Xây dựng mô hình Machine Learning dự báo cháy rừng ở các tỉnh Tây Nguyên dựa vào dữ liệu lịch sử thời tiết.

Nguyễn Đại Kỳ 19521731
Văn Viết Hiếu Anh 19521225
Lê Văn Phước

Ngày 2 tháng 8 năm 2021

Môn học:	CS114 - Máy học
Giảng viên hướng dẫn:	Lê Đình Duy
	Phạm Nguyễn Trường An

Mục lục

1 Tổng quan

Bài viết là về quá trình thực nghiệm nghiên cứu các model Machine Learning với mục đích chọn ra mô hình tối ưu để dự đoán mức độ cháy rừng dựa vào dữ liệu thời tiết trong lịch sử của từng địa phương. Với mục đích hổ trợ trong việc dự đoán để phục vụ trong công tác phòng chống cháy rừng ở nước ta. Vì bài viết là ghi chép của quá trình thực nghiệm nên sẽ có nhiều phương pháp được đưa ra sử dụng.

1.1 Mô tả bài toán

Ở nước ta có 3 thảm họa lớn nhất, gây thiệt hại lớn hàng năm về cả người và của. Cùng với lũ lụt và hạn hán, cháy rừng là một thảm họa gây thiệt hại không chỉ về kinh tế mà còn cả con người và hệ sinh thái. Theo thống kê của Cục Kiểm lâm từ năm 1992 đến 2006, trung bình mỗi năm xảy ra 1254 vụ cháy rừng gây thiệt hại khoảng 6646 ha rừng, trong đó có 2854 ha là rừng tự nhiên và 3791 ha là rừng trồng. Bên cạnh việc nâng cao năng lực phòng cháy chữa cháy rừng (PCCCR) cho lực lượng kiểm lâm như đầu tư trang thiết bị, cơ sở vật chất, xây dựng cơ chế điều hành phối hợp và tuyên truyền nâng cao nhận thức trách nhiệm của chủ rừng và người dân, công tác cảnh báo nguy cơ cháy rừng cũng như tổ chức phát hiện sớm và thông báo kịp thời điểm cháy rừng là rất cần thiết.

Từ đầu năm 2007, Cục Kiểm lâm (Bộ Nông nghiệp và Phát triển Nông thôn) đã lắp đặt và vận hành trạm thu ảnh viễn thám MODIS tại Hà Nội với mục đích chính là phát hiện sớm các điểm cháy rừng (hotspots) trên toàn lãnh thổ Việt Nam. Hệ thống trạm thu của TeraScan đã tự động thu nhận, xử lý và sao lưu dữ liệu ảnh MODIS hàng ngày từ 2 vệ tinh TERRA và AQUA với mô-đun Vulcan tự động xử lý và tạo ra dữ liệu các điểm cháy sử dụng thuật toán ATBD-MOD14 [**?**].

Hệ thống này cung cấp dữ liệu về điểm cháy ghi nhận được từ vệ tinh và lưu lại thời gian và tọa độ cháy. Từ khi bắt đầu lắp đặt đến nay hệ thống dữ liệu cháy của cục kiểm lâm được ghi lại được gần 1 triệu điểm cháy. Nhờ lượng dữ liệu này việc xây dựng một hệ thống tự động phân tích mức độ cháy rừng dựa vào các đặc trưng cơ bản của dữ liệu khí tượng thủy văn là hoàn toàn có cơ sở và khả quan.

Input của bài toán là dữ liệu lịch sử thời tiết trong vòng 1 tháng trở lại và từ đó để model đánh giá địa phương đó vào ngày mai có mức độ cháy được đánh giá ở thang nào.

Chưa làm xong đâu, làm phụ đi

1.2 Mô tả dữ liệu

Nguồn dữ liệu của nhóm đến từ các website gồm 3 website chính:

- firewatchvn.kiemlam.org.vn: là Hệ thống theo dõi cháy rừng trực tuyến thuộc Cục Kiểm Lâm - Tổng cục Lâm Nghiệp
- weather.com: là website của The Weather Channel (TWC) - IBM [**?**]
- worldweatheronline.com

Trong 3 nguồn dữ liệu thì chỉ có worldweatheronline.com sử dụng SSR(Server-side render) còn 2 nguồn còn lại đều sử dụng Ajax để truyền dữ liệu qua lại giữa server.

