Np	Proper noun
Nc	Classifier noun
Nu	Unit noun
N	Noun
Ny	Abbreviated noun
Nb	(Foreign) borrowed noun
V	Verb
Vb	(Foreign) borrowed verb
A	Adjective
P	Pronoun
R	Adverb
L	Determiner
M	Numeral/Quantity
E	Preposition
C	Subordinating conjunction
Cc	Coordinating conjunction
I	Interjection/Exclamation
T	Particle/Auxiliary, modal words
Y	Abbreviation
Z	Bound morpheme
X	Un-definition/Other
CH	Punctuation and symbols

MÔ TẢ DỮ LIỆU GÁN NHÃN TỪ LOẠI

1. Dữ liệu training (huấn luyện)

Dữ liệu này được lưu trong 2 file nén:

Trainset-POS-1.zip: file chứa khoảng 20,000 câu đã gán nhãn từ loại

Trainset-POS-2.zip: file chứa khoảng 7,000 câu đã gán nhãn từ loại

Chú ý:

- Các âm tiết của từ ghép được nối bằng dấu gạch dưới '_'.
- Từ được phân tách với nhãn bằng dấu gạch chéo '/'
- Kèm với 2 file trên là 2 file nén (tên file có từ raw) chứa dữ liệu đã bỏ đi nhãn từ loại

Tập nhãn:

STT	Nhãn	Tên	Ví dụ
1.	N	Danh từ	tiếng, nước, thủ đô, nhân dân, đồ đạc, cây cối, chim muông
2.	Np	Danh từ riêng	Nguyễn Du, Việt Nam, Hải Phòng, Trường Đại học Bách khoa Hà Nội, Mộc tinh, Hoả tinh, Phật, Đạo Phật
3.	Nc	Danh từ chỉ loại	con, cái, đứa, bức
4.	Nu	Danh từ đơn vị	mét, cân, giờ, nắm, nhúm, hào, xu, đồng
5.	Ni	Danh từ ký hiệu	A1, A4, 60A, 60B, 20a, 20b, ABC, ABCD
6.	V	Động từ	ngủ, ngồi, cười; đọc, viết, đá, đặt; thích, yêu, ghét, giống, muốn
7.	A	Tính từ	tốt, xấu, đẹp; cao, thấp, rộng
8.	P	Đại từ	tôi, chúng tôi, hắn, nó, y, đại nhân, đại ca, huynh, đệ
9.	L	Định từ	mỗi, từng, mọi, cái; các, những, mấy
10.	M	Số từ	một, mười, mười ba; dăm, vài, mươi; nửa, rưỡi
11.	R	Phó từ	đã, sẽ, đang, vừa, mới, từng, xong, rồi; rất, hơi, khí, quá
12.	E	Giới từ	trên, dưới, trong, ngoài; của, trừ, ngoài, khỏi, ở
13.	C	Liên từ (thường là chính phụ)	vì vậy, tuy nhiên, ngược lại
14.	Cc	Liên từ đẳng lập	và, hoặc, với, cùng
15.	I	Thán từ	ôi, chao, a ha
16.	T	Trợ từ, tình thái từ (tiểu từ)	à, a, á, ạ, ấy, chắc, chăng, cho, chứ
17.	B	Từ tiếng nước ngoài (hay từ vay mượn). Khi gán nhãn ngữ liệu, nhãn từ tiếng nước ngoài thường là nhãn kép. Chẳng hạn nếu từ là chat thì	Internet, email, video, chat

		nhãn của nó là Vb, video thì nhãn là Nb. Qua thống kê trong kho ngữ liệu thấy có: Ab Cb Eb Mb Nb Pb Vb.	
18.	Y	Từ viết tắt. Khi gán nhãn ngữ liệu, nhãn từ viết tắt thường là nhãn kép. Chẳng hạn nếu từ viết tắt là HIV thì nhãn của nó là Ny vì HIV viết đầy đủ thì là cụm danh từ. Qua thống kê trong kho ngữ liệu thấy có: Ny, Vy, Xy.	OPEC, WTO, HIV
19.	X	Các từ không phân loại được	
20.	Z	Yếu tố cấu tạo từ	bất, vô, phi
21.	CH	Nhãn dành cho các loại dấu (nhiều nhất là dấu câu) và một số ký hiệu khác	. ! ? , ; :

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

Dữ liệu này được lưu trong file nén:

Testset-POS-raw.zip: chứa khoảng 2,000 câu

Chú ý: Dữ liệu này đã được tách câu, tách từ

3. **Yêu cầu:**

Sau khi nhận được ngữ 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

Vũ Xuân Lương: vuxuanluong@gmail.com;

Nguyễn Thị Minh Huyền: ntmhuyen@gmail.com

Chú ý:

- Kết quả gán nhãn từ loại được lưu trong các file cùng tên với file test nhưng đuôi là pos
- Từ được phân tách với nhãn bằng dấu gạch chéo '/' 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.