CẤU TRÚC DỮ LIỆU VÀ GIẢI THUẬT I
Bài tập lớn 1
Assignment 1

1. Yêu cầu:

Xây dựng chương trình tách từ tố cho file văn bản tiếng Việt và nhận dạng một số loại từ tố thông dụng (số nguyên, số thực, số điện thoại, địa chỉ website, ngày tháng).

Input: đầu vào là 1 file văn bản tiếng Việt.

Ouput: đầu ra là văn file văn bản tương ứng đã được từ tố hóa (tokenized) và các từ tố đặc biệt được gán nhãn.

2. Mô tả:

a) Mục tiêu của bài tập là giúp sinh viên thành thạo kĩ năng về làm việc với file văn bản và xâu kí tự, với biểu thức chính qui (regular expression) trong Java. Đây cũng là bài toán thực tế trong các ứng dụng liên quan đến xử lý văn bản.

b) Ví dụ:

Đầu vào là 1 file:

dsa_input_assign1.txt

Theo thống kê đến ngày 6/8, bão số 1 đã làm 7 người chết và mất tích, 63 người bị thương; làm 2.989 nhà bị đổ sập hoàn toàn, trên 82.650 nhà bị tốc mái, hư hỏng, 511 nhà bị ngập nước; 1.316 tàu thuyền bị chìm, hư hỏng tại khu vực cửa sông; 216.194 ha lúa bị ngập, 28.372 ha rau màu bị hư hại… tổng thiệt hại ước tính trên 6.442 tỷ đồng. Nguồn: http://baochinhphu.vn/Tin-noi-bat/Rut-kinh-nghiem-tu-bao-so-1-so-2/284860.vgp Số điện thoại nóng phụ trách bão lũ: 0901234567

Kết quả trả về sẽ là:

dsa_input_assign1.out

Theo thống kê đến ngày 6/8[DATE] , bão số 1[NUM] đã làm 7[NUM] người chết và mất tích , 63[NUM] người bị thương ; làm 2.989[NUM] nhà bị đổ sập hoàn toàn , trên 82.650[NUM] nhà bị tốc mái , hư hỏng , 511[NUM] nhà bị ngập nước ; 1.316[NUM] tàu thuyền bị chìm , hư hỏng tại khu vực cửa sông ; 216.194[NUM] ha lúa bị ngập , 28.372[NUM] ha rau màu bị hư hại … tổng thiệt hại ước tính trên 6.442[NUM] tỷ đồng . Nguồn : http://baochinhphu.vn/Tin-noi-bat/Rut-kinh-nghiem-tu-bao-so-1-so-2/284860.vgp[URL] Số điện thoại nóng phụ trách bão lũ : 0901234567[PHONE]

3. Hướng dẫn

a)

Tách từ tố là gì:

Tách từ tố (tokenize) tức là tách các từ tố trong một văn bản và cho cách nhau bởi dấu cách. Nhiệm vụ này cần phải thực hiện vì trong cách viết bình thường thì các từ tố và các kí hiệu bị dính với nhau.

Nhận dạng một số từ tố có format đặc biệt:

- số (bao gồm số nguyên và số thực)

- số điện thoại
- ngày tháng
- địa chỉ website (url)

b) Kĩ thuật để tách từ tố và nhận dạng từ tố là sử dụng biểu thức chính qui. Tham khảo các tài liệu sau:
http://www.tutorialspoint.com/java/java_regular_expressions.htm
(tiếng Việt: http://vietjack.com/java/regular_expression_trong_java.jsp)
https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html
http://beginnersbook.com/2014/08/java-regex-tutorial/

Các trao đổi hỏi đáp liên quan đến bài tập thực hiện trên forum của môn học trên sakai.

4. Qui định:
Mọi hình thức sao chép (copy của bạn và của các tool trên mạng) mã sẽ bị 0 điểm. Sinh viên cho bạn copy cũng sẽ bị 0 điểm.

5. Đánh giá:
Giảng viên sẽ đánh giá dựa trên:
- Test case: đưa file input và kiểm tra kết quả trên file ouput.
- Kiểm tra code.

Nộp bài theo thời hạn như sau:

Thời hạn nộp	**Mô tả**	**Trường hợp không nộp đúng hạn**
Lần 1: 12:00 pm (tuần 7) 29/9/2016	Chạy được input, output đúng khuôn dạng, kết quả có thể chưa hoàn toàn chính xác	Trừ 2 điểm
Lần 2: 12:00 pm (tuần 9) 20/10/2016	Gần như hoàn chỉnh, chạy qua các test phức tạp.	0 điểm toàn bài
Lần 3: 12:00 pm (tuần 11) 3/11/2016	Tối ưu lại code, sửa lỗi.	0 điểm toàn bài.