1.2.1 Weather Data

Việc tìm các nguồn dữ liệu khác về thời tiết ngoài 2 nguồn trên đã được thực hiện song các nguồn này đều có những điểm thiếu rất quan trọng ví dụ như API của website chỉ cung cấp trong 1 năm trở lại hay các website này không cung cấp đủ nhiều địa phương mà chỉ cung cấp dữ liệu ở những thành phố cụ thể.

Có thể nói việc lấy dữ liệu thời tiết là công đoạn gây ra nhiều khó khăn nhất. Đa phần dữ liệu lịch sử thời tiết là rất lớn và các công ty hay tập đoàn công nghệ đều dùng để bán chứ không public trên website của họ. Ngay cả trên giao diện chính của weather.com của IBM cũng chỉ hiển thị dữ liệu thời tiết trong 2 năm trở lại (tức là 2021 và 2020).

Tuy nhiên vì sử dụng Ajax nên sau khi phân tích và ghi lại các request mà website gửi đi cũng như các response nhận về, việc có thể tìm được các cổng API và phương thức giao tiếp với server, từ đó dùng vào việc khai thác dữ liệu tự động trong nhiều năm trước nữa là hoàn toàn có hy vọng.

Về việc tìm nguồn dữ liệu tương tự đã được khai thác trước, nhóm đã từng thử tìm kiếm nhưng để đạt được yêu cầu chi tiết đến từng địa phương với thời gian kéo dài thì không tìm được dữ liệu nào đạt yêu cầu. Ngay cả khi join vào Slack của Call For Code năm nay để xin hỗ trợ vì đề tài này

có liên quan đến cuộc thi thì phía ban tổ chức cuộc thi cũng trả lời rằng dữ liệu này không cung cấp cho thí sinh. Ngoài ra nhóm cũng đã thử gửi mail cho Trung tâm Dự báo khí tượng thuỷ văn quốc gia nhưng cũng không nhận được phản hồi. Hiển nhiên, việc tự đi thu thập dữ liệu là tất yếu.

1.2.2 Fire Data

firewatchvn.kiemlam.org.vn là website tạo ra ý tưởng cho nhóm. Website này cung cấp giao diện tra cứu dữ liệu về các điểm cháy vào từng thời gian cụ thể. Tuy nhiên điểm yếu của website này là xây dựng quá nhiều tính năng và sử dụng Ajax nên dùng những công cụ như Beautiful Soup thì không thể thu thập còn nếu dùng những công cụ như Selenium hay Puppeteer thì tốc độ quá chậm (dữ liệu này kéo dài từ 1/1/2008 đến 5/12/2020 nếu tra cứu từng ngày trên 700 quận, huyện, thành phố thuộc tỉnh,... thì sẽ mất rất nhiều thời gian). Điều này bắt buộc nhóm phải phân tích API mà website đã lấy dữ liệu điểm cháy để tăng tốc độ lấy dữ liệu. Bởi vì những thông tin như bản đồ của địa điểm lấy dữ liệu là không cần thiết, ta hoàn toàn có thể lấy bản đồ địa hình của cả trái đất chỉ cần dùng tọa độ. Việc lấy những thông tin nặng như bản đồ cần thời gian tải rất lâu nên việc tìm ra API mà website sử dụng cũng là cần thiết.

2 Các nghiên cứu trước

2.1 Predicting Australia wildfires with weather data a Call For Code spot challenge of IBM [?]

Mục đích của cuộc thi này là dự đoán kích thước của khu vực cháy tính bằng km2 theo khu vực ở Úc cho mỗi ngày trong tháng 2 năm 2021 bằng cách sử dụng dữ liệu có sẵn cho đến ngày 29 tháng 1. Bài toán chuỗi thời gian dựa trên dữ liệu hàng ngày do Pairs Geoscope cung cấp. Dữ liệu được cung cấp bao gồm:

- Những trận cháy rừng trong lịch sử
- Thời tiết lịch sử
- Các dự báo thời tiết lịch sử
- Chỉ số thảm thực vật lịch sử
- Các loại đất của các địa điểm xảy ra cháy rừng.

