**Dự báo lượng mạng Internet theo chu kỳ giờ/ngày**

1. **Giới thiệu**

**1.1 Tổng quan bài toán**

Trong thời đại số, các hệ thống máy chủ và mạng máy tính đóng vai trò quan trọng trong việc duy trì hoạt động liên tục của các dịch vụ trực tuyến, từ website, ứng dụng di động, đến hệ thống dữ liệu lớn. Một trong những bài toán quan trọng mà các nhà quản lý hệ thống quan tâm là: **dự báo lưu lượng truy cập Internet**. Việc dự đoán trước được các khoảng thời gian cao điểm, thấp điểm giúp tối ưu hóa phân bổ tài nguyên, đảm bảo hiệu suất vận hành, đồng thời giảm thiểu nguy cơ gián đoạn dịch vụ.

**1.2 Mục tiêu**

Phân tích chu kỳ cao điểm/lặp lại trong dữ liệu.

Dự báo lưu lượng mạng trong tương lai gần.

Trực quan hóa các chu kỳ nổi bật giúp hiểu hành vi mạng.

**1.3 Các công việc liên quan**

Tiền xử lý dữ liệu

Phân tích phổ tần số bằng biến đổi Fourier nhanh (FFT)

Ước lượng và dự báo lưu lượng tương lai

Phân tích chu kỳ theo giờ và theo ngày trong tuần

1. **Cơ sở lý thuyết**

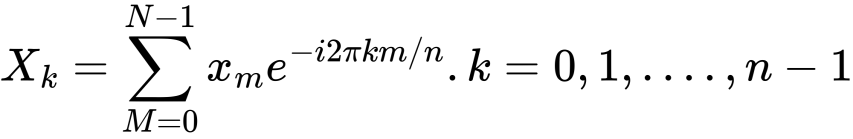
**2.1 Biến đổi Fourier nhanh (FFT)**

Biến đổi Fourier nhanh ( FFT ) là [thuật toán](https://en.wikipedia.org/wiki/Algorithm" \o "Thuật toán) tính toán [biến đổi Fourier rời rạc](https://en.wikipedia.org/wiki/Discrete_Fourier_transform" \o "Biến đổi Fourier rời rạc) (DFT) của một chuỗi hoặc nghịch đảo của nó (IDFT). [Biến đổi Fourier](https://en.wikipedia.org/wiki/Fourier_transform" \o "Biến đổi Fourier) chuyển đổi tín hiệu từ miền gốc của nó (thường là thời gian hoặc không gian) thành biểu diễn trong [miền tần số](https://en.wikipedia.org/wiki/Frequency_domain" \o "Miền tần số) và ngược lại.

Nói một cách đễ hiểu : **DFT** giúp bạn phân tích tín hiệu rời rạc (dữ liệu dạng chuỗi, như lưu lượng mạng, âm thanh số, v.v.) thành các thành phần tần số. Nhưng tính DFT trực tiếp rất tốn thời gian: phải thực hiện khoảng n2 phép tính cho n điểm. **FFT** giảm số phép tính xuống chỉ khoảng nlogn , nhanh hơn rất nhiều, nhờ chia nhỏ bài toán.

**2.2 Công thức DFT**

Cho x0,x1,...xn-1 là số phức. DFT được định nghĩa bởi công thức



wps : Giá trị ở tần số k

wps: nhân tử xoay (twiddle factor), chứa pha và biên độ.

**2.3 Ý tưởng**

Ý tưởng của FFT là dùng phương pháp chia để trị:

+ Với một chuỗi x[n] dài N, thay vì tính DFT trực tiếp (rất tốn công), FFT chia dãy ra thành một dãy chuỗi lẻ và dãy chuỗi chẵn

+ Sau đó tính riêng DFT của mỗi chuỗi rồi kết hai phần lại bằng tính chất đối xứng. Kết quả giảm số phép tính n2 còn nlogn

1. **Thực nghiệm**

**3.1 Giới thiệu dữ liệu**

Link dữ liệu : [https://www.kaggle.com/datasets/raminhuseyn/web-traffic-time-series-dataset](https://www.kaggle.com/datasets/raminhuseyn/web-traffic-time-series-dataset" \t "https://assignments.onenote.com/classes/b53d18ae-b8b0-4004-bddd-194098eb98b5/assignments/_blank)

web\_traffic.csv là tập dữ liệu ghi nhận lưu lượng truy cập Internet từ một máy chủ trong vòng 2 tuần ( 2793 dòng x 2 cột), được thu thập theo từng mốc thời gian cụ thể.

web\_traffic.csv có những cột như :

