

Tài liệu hướng dẫn

Cài đặt/biên dịch/chạy chương trình

Hà Nội, 2020

Mục lục

Mục lục	1
Hướng dẫn tải và cài đặt các gói cần thiết	2
Đối với Ubuntu	2
Cài đặt Python	2
Cài đặt Pip3 cho Python	2
Cài đặt NumPy	2
Cài đặt Pandas	3
Cài đặt Matplotlib	3
Cài đặt NLTK	3
Cài đặt SciPy	4
Cài đặt Sklearn	4
Đối với windows	4
Cài đặt Python	4
Cài đặt NumPy, Pandas, Matplotlib, SciPy	4
Cài đặt NLTK	4
Cài đặt Sklearn	5
Cách thực thi chương trình	5
Đối với Ubuntu	6
Đối với Windows	6

Hướng dẫn tải và cài đặt các gói cần thiết

- Các phần mềm và gói sử dụng trong chương trình:
 - Python version $\geq 3.6.9$
 - Pip3
 - NumPy
 - Pandas
 - Matplotlib
 - NLTK
 - SciPy
 - Sklearn

Đối với Ubuntu

Cài đặt Python

Đối với Ubuntu, Python được tích hợp sẵn. Để kiểm tra phiên bản Python hiện tại, ta mở terminal bằng cách ấn Ctrl + Alt + T rồi gõ

```
python3 -V
```

Nếu phiên bản hiện tại thấp hơn 3.6.9 thì ta chạy câu lệnh

```
sudo apt-get install python3.8
```

Cài đặt Pip3 cho Python

Trường hợp python của bạn chưa tích hợp sẵn pip thì bạn có thể cài đặt pip bằng lệnh

```
sudo apt install python3-pip
```

Cài đặt NumPy

Ta cài numpy thông qua câu lệnh

```
sudo pip3 install numpy
```

Cài đặt Pandas

Pandas là thư viện mã nguồn mở với hiệu năng cao cho phân tích dữ liệu trong Python được phát triển bởi Wes McKinney trong năm 2008. Ta có thể cài nó thông qua câu lệnh

```
sudo pip3 install pandas
```

Cài đặt Matplotlib

Thư viện Matplotlib giúp ta trực quan hóa dữ liệu.

```
sudo pip3 install matplotlib
```

Cài đặt NLTK

NLTK là một thư viện Python phổ biến được sử dụng cho Natural Language Processing.

```
sudo pip3 install nltk
```

Khi đã cài đặt NLTK, ta cài đặt các gói NLTK bằng cách chạy lần lượt các mã sau

```
python3
import nltk
nltk.download()
d
averaged_perceptron_tagger
d
stopwords
d
treebank
d
wordnet
```

Cài đặt SciPy

```
sudo pip3 install scipy
```

Cài đặt Sklearn

Scikit-learn là một trong những thư viện mã nguồn mở Machine Learning viết bằng Python và được đông đảo mọi người sử dụng nhất hiện nay.

```
sudo pip3 install scikit-learn
```

Đối với windows

Cài đặt Python

Đầu tiên ta tải bộ cài python mới nhất cho windows [tại đây](#).

Chọn thêm Add Python 3.8 to PATH rồi ấn Install Now

Cài đặt NumPy, Pandas, Matplotlib, SciPy

Mở cmd bằng cách ấn phím windows, gõ cmd rồi ấn enter, ta cài các gói trên qua các câu lệnh

```
python -m pip install numpy  
python -m pip install pandas  
python -m pip install matplotlib  
python -m pip install scipy
```

Cài đặt NLTK

NLTK là một thư viện Python phổ biến được sử dụng cho Natural Language Processing.

```
python -m pip install nltk
```

Khi đã cài đặt NLTK, ta cài đặt các gói NLTK bằng cách chạy mã sau

```
python
import nltk
nltk.download()
```

Cửa sổ tải xuống NLTK mở lên. Nhấp vào tab All Packages tải những gói có tên Averaged Perceptron Tagger, Stopwords Corpus, Penn Treebank Sample, WordNet.