Trong đó dữ liệu thời tiết lịch sử bao gồm các trường như **vùng**, **thời gian**, **lượng mưa (mm/ngày)**, **Độ ẩm tương đối (%)**, **Hàm lượng nước trong đất** (m^3/m^3), **bức xạ mặt trời (MJ/ngày)**, **nhiệt độ (C)**, **tốc độ gió (m/s)**. Đây là dữ liệu chính áp dụng vào bài toán của nhóm. Tuy nhiên vì khó khăn trong việc tìm kiếm dữ liệu nên nhóm chỉ đáp 4/6 tiêu chí mà các chuyên gia đã đưa ra đó là lượng mưa, độ ẩm, nhiệt độ, tốc độ gió. Từ những nghiên cứu các chuyên gia đưa ra các biến ảnh hưởng đến cháy rừng như:

- Lãnh thổ:
  - Khu vực: Số lượng và cường độ đám cháy khác nhau ở các khu vực khác nhau. Các sự kiện ở các khu vực lân cận có thê ảnh hưởng đến cháy rừng trong một lãnh thổ nhất định. Thành phần tự phục hồi(sự diện diện của cháy rừng trong những ngày trước đó).
  - Tính theo mùa: Cháy rừng đặc biệt dữ dội trong "Mùa cháy rừng" kéo dài từ tháng 10 đến tháng 12. Quan sát này sẽ ảnh hưởng đến cách thức phân chia tập dữ liệu huấn luận và kiểm tra.
- Điều kiện đất và khí quyển:
  - Thời tiết và đất đai: Thời tiết và Hạn hán có liên quan chặt chẽ đến hỏa hoạn. So sánh giữa lịch sử và dự báo thời tiết trong lịch sử.
  - Thảm thực vật: Có mối tương quan thuận giữa sự thay đổi chỉ số thảm thực vật và cường độ cháy.
  - Sử dụng đất: Việc sử dụng đất có thể liên quan đến việc dự đoán sự kéo dài của cháy rừng. Tuy nhiên, dữ liệu này chỉ có sẵn dưới dạng một hàng – cho mọi khu vực, do đó nó không được đưa vào mô hình.

Theo báo các tổng kết cuộc thi thì mô hình được sử dụng là Convolutional Neural Network (Windowing Dataset, Conv1D Layers,...) cho ra kết quả RMSE: 19.96, MAE: 6.94, TOT: 9.54 Qua nghiên cứu của cuộc thi này nhóm rút ra được các biến ảnh hưởng đến cháy rừng, cách khai thác lấy dữ liệu từ thực tế, mô hình đào tạo phù hợp với bài toán.

2.2 AI Climate data for predicting fire frequency in California [?]

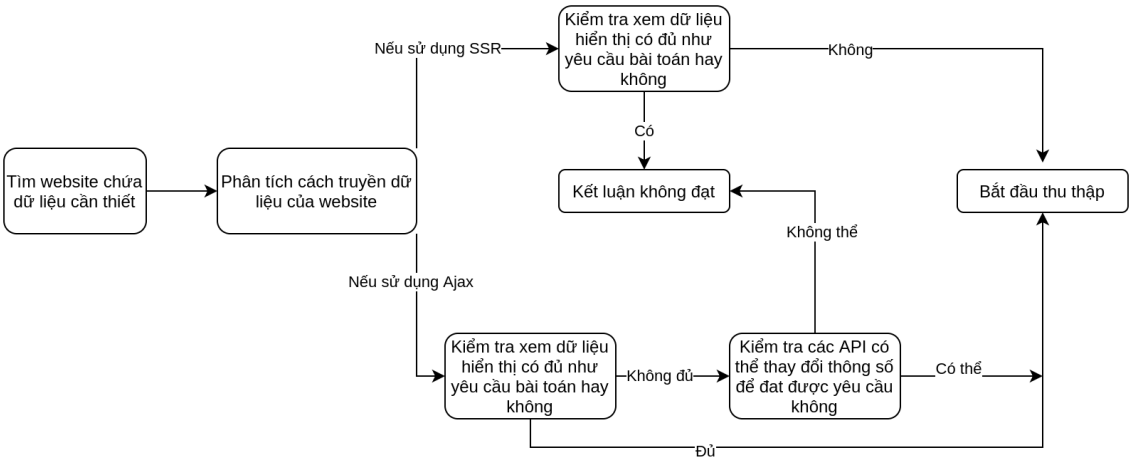
Mục đích của nghiên cứu này là dự đoán tần suất xảy ra cháy rừng có thể giúp lập kế hoạch khẩn cấp và chủ động quản lý rủ ro thiên tai. Dữ liệu về đám cháy được thu từ Kaggle(`https://www.kaggle.com/rtatman/188-million-us-wildfires`), bao gồm khoảng 1,9 triệu vụ cháy rừng ở Mỹ được tham chiếu trong giai đoạn 1992-2015. Ngoài ra, dữ liệu khí hậu từ Copernicus ERA5 được tải xuống và sử dụng cho bài toán này. Bài nghiên cứu này cũng đưa ra kết luận về việc cháy rừng thường xảy ra hàng tháng hoặc theo mùa. Ở đây, các chuyên gia đưa ra các trường dữ liệu thời tiết được lấy là $\mathbf{total}_precipitation, \mathbf{2m}_temperature, \mathbf{2m}_dewpoint_temperature, \mathbf{10m}_wind_speed, \mathbf{volumetric}_soil_water, \mathbf{potential}_evapor$