Timesamp : thời gia ghi nhận lưu lượng truy cập

Trafficcount : Số lượng yêu cầu truy cập

# 3.2 Tiền xử lý

Chuyển cột Timesamp thành dạng datetime

df[ 'Timestamp' ] = pd.to\_datetime(df[ 'Timestamp' ])

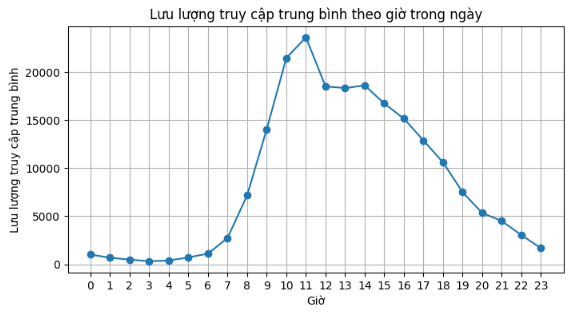
Tạo thêm cột giờ và cột tuần để vẽ trực quan hóa

df[ 'hour' ] = df[ 'Timestamp' ].dt.hour

df[' weekday' ] = df[ 'Timestamp'].dt.day\_name()

**3.3 Trực quan hóa**

**3.3.1 Lưu lượng truy cập trung bình theo giờ trong ngày**



Có thể thấy rằng từ:

+ 0h - 6h lưu lượng thấp -> ít người dùng internet vào ban đêm

+ 7h - 9h bắt đầu tằng mạnh -> người dùng bắt đầu online(học tập , làm việc)

+ 11h -12h đạt đỉnh -> có thể do nghỉ trưa, nhiều người lượt web,xem tin tức

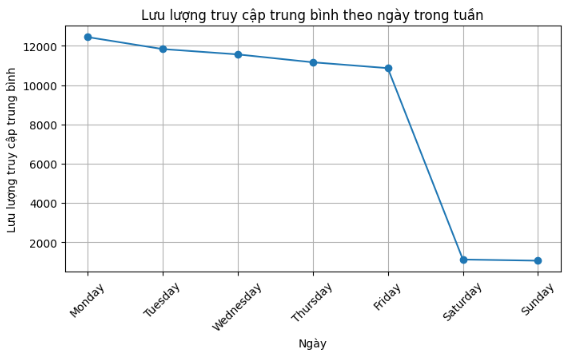
+ 13h - 17h -> duy trì mức cao nhưng giảm dần

+ Sau 18h tiếp tục giảm -> có thể mọi người nghỉ ngơi

+ Sau 23h giảm hẳn -> ít người dùng

Từ những ý trên giúp ta xác định **giờ cao điểm để** phân bổ tài nguyên hệ thống, tối ưu server, tránh quá tải

**3.3.2 Lưu lượng truy cập trung bình theo ngày trong tuần**



Có thể thấy rằng từ :

+ Thứ 2 - 6 : Lưu lượng duy trì cao, dao động ~10,500–12,000 → cao điểm do nhu cầu làm việc, học tập, sử dụng dịch vụ online.

+ Thứ 7 - CN : Giảm đột ngột xuống chỉ khoảng 1,500 → cực thấp so với ngày thường

**3.4 Mô hình FFT**

traffic = df['TrafficCount'].values

n = len(traffic)

dt = 1

freqs = np.fft.fftfreq(n, d=dt)

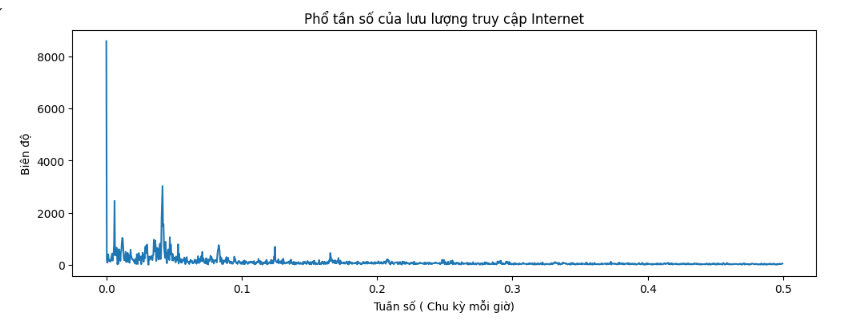
fft\_values = np.fft.fft(traffic)

amplitude = np.abs(fft\_values) / n

positive\_freqs = freqs[:n//2]

positive\_amplitude = amplitude[:n//2]

Dầu tiên lấy dữ liệu cột [‘TrafficCount’], áp dụng biến đổi Fourier nhanh (FFT) để phân tích phổ tần số, chỉ giữ phần tần số dương, rồi tìm tần số mạnh nhất (dominant frequency) có biên độ lớn nhất. Cuối cùng, nó tính tần số góc ω và in ra tần số nổi bật cùng chu kỳ lặp lại (giờ).



