Hướng Dẫn Cài Đặt
Hệ Thống Tách Từ Không Giám Sát
Dựa Vào Mô Hình Markov Ẩn

Hướng dẫn này được viết cho hệ điều hành ubuntu 16.04

1. Cài đặt các thư viện cần thiết

Chương trình được viết bằng ngôn ngữ lập trình python, phiên bản 2.7.12 (Chú ý rằng nếu sử dụng python 3 có thể gây lỗi do không tương thích phiên bản).

Cài đặt các thư viện cần thiết bằng các câu lệnh sau:

- Cài đặt pip

```
sudo apt-get -y install python-pip
pip install --upgrade pip
```

- Cài đặt cơ sở dữ liệu mongodb:
  - Cài đặt mongodb trên ubuntu, tham khảo link sau: https://docs.mongodb.com/v3.0/tutorial/install-mongodb-on-ubuntu/
  - Cài đặt thư viện cho python:

```
pip install pymongo
```

- Cài đặt thư viện nltk:

```
pip install numpy scipy matplotlib ipython jupyter
pandas sympy nose
pip install nltk
```

(sử dụng bản 3.3 được công bố tháng 5 năm 2018. Nếu sử dụng phiên bản cũ hơn có thể xảy ra lỗi)

- Cài đặt web frameword flask:

```
pip install Flask
```

(Có thể tham khảo thêm tại http://flask.pocoo.org/docs/1.0/installation/)

2. Khởi động hệ thống

Bước 1: Di chuyển vào thư mục chứa mã nguồn bằng lệnh cd:

```
cd path_to_project/graduation-project/word_recognition
```

(Chú ý không được thay đổi tên thư mục cũng như cấu trúc file. Nếu thay đổi có thể dẫn đến lỗi)

Bước 2: Khởi chạy server:

```
python ui/ui.py
```

(Thời gian khởi động server sẽ hoi lâu vì hệ thống load các file cần thiết)

Sau khi đã khởi động server, truy cập vào đường dẫn localhost:5000 để thử nghiệm chương trình.

3. Huấn luyện hệ thống

Dữ liệu hiện tại đã được huấn luyện từ trước, nếu muốn thử quá trình huấn luyện tiến hành theo các bước sau:

Bước 1: Huấn luyện mô hình Markov ẩn:

```
python hmm/hmm_written_by_me.py
```

Bước 2: Thực hiện việc thống kê tần số xuất hiện để phục vụ cho việc tính PMI trong quá trình decode:

```
python preprocessing.py
```

Hệ thống sẽ thực hiện việc huấn luyện trên tập dữ liệu của wikipedia tiếng Việt

Sau khi đã huấn luyện hệ thống, nếu muốn thực hiện thí nghiệm đánh giá hệ thống, chạy lệnh:

```
python hmm/evaluate_hmm_by_me.py
```

Hệ thống sẽ được đánh giá trên toàn bộ tập VLSP