**MÔ TẢ DỮ LIỆU TÁCH TỪ**

1. **Dữ liệu training (huấn luyện)**

Dữ liệu này được lưu trong 2 file nén:

Trainset-Segmentation-1.zip: file chứa khoảng 70,000 câu đã tách từ

Trainset-Segmentation-2.zip: file chứa khoảng 7,000 câu đã tách từ

Chú ý:

* Các âm tiết của từ ghép được nối bằng dấu gạch dưới ‘\_’.
* Trong các file dữ liệu train có thể vẫn còn các thẻ HTML, khi đọc vào thì bỏ qua.
* Kèm với 2 file trên là 2 file nén (tên file có từ raw) chứa dữ liệu đã tách câu và tách từ tố (âm tiết, số, dấu câu, …).

1. **Dữ liệu test (đánh giá): sẽ gửi vào 24/10**

Dữ liệu này được lưu trong file nén:

Testset-Segmentation-raw.zip: chứa khoảng 3,000 câu

Chú ý: Dữ liệu này đã được tách câu, tách từ tố (âm tiết, số, dấu câu, …).

1. **Yêu cầu:**

Sau khi nhận được dữ liệu test, bạn hãy chạy chương trình của mình và submit kết quả (file nén) vào 28/10 theo địa chỉ email:

Nguyễn Phương Thái: [thainp@vnu.edu.vn](mailto:thainp@vnu.edu.vn)

Vũ Xuân Lương: [vuxuanluong@gmail.com](mailto:vuxuanluong@gmail.com);

Nguyễn Thị Minh Huyền: [ntmhuyen@gmail.com](mailto:ntmhuyen@gmail.com)

Chú ý:

* Kết quả tách từ được lưu trong các file cùng tên với file test nhưng đuôi là seg
* Các âm tiết của từ ghép được nối bằng dấu gạch dưới ‘\_’ giống trong dữ liệu huấn luyện
* Số lượng câu, trật tự các câu trong file kết quả hoàn toàn giống như trong file test

**Acknowledgement**

Dữ liệu này chỉ được dùng cho mục đích nghiên cứu phát triển. Người dùng không được tự ý phân phối lại. Nguồn gốc dữ liệu này từ:

* Đề tài VLSP: <http://vlsp.vietlp.org:8080/demo/>
* Bổ sung thêm 12,000 câu được tách từ, gán nhãn từ loại do: Vietlex, Trường ĐH Công nghệ, Trường ĐH Khoa học Tự nhiên phối hợp xây dựng. Đề tài KC.01.20/11-15 tài trợ một phần.