Chương này mô tả quá trình thu thập dữ liệu. Nếu dữ liệu crawling tự động thì mô tả cách viết crawler, các khó khăn gặp phải và các số liệu liên quan. Nếu dữ liệu thu thập thủ công thì mô tả các tiêu chí đặt ra để thống nhất trong nhóm khi thu thập. Làm sao để đảm bảo bộ dữ liệu thu thập thủ công có thể khớp gần giống với ngữ cảnh ứng dụng của bài toán.

Sau đó mô tả các thông số chi tiết của bộ dữ liệu, kèm theo ví dụ minh họa rõ ràng. Bài toán đặt ra các trường hợp dữ liệu nào là khó xử lý, có bao nhiêu mẫu dữ liệu thuộc trường hợp đó, chụp vài mẫu dữ liệu khó đó vào báo cáo để minh họa.

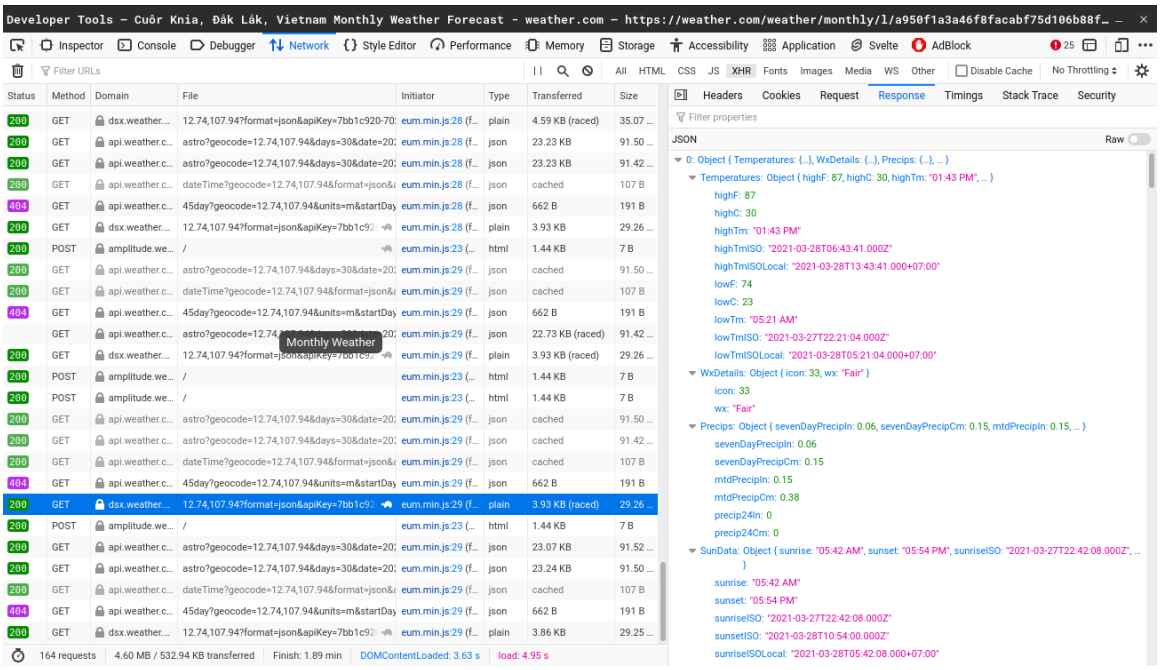
3.1 Quá trình thu thập dữ liệu

Tất cả dữ liệu của nhóm đều bắt đầu bằng 1 quy trình chung để kiểm tra xem nguồn dữ liệu có đáp ứng các yêu cầu hay không.

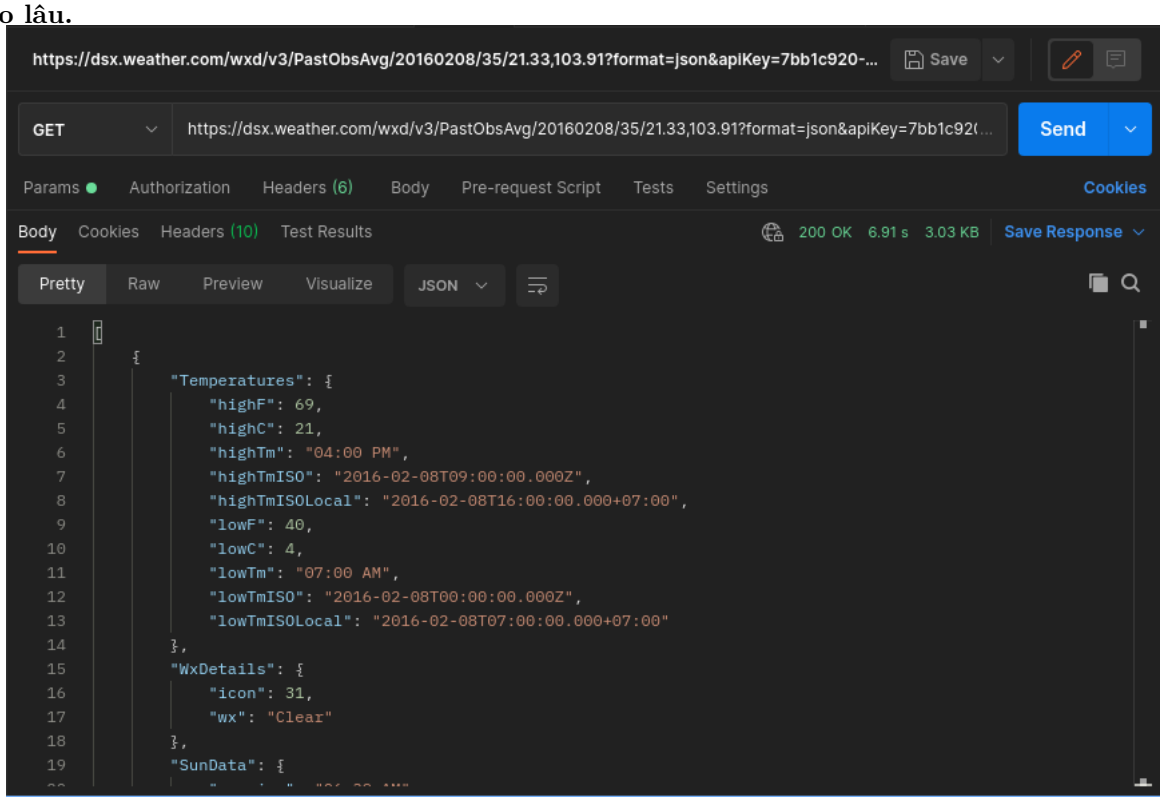


3.1.1 weather.com

Để tìm ra API phục vụ cho việc crawl ta vào trang Monthly của một địa phương cụ thể sau đó navigate đến các trang chứa dữ liệu của các tháng trước (việc này giúp cho website sử dụng các API yêu cầu dữ liệu thời tiết quá khứ). Kiểm tra các response trong filter XHR ta có thể tìm ra những response tốt nhất chứa dữ liệu cần thiết.



