong nhiều năm qua, trong khi nghịch lý ở khâu chế biến, phân phối vẫn là vấn đề khiến cơ quan
9 Title: GS Nguyễn Minh Thuyết: 'Không thể nhập khẩu sách giáo khoa'
10 Text: "Việc nhập khẩu chương trình, sách giáo khoa nước ngoài nếu chỉ ở môn Ngoại ngữ, Khoa học tự nhiên, Kỹ thuật. Nhưng cơ cấu hệ thống giáo dục các n
11 Title: Triều Tiên tung video tấn công giả định vào Nhà Trắng
12 Text: Truyền thông Triều Tiên đăng hình ảnh mô phỏng các cuộc tấn công vào nhiều mục tiêu chiến lược của Mỹ, khi căng thẳng leo thang tại khu vực.
13 Title: Fellaïni nhận thẻ đỏ, Man Utd cấm hòa Man City
14 Text: Đội bóng của HLV Mourinho đã chơi phòng ngự để có được một điểm và giữ cho đối thủ cạnh tranh top 4 trong tầm mắt.
15 Title: Tân Tào bị nghi ngờ khả năng hoạt động
16 Text: Đơn vị kiểm toán lưu ý về khả năng hoạt động của Tân Tào do giá trị vay nợ của công ty xấp xỉ 1.500 tỷ, trong đó nhiều khoản vay đã quá hạn thanh toá
17 Title: Gà rừng đột biến gen lông trắng, chân xanh 5 triệu đồng cặp
18 Text: Từ cặp gà rừng đột biến gen lông trắng, chân xanh và vàng, sau ba năm nuôi anh Nguyễn Bảo Ngọc ở Khánh Hòa có hơn 100 con, bán khoảng 5 triệu đồng mỗ
19 Title: Syria muốn mua tên lửa mới nhất của Nga để đối phó Mỹ
20 Text: Tổng thống Syria Bashar al-Assad cho biết Damascus đang đàm phán với Moscow để đặt mua các hệ thống phòng không mới nhất.
21 Title: Du khách tắm biển, ăn hải sản trở lại ở Hà Tĩnh, Quảng Bình
22 Text: Lượng khách đặt phòng ở các tỉnh Quảng Bình, Hà Tĩnh, Quảng Trị dịp lễ 30/4 năm nay tăng cao so với năm trước, nhiều khách sạn ven biển đã kín phòng.
23 Title: Khán đài 21.000 chỗ ngồi xem pháo hoa ở Đà Nẵng
24 Text: Khán đài Lễ hội pháo hoa quốc tế Đà Nẵng dự kiến phục vụ hơn 100.000 lượt người trong 5 đêm và hoàn thành sau hai tháng thi công.
25 Title: Người tham gia Thiên Ngọc Minh Uy điều động vì không đối được tiền
26 Text: Châu chực cả ngày tại trụ sở Công ty Thiên Ngọc Minh Uy (quận Cầu Giấy, Hà Nội) nhiều người đã từng tham gia vào hệ thống bán hàng đa cấp này vẫn khó
27 Title: Nhân viên ngân hàng: 'Tên cướp nạt nộ, liên tục dọa bắn'
28 Text: Kẻ bit mặt lao vào gào sùng hô to "cướp ngân hàng đây", rồi đưa hai túi xách buộc các nhân viên gom tiền bỏ vào, liên tục dọa bắn nếu ai chống cự.
29
```