dominant\_idx = np.argmax(positive\_amplitude[1:]) + 1

dominant\_freq = positive\_freqs[dominant\_idx]

omega = 2 \* np.pi \* dominant\_freq

print(f"Tần số: {omega:.4f}")

print(f"Tần số mạnh nhất : {dominant\_freq:.4f} cycles/hour → Chu kỳ ≈ {1/dominant\_freq:.2f} hours")



**3.5 Dự đoán**

def sin\_func(t, A, omega, phase, offset):

    return A \* np.sin(omega \* t + phase) + offset

t = np.arange(n)

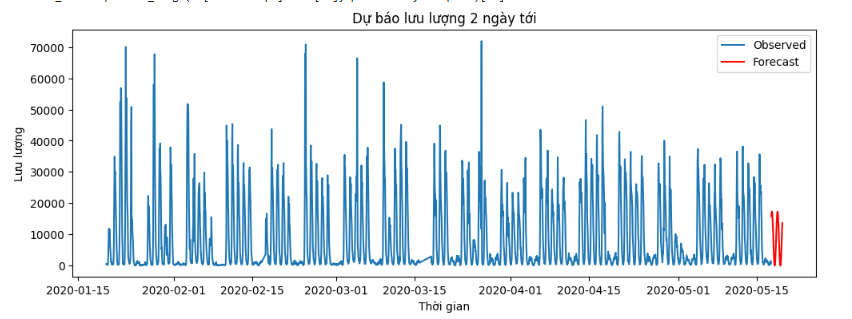
popt, \_ = curve\_fit(sin\_func, t, traffic, p0=[np.std(traffic), omega, 0, np.mean(traffic)])

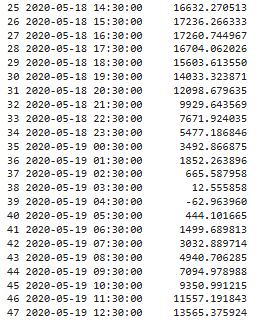
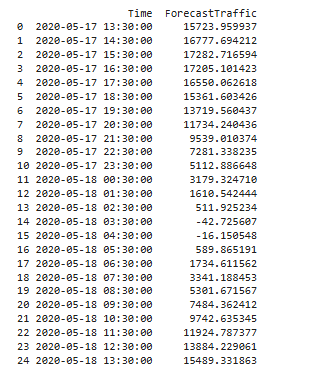
# Dự báo 48 giờ tới

future\_t = np.arange(n, n + 48)

future\_traffic = sin\_func(future\_t, \*popt)

Đầu tiên định nghĩa hàm sin\_func với các tham số biên độ A, tần số góc ω, pha phase, và độ lệch offset. Tiếp theo, sử dụng hàm curve\_fit để khớp hàm sin này vào dữ liệu thực tế, với các giá trị khởi tạo dựa trên thống kê từ dữ liệu (độ lệch chuẩn, trung bình và tần số góc đã tính). Sau khi tìm được các tham số tối ưu, mã tạo ra một dãy thời gian tương lai 48 giờ và dùng hàm sin đã khớp để dự báo lưu lượng traffic trong khoảng thời gian đó.





**IV Kết luận**

Trong bài báo cáo này, em đã tiến hành phân tích và dự báo lưu lượng truy cập Internet dựa trên dữ liệu chuỗi thời gian thu thập theo giờ. Bằng cách áp dụng biến đổi Fourier nhanh (FFT), đã xác định được các chu kỳ lặp lại nổi bật trong dữ liệu lưu lượng, từ đó giúp hiểu rõ hơn về hành vi sử dụng Internet của người dùng theo giờ và ngày trong tuần.

Việc trực quan hóa dữ liệu cho thấy rõ các khung giờ cao điểm (như 11h – 12h trưa), cũng như sự sụt giảm đáng kể vào cuối tuần. Dựa vào phân tích phổ tần số và mô hình hàm sin, em đã xây dựng được mô hình dự báo lưu lượng trong tương lai một cách tương đối hiệu quả.

Kết quả này có thể được ứng dụng trong việc tối ưu hóa tài nguyên hệ thống mạng, lên lịch bảo trì hệ thống hợp lý, và nâng cao chất lượng dịch vụ bằng cách chuẩn bị sẵn sàng cho các thời điểm có lưu lượng cao.

Trong tương lai, có thể mở rộng nghiên cứu bằng cách thử nghiệm với các mô hình dự báo phức tạp hơn như ARIMA, LSTM hoặc kết hợp thêm các yếu tố thời tiết, sự kiện,… để tăng độ chính xác cho việc dự báo lưu lượng truy cập.

Link Github: <https://github.com/waanuu/TH4>