Tiếp theo đó khi đã xác nhận API này đủ điều kiện ra kiểm tra ý nghĩa của các thông số trong request, thử thay đổi thông số để kiểm tra dữ liệu kéo dài được đến bao lâu.



Vì dữ liệu chấp nhận đến 2014 nên sẽ tiến hành crawl dữ liệu. Nhóm sử dụng Scrappy để crawl vì công cụ này hỗ trợ async cho phép gửi nhiều request cùng lúc cùng với đó là có pipeline để lưu dữ liệu nên giảm rất nhiều quá trình cài đặt.

Đầu tiên ta tạo tải dữ liệu danh sách chứa tất cả các xã cần crawl dữ liệu, làm tròn các tọa độ đến 2 chữ số thập phân(theo yêu cầu của API, xử lý bằng numpy trước sẽ nhanh hơn việc xử lý trong vòng lặp).

```
class IBMWeatherScapper(Spider):
    name = 'ibm-weather'

    df_ward_taynguyen_fire = pandas.read_csv(
        'https://raw.githubusercontent.com/.../taynguyen_wards_longlat.csv')
```

```
df_ward_taynguyen_fire['long'] = numpy.round(
    df_ward_taynguyen_fire['long'], 2)
df_ward_taynguyen_fire['lat'] = numpy.round(
    df_ward_taynguyen_fire['lat'], 2)
```

Trong hàm start_requests của Spider với mỗi xã ta tạo 1 request vào ngày 1/1/2014 với tọa độ của xã đó, cùng với đó là lưu kèm 1 meta để nhận diện xã đó trong parse function.

```
def start_requests(self):
requests = []
for i, ward in self.df_ward_taynguyen_fire.iterrows():
    requests.append(request.Request(
        url=f"https://dsx.weather.com/wxd/v3/PastObsAvg/20140101/35/\
        {ward['lat']:.2f},{ward['long']:.2f}\
        ?format=json\
        &apiKey=7bb1c920-7027-4289-9c96-ae5e263980bc&\
        fbclid=IwAR1IgpD8qPU6ZaHqDnZT1tMl95Y4G8gfGvIYmpM3CGGIqyjwQeaAmbZZ8SE",
        meta={
            'ward_code': ward['ward_code'],
            'long':ward['long'],
            'lat':ward['lat']
        },
        callback=self.parse
    ))
return requests
```

Vì đây là API nên dữ liệu được thể hiện dưới dạng json rất rõ ràng.


```
{
    "Temperatures": {
        "highF": 72,
        "highC": 22,
        "highTm": "01:00 PM",
        "highTmISO": "2014-01-01T06:00:00.000Z",
        "highTmISOLocal": "2014-01-01T13:00:00.000+07:00",
        "lowF": 68,
        "lowC": 20,
        "lowTm": "01:00 AM",
        "lowTmISO": "2013-12-31T18:00:00.000Z",
        "lowTmISOLocal": "2014-01-01T01:00:00.000+07:00"
    },
    "WxDetails": { "icon": 27, "wx": "Mostly Cloudy" },
    "SunData": {
        "sunrise": "06:05 AM",
        "sunset": "05:26 PM",
        "sunriseISO": "2013-12-31T23:05:00.000Z",
        "sunsetISO": "2014-01-01T10:26:00.000Z",
        "sunriseISOLocal": "2014-01-01T06:05:00.000+07:00",
        "sunsetISOLocal": "2014-01-01T17:26:00.000+07:00"
    },
    "Moon": {
        "moonriseLocal": "2014-01-01T05:34:00.000+07:00",
        "moonsetLocal": "2014-01-01T17:25:00.000+07:00",
        "moonriseISO": "2013-12-31T22:34:00.000Z",
        "moonsetlISO": "2014-01-01T10:25:00.000Z"
    }
}
```


