NATURAL LANGUAGE PROCESSING

Lecturer: Doctor Bui Thanh Hung
Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city
Email: hung.buithanhcs@gmail.com (buithanhhung@iuh.edu.vn)
Website: https://sites.google.com/site/hungthanhbui1980/

Bài 1:
Sử dụng CoreNLP ở đường link https://stanfordnlp.github.io/CoreNLP/
thực hiện các yêu cầu sau:
Vào một file văn bản tiếng Anh với nội dung ở đường link sau:
https://www.nytimes.com/2021/06/29/science/indigenous-data-microbiome-science.html

1- Ra gán nhãn cho từng từ (PARTS OF SPEECH) trong văn bản đó, lưu kết quả ra 1 file bao gồm cả thống kê các nhãn và số lượng của từng loại ở phần đầu file.
2- Ra các tên riêng (NAMED ENTITIES) trong văn bản đó, lưu kết quả ra 1 file với mỗi dòng trong file theo thứ tự sau: STT (câu), câu, danh sách tên riêng.
3- Kết quả của DEPENDENCY PARSES trong 1 file và đề xuất sử dụng kết quả này trong 1 bài toán cụ thể
4- Kết quả của COREFERENCE trong 1 file và đề xuất sử dụng kết quả này trong 1 bài toán cụ thể

Bài 2:
Sử dụng vnCoreNLP ở đường link https://github.com/vncorenlp/VnCoreNLP
Sử dụng file văn bản tiếng Việt với nội dung ở đường link sau:
https://tuoitre.vn/hanh-trinh-vuon-den-do-thi-thong-minh-20190713235201371.htm
thực hiện các yêu cầu:

1. Tách từ lưu kết quả ra 1 file bao gồm cả thống kê các từ và số lượng của từng từ ở phần đầu file.
2. Ra gán nhãn cho từng từ (PARTS OF SPEECH) trong văn bản đó, lưu kết quả ra 1 file bao gồm cả thống kê các nhãn và số lượng của từng loại ở phần đầu file.
3. Ra các tên riêng (NAMED ENTITIES) trong văn bản đó, lưu kết quả ra 1 file với mỗi dòng trong file theo thứ tự sau: STT (câu), câu, danh sách tên riêng.

4. Kết quả của DEPENDENCY PARSES trong 1 file và cho biết các mẫu DEPENDENCY PARSES phổ biến nhất (top 2) trong tất cả các câu.

Bài 3:
Cho văn bản là tiểu thuyết Dracula của Bram Stoker ở đường link sau:
https://www.gutenberg.org/files/345/345-h/345-h.htm
Hãy thực hiện các yêu cầu sau:

1. Tìm 5 từ đơn (unigram) có tần suất xuất hiện từ 1000 trở lên.
2. Tìm 3 từ ghép (word pairs) có tần suất từ 200 trở lên.
3. Tìm 1 từ ghép của 3 từ (trigram) có tần suất từ 10 trở lên.
4. Cho một từ (từ đơn- unigram) ký hiệu là w2 và ba từ khác nhau (từ đơn-unigram) ký hiệu là w1, hãy tính xác suất có điều kiện P(w2|từ trước đó là w1). Chọn các từ sao cho xác suất của ít nhất hai cặp là lớn hơn 0.
5. Chọn một từ tùy ý từ dữ liệu. Hãy cho biết xác suất để từ đó là "godalming" là bao nhiêu? Và xác suất để từ đó là "godalming" khi từ trước đó là "lord" là bao nhiêu?