# BÁO CÁO MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN – CS221.M11

## Bài tập nhóm số 1

**Bùi Trần Ngọc Dũng – 19521385**

**Nguyễn Đăng Minh – 19520164**

1. **Thu thập dữ liệu**

   Nhóm chúng em tiến hành thu thập 56 câu tiêu đề của trên trang báo điện tử https://vnexpress.net/ chuyên mục đời sống – pháp luật. Sau đó, tiến hành chuẩn hóa các tiêu đề bằng cách loại bỏ các ký tự, dấu câu không cần thiết và các chữ in hoa (trừ địa danh và tên riêng).

   Nhóm cũng tìm kiếm và kết hợp xây dựng được một bộ từ điển Tiếng Việt khá lớn với 91970 từ vựng khác nhau (một hay nhiều âm tiết) bao gồm các địa danh và tên riêng phổ biến.

2. **Xác định từ ghép và thống kê tần suất**

   Nhóm thực hiện tách từ ghép và đếm tần suất của từng từ bằng phương pháp thủ công, sau đó sẽ sử dụng phương pháp tách từ tự động để kiếm chứng.

   File xử lý thủ công :

   + **Tu_ghep.txt** chứa các từ ghép được tách

   + **Tansuat.csv** chứa tất cả các từ và số lần xuất hiện của chúng trong toàn bộ ngữ liệu

3. **Phương pháp tách từ Maximum Matching**

   **Ý tưởng:** Phương pháp này được gọi là so khớp tối đa từ trái sang phải (hoặc ngược lại). Nó sẽ duyệt một câu từ trái sang phải (hoặc ngược lại) và chọn ra từ ghép có độ dài được định nghĩa lớn nhất có mặt trong một từ điển từ vựng được cho sẵn. Quá trình này được lặp đi lặp lại cho đến khi độ dài giảm dần cho đến hết câu

   **Ưu điểm:**

   - Thuật toán đơn giản và dễ hiểu

- Tuy nó sử dụng chiến lược vét cạn nhưng trong thực tế thuật toán này chạy rất nhanh.
- Phù hợp với bộ dữ liệu nhỏ của nhóm

**Nhược điểm:**

- Nếu các từ không có trong từ điển thì chắc chắn thuật toán sẽ thất bại, không giải quyết được các trường hợp nhập nhằn: có dấu câu, in hoa thường lẫn lộn,…
- Ngoài ra, cách vận hành của thuật toán từ trái sang phải hay phải sang trái cũng có thể cho ra kết quả không nhất quán

## 4. Source code

- Nhóm cung cấp file ***.zip** với đầy đủ thông tin như sau:

  + Gọi **main.py** để chạy toàn bộ mã nguồn

  + **VNDictionary.txt, diadanh.txt** và **ten.txt** chứa các từ vựng trong từ điển. Thực hiện chỉnh sửa bằng cách thêm từ mới vào đúng format. Sau đó, gọi **CreateDic.py** để cập nhật lại file **VietNamDictionary.pkl**

  + **MaximumMatchingSegmentation.py** chứa các hàm xử lý chính

  + **raw_sentences.txt** là ngữ liệu nhóm đã thu thập

  + **result.txt** là kết quả sau khi chạy thuật toán

- Cách chạy:

**python main.py --path_data [your_file_txt] --maxlen [longest lengh word]**

## 5. Reference

**https://github.com/roy-a/Roy_VnTokenizer**

**http://tailieuso.udn.vn/bitstream/TTHL_125/6952/1/VilavongSouksan.TT.pdf**

**https://github.com/UITTrinhQuangTruong/CS221.L11**