Việc còn lại là lấy ra và lưu lại dưới file csv. Sau đó lấy ngày cuối cùng trong dữ liệu trả về xem có bằng ngày hôm nay không, nếu không thì tiếp tục tải dữ liệu còn nếu có thì dừng lại.

```
def parse(self, response, **kwargs):
    json = response.json()
    ward_code = response.meta['ward_code']
    long = response.meta['long']
    lat = response.meta['lat']
    for d in json:
        yield {
            'ward':ward_code,
            'date':d.get('Temperatures', {}).get('highTmISOLocal','')[:10],
            'highC': d.get('Temperatures', {}).get('highC',''),
            'lowC': d.get('Temperatures', {}).get('lowC',''),
            'sun_rise': d.get('SunData', {}).get('sunrise',''),
            'sun_set': d.get('SunData',{}).get('sunset',''),
            'sevenDayPrecipCm':d.get('Precips',{}).get('sevenDayPrecipCm',''),
            'mtdPrecipCm':d.get('Precips',{}).get('mtdPrecipCm',''),
            'precip24Cm':d.get('Precips',{}).get('precip24Cm',''),
        }

    last = json[-1]
    recentDate = datetime.strptime(
        last['Temperatures']['highTmISO'][:10], '%Y-%M-%d')
    next = recentDate + timedelta(days=1)

    if recentDate.date() == datetime.today():
        return
    else:
        yield response.follow(
            url=f"https://dsx.weather.com/wxd/v3/PastObsAvg/ \
            {next.strftime('%Y%M%d')}/35/{lat:.2f},{long:.2f}\
            ?format=json\
            &apiKey=7bb1c920-7027-4289-9c96-ae5e263980bc\
            &fbclid=IwAR1IgpD8qPU6ZaHqDnZT1tMl95Y4G8gfGvIYmpM3CGGIqyjwQeaAmbZZ8SE",
            meta={
                    'ward_code': ward_code,
                    'long':long, 'lat':lat
                },
                callback=self.parse
        )
```

Dữ liệu từ website này cung cấp có độ chính xác đến từng tọa độ, có nghĩa là chỉ cần cung cấp tọa độ (làm tròn đến 2 chữ số thập phân) thì server sẽ trả về thời tiết tại điểm đó tùy vào thời gian mà ta muốn. Tuy nhiên điểm yếu của dữ liệu này là chỉ cung cấp các đặc tính cơ bản nhất của thời tiết tại địa điểm đó gồm: nhiệt độ cao nhất và thấp nhất trong ngày, thời gian mặt trời mọc và lặn, lượng mưa (tích lũy trong 7 ngày, trong 1 tháng hoặc chỉ ngày hôm đó).

Sau khoảng nhiều ngày khai thác và xử lý, nhóm đã lấy được dữ liệu của 5 tỉnh Tây Nguyên vào từng xã từng ngày kéo dài từ 1/1/2014 đến 8/6/2021. Dữ liệu gồm các trường cơ bản sau (lưu ý các trường này đã được đổi tên so với khi crawl dữ liệu để thống nhất các bộ dữ liệu với nhau nhằm dễ dàng cho việc nghiên cứu):

- ward: Mã của xã, phường, thị trấn, thị xã. Vì sẽ có những địa điểm trùng tên nên nhóm sử dụng mã để phân biệt các địa phương (mã này cung cấp bởi API của firewatchvn.kiemlam.org.vn [?])
- date: là ngày mà record được ghi lại.
- max/min: là nhiệt độ cao nhất và thấp nhất được ghi nhận trong ngày(celcius).
- sunrise/sunset: là thời gian mặt trời mọc và lặn.
- 7_rain: lượng mưa tổng tính từ ngày chủ nhật gần nhất trước đó (cm)
- m_rain: lượng mưa tổng tính từ ngày 1 của tháng đó(cm)

- **24_rain: lượng mưa ghi nhận trong ngày(cm)**

Dưới đây là một đoạn mẫu trong dữ liệu.