Cài đặt Sklearn

Scikit-learn là một trong những thư viện mã nguồn mở Machine Learning viết bằng Python và được đông đảo mọi người sử dụng nhất hiện nay.

```
python -m pip install scikit-learn
```

Cách thực thi chương trình

Cấu trúc thư mục ta đã tải về như sau

```
NaiveBayes/
  datasets/
    emails.csv
  model/
    classifier.py
    get_data.py
    get_dict.py
```

Giả sử ta tải source codes ở thư mục Downloads (nếu tải về ở thư mục khác cách làm cũng tương tự)

Ta thấy có 2 thư mục con bên trong thư mục chính đó là datasets và model. Trong thư mục datasets có chứa file emails.csv là tập dữ liệu lấy từ [kaggle](https://www.kaggle.com/wordsnips/embeddings). Trong thư mục model chứa lần lượt các file classifier.py - dùng để phân loại email, get_data.py - các bước tiền xử lý dữ liệu được thực hiện ở file này, get_dict.py - tạo từ điển.

Đối với Ubuntu

- Với Ubuntu ta mở terminal bằng cách ấn tổ hợp phím Ctrl + Alt + T rồi chạy câu lệnh (chú ý đường dẫn file khi giải nén)

```
cd Downloads/NaiveBayes/model
```

- Thực thi file get_data.py đầu tiên

```
python3 get_data.py
```

Sau khi chạy câu lệnh trên ta sẽ thu được biểu đồ phân bố lớp của các nhãn trong từng tập dữ liệu, đồng thời các file train, test, val cũng được tạo ra trong thư mục datasets.

- Tiếp theo ta thực thi file classifier.py

```
python3 classifier.py
```

Ta sẽ nhận được đầu ra như sau

```
toan@fixcer:~/Downloads/NaiveBayes/model$ python3 classifier.py
Confusion Matrix:
[[188  22]
 [ 22 685]]

Accuracy = 95.202, Precision = 89.524, Recall = 89.524, F1 = 89.524
```

Như vậy ta đã thực thi xong chương trình.

Đối với Windows

- Với Windows ta mở cmd rồi chạy câu lệnh (chú ý đường dẫn file khi giải nén)

```
cd Downloads/NaiveBayes/model
```

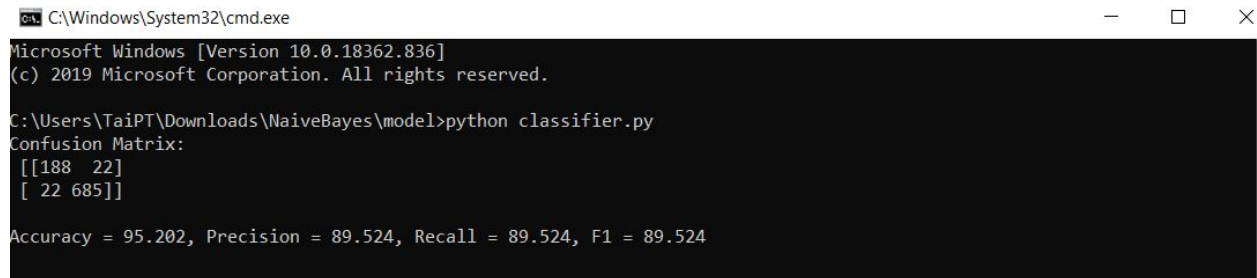
- Thực thi file get_data.py đầu tiên

```
python get_data.py
```

Sau khi chạy câu lệnh trên ta sẽ thu được biểu đồ phân bố lớp của các nhãn trong từng tập dữ liệu, đồng thời các file train, test, val cũng được tạo ra trong thư mục datasets.

- Tiếp theo ta thực thi file classifier.py

```
python classifier.py
```



```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.18362.836]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\TaiPT\Downloads\NaiveBayes\model>python classifier.py
Confusion Matrix:
[[188  22]
 [ 22 685]]

Accuracy = 95.202, Precision = 89.524, Recall = 89.524, F1 = 89.524
```

Như vậy ta đã thực thi xong chương trình.