Phishing Url

Note: Định dạng tệp dữ liệu csv:

- 40 trường dữ liệu (tính cả cột số thứ tự)

|  |  |  |  |
| --- | --- | --- | --- |
| # | Cột | Kiểu | Mô tả |
| 1 |  | int | Số thứ tự |
| 2 | url | string |  |
| 3 | Label | int | Nhãn dữ liệu:  0: Sạch  1: Phishing |
| 4 | url\_length | int |  |
| 5 | subdomain | string |  |
| 6 | domain | string |  |
| 7 | tld | string |  |
| 8 | tld\_digit\_count | int |  |
| 9 | domain\_digit\_count | int |  |
| 10 | subdomain\_digit\_count | int |  |
| 11 | tld\_length | int |  |
| 12 | subdomain\_length | int |  |
| 13 | domain\_length | int |  |
| 14 | isKnown\_tld | bool |  |
| 15 | issubdomainwww | bool |  |
| 16 | count\_of\_special\_char | int |  |
| 17 | creation\_date | date |  |
| 18 | expiration\_date | date |  |
| 19 | last\_updated | date |  |
| 20 | registrar | string |  |
| 21 | word\_process | string |  |
| 22 | count . | int |  |
| 23 | count - | int |  |
| 24 | count @ | int |  |
| 25 | count // | int |  |
| 26 | today\_date | date |  |
| 27 | creation\_date\_days | date |  |
| 28 | expiration\_date\_days | date |  |
| 29 | last\_updated\_days | date |  |
| 30 | domaincount - | int |  |
| 31 | domaincount . | int |  |
| 32 | domaincount / | int |  |
| 33 | domaincount @ | int |  |
| 34 | domaincount ? | int |  |
| 35 | domaincount & | int |  |
| 36 | domaincount = | int |  |
| 37 | domaincount \_ | int |  |
| 38 | None | int |  |
| 39 | http | int |  |
| 40 | https | int |  |

1. Module 1 - xử lý dữ liệu – preprocess.py

Gồm 2 hàm

+ train\_preprocess: Tiền xử lý cho dữ liệu huấn luyện (đầu ra dùng cho Module 2)

- Đầu vào:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Biến | Kiểu | Giá trị mặc định | Mô tả |
| 1 | data\_path | string | - | Đường dẫn file tập dữ liệu |
| 2 | model\_type | string | “sequential | loại mô hình |
| 3 | feature\_set | string | “full\_set” | loại tập features |
| 4 | test\_size | float | 0.3 | kích thước tập test |
| 5 | random\_state | int | random.randrange(100) | độ ngẫu nhiên của tập dữ liệu |
| 6 | number\_records | int | 100 |  |

- Đầu ra: 1 Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| # | Key | Value | Mô tả |
| 1 | 'X\_train' | array | Dữ liệu train của Module 2 |
| 2 | 'X\_test': | array | Dữ liệu test của Module 2 |
| 3 | 'y\_train': | array | Dữ liệu train của Module 3 |
| 4 | 'y\_test': | array | Dữ liệu test của Module 3 |
| 5 | 'model\_type' | string | Loại mô hình |
| 6 | 'message' | string | Thông báo |

+ inference\_preprocess: Tiền xử lý cho dữ liệu inference (đầu ra dùng cho Module 4)

- Đầu vào

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Biến | Kiểu | Giá trị mặc định | Mô tả |
| 1 | data\_path | string | none | Đường dẫn file tập dữ liệu |
| 2 | feature\_set | string | “full\_set” | loại tập features |

- Đầu ra: 1 Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| # | Key | Value | Mô tả |
| 1 | 'dataset' | array | Tập dữ liệu cho Module 4 |
| 2 | 'message' | string | Thông báo |

2. Module 2 - huấn luyện – train.py

- Đầu vào:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Biến | Kiểu | Giá trị mặc định | Mô tả |
| 1 | X\_train | string | - | Dữ liệu train của Module 2 |
| 2 | y\_train | string | - | Dữ liệu test của Module 2 |
| 3 | epoch\_num | string | - | số epoch |
| 4 | batch\_size | float | - | kích thước batch |
| 5 | model\_type | int | - | loại mô hình |
| 6 | model\_config | int | None | cấu hình mô hình |

- Đầu ra: 1 Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| # | Key | Value | Mô tả |
| 1 | 'output\_folder' | string | Thư mục lưu checkpoint – nơi lưu dữ liệu đầu vào cho Module 3, 4 |
| 2 | 'message' | string | Thông báo |

- Kết quả:

* Checkpoint lưu trong thư mục './output/'+datetime.now().strftime("%d %b %Y %Hh%M"), định dạng “.hdf5”

3. Module 3 - kiểm thử - test.py

- Đầu vào:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Biến | Kiểu | Giá trị mặc định | Mô tả |
| 1 | output\_folder | string | - | Thư mục lưu checkpoint – đầu ra của Module 2 |
| 2 | X\_test | array | - | Dữ liệu test của Module 3 |
| 3 | y\_test | array | - | Nhãn test của Module 3 |
| 4 | epoch\_num | string | None | số epoch |

- Đầu ra: 1 Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| # | Key | Value | Mô tả |
| 1 | 'message' | string | Thông báo Acc, Loss |

4. Module 4 - inference – inference.py

- Đầu vào:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Biến | Kiểu | Giá trị mặc định | Mô tả |
| 1 | output\_folder | string | - | Thư mục lưu checkpoint – đầu ra của Module 2 |
| 2 | data\_path | array | - | Dữ liệu cần kiểm tra phishing |
| 3 | epoch\_num | string | None | số epoch |

- Đầu ra: 1 Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| # | Key | Value | Mô tả |
| 1 | 'message' | string | Thư mục chưa inference |
| 2 | 'inference\_result' | list | Kết quả đánh giá phishing |