ward	date	max	min	sunrise	sunset	7_rain	m_rain	24_rain
24727.0	2016-10-17	30	25	05:36 AM	05:28 PM	5.94	30.91	1.82
24727.0	2016-10-18	30	25	05:36 AM	05:28 PM	7.54	32.66	1.75
24727.0	2016-10-19	30	25	05:36 AM	05:27 PM	7.54	32.96	0.3
24761.0	2019-10-14	32	24	05:36 AM	05:30 PM	0.02	0.07	0.0
24761.0	2019-10-15	32	25	05:36 AM	05:30 PM	0.02	0.07	0.0
24761.0	2019-10-16	32	25	05:36 AM	05:29 PM	0.02	0.07	0.02
24761.0	2019-10-17	32	25	05:36 AM	05:29 PM	0.02	0.07	0.0
24761.0	2019-10-18	32	25	05:36 AM	05:28 PM	0.02	0.07	0.0
24761.0	2019-10-19	32	26	05:36 AM	05:28 PM	0.02	0.07	0.0
24761.0	2019-10-20	32	26	05:36 AM	05:27 PM	0.02	0.1	0.0
24761.0	2019-10-21	33	25	05:36 AM	05:27 PM	0.02	0.1	0.0
24761.0	2019-10-22	32	25	05:37 AM	05:26 PM	0.02	0.1	0.0
24761.0	2019-10-23	33	25	05:37 AM	05:26 PM	0.0	0.1	0.0
24761.0	2019-10-24	33	25	05:37 AM	05:25 PM	0.0	0.1	0.0

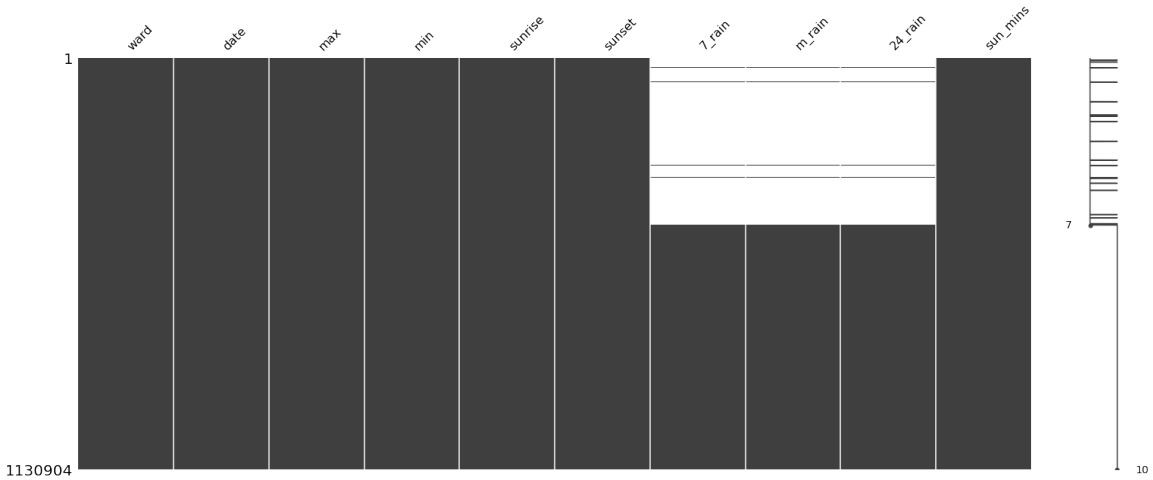
Dưới đây là các thông số của bộ dữ liệu

	ward	max	min	7_rain	m_rain	24_rain
count	1130904	1130904	1130904	679514	679514	679514
mean		30.60013	24.66924	33.813765	82.088195	4.797731
std		2.836824	2.207649	71.482188	153.732438	16.574482
min		18	14	0	0	0
25%		29	23	0.5	2.2	0
50%		31	25	8.1	21.3	0
75%		32	26	34.7	88.6	2.7
max		46	36	903.7	1300.7	419.1

Kết quả trả về từ hàm nunique():

ward	592
date	2739
max	25
min	21
sunrise	75
sunset	74
7_rain	644
m_rain	976
24_rain	231

Dữ liệu bị trống sau khi sắp xếp theo ngày và địa phương(theo thứ tự ngày trước địa phương sau, tăng dần):



3.1.2 worldweatheronline.com

3.1.3 firewatchvn.kiemlam.org.vn

3.2 Xây dựng bộ dữ liệu

3.2.1 Imputation of Data

3.2.2 Creation of Dataset

4 Training và đánh giá modelods

5 Ứng dụng và hướng phát triển