**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

-----📖-----

# BÁO CÁO ĐỒ ÁN

## "PHÂN TÍCH SỰ PHẢN ỨNG CỦA NGƯỜI DÙNG TWITTER ĐỐI VỚI TUYÊN BỐ CỦA ELON MUSK VỀ BITCOIN NGÀY 13/05/2021"

**Môn học: Phân tích dữ liệu web**

**GVHD: Thầy Đặng Nhân Cách**

**Nhóm thực hiện: 5Deiz**

| STT | Họ và tên | MSSV | Chức vụ |
|---|---|---|---|
| 01 | Trần Trí Tín | K184111429 | Nhóm trưởng |
| 02 | Trần Minh Nghĩa | K184111395 | Thành viên |
| 03 | Huỳnh Trí Vĩ | K184111436 | Thành viên |
| 04 | Lê Thị Kiều Ly | K184111383 | Thành viên |

*TP. Hồ Chí Minh, ngày 20 tháng 05 năm 2021*

# MỤC LỤC

# DANH MỤC HÌNH ẢNH

# LỜI CẢM ƠN

Trong thời gian nỗ lực thực hiện đồ án môn học, nhóm chúng em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và hướng dẫn của Thầy, bạn bè.

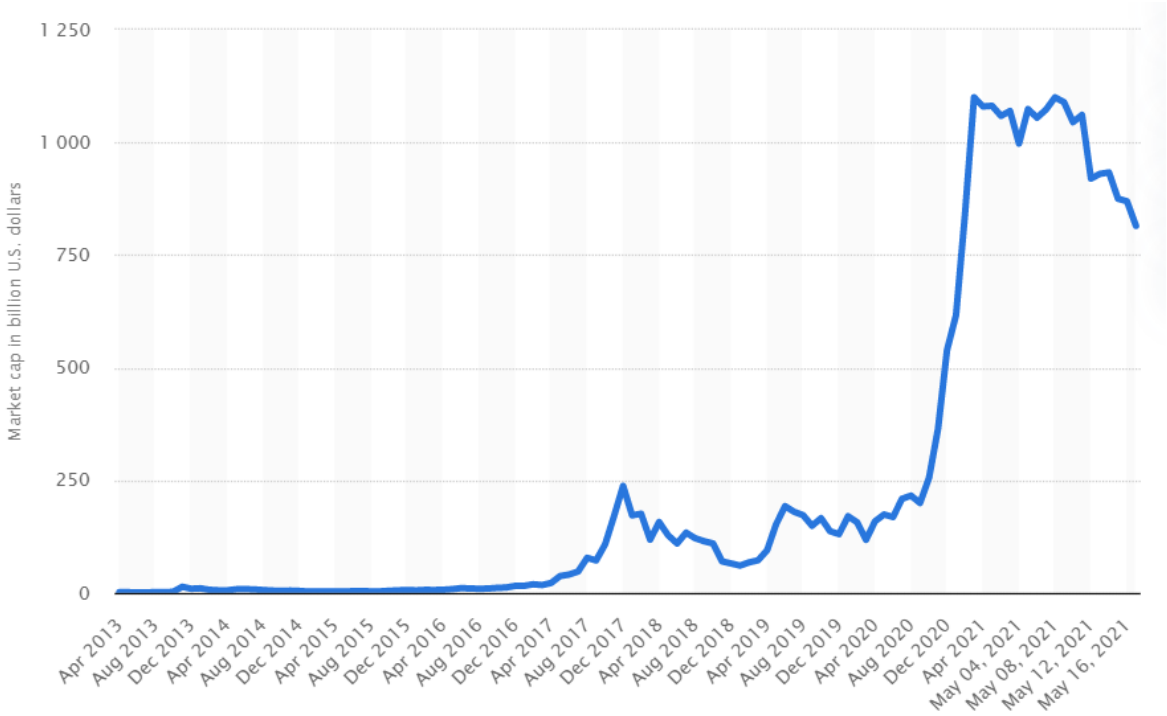
Lời đầu tiên, nhóm của em xin gửi lời cảm ơn chân thành đến Thầy Đặng Nhân Cách – giảng viên hướng dẫn môn Phân tích dữ liệu web – đã trang bị kiến thức nền tảng, hướng dẫn và tạo điều kiện tốt nhất để giúp đỡ nhóm hoàn thành đồ án "Phân tích sự phản ứng của người dùng Twitter đối với tuyên bố của Elon Musk về Bitcoin ngày 13/05/2021" - một cách trọn vẹn nhất trong khả năng của nhóm.

Mặc dù đã có cố gắng hoàn thiện đồ án trong phạm vi và khả năng cho phép nhưng chắc chắn không thể tránh khỏi những thiếu sót. Nhóm mong nhận được góp ý từ Thầy và các bạn để hoàn thiện hơn trong tương lai.

Nhóm xin trân trọng cảm ơn!

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## 1.1. Lý do chọn đề tài




*Hình 1: Biểu đồ vốn hóa thi trường của Bitcoin (USD) từ 04/2012 đến 16/05/2021*

### 1.1.1. Tiềm năng của Bitcoin

Bitcoin có thể nói là một trong những đồng tiền ảo phổ biến nhất, thu hút được nhiều sự quan tâm nhất từ những người chơi bitcoin cho đến những người chỉ dừng lại ở mức độ quan tâm.

Vào tháng 3 năm 2021, vốn hóa thị trường Bitcoin đạt mức cao nhất mọi thời đại. Vốn hóa thị trường trong thời gian ngắn đạt hơn 1.000 tỷ đô la Mỹ vào tháng 5 năm 2021 đô la. Vốn hóa thị trường Bitcoin đã tăng từ khoảng một tỷ đô la Mỹ vào năm 2013 lên gấp vài lần số tiền này kể từ khi nó trở nên phổ biến vào năm 2017.

| Characteristic | Price (in U.S. dollars) | 24h price change (%) |
|---|---|---|
| yearn.finance *(YFI)* | 52,211.36 | -19.26 |
| Wrapped Bitcoin *(WBTC)* | 40,111.65 | -0.12 |
| Bitcoin BEP2 *(BTCB)* | 39,804.08 | -1.09 |
| Bitcoin *(BTC)* | 39,797.01 | -0.89 |
| Maker *(MKR)* | 3,734.75 | -9.34 |
| Ethereum *(ETH)* | 2,712.07 | -8.2 |
| Bitcoin Cash *(BCH)* | 798.73 | -16.81 |
| Compound *(COMP)* | 518.48 | -14.75 |
| Aave *(AAVE)* | 458.61 | -24.69 |
| Kusama *(KSM)* | 420.25 | -20.54 |
| Binance Coin *(BNB)* | 369.73 | -14.98 |
| Monero *(XMR)* | 240.49 | -18.04 |
| Dash *(DASH)* | 216.45 | -18.48 |

*Hình 2: Thống kê 100 đồng tiền ảo có giá trị cao nhất toàn cầu tính đến ngày 20/05/2021*

Theo bảng thống kê trong top 100 đồng tiền ảo có giá trị cao nhất toàn cầu tính đến ngày 20/05/2021 thì có thể thấy giá trị của bitcoin đứng thứ 4 cho thấy sự ảnh hưởng cũng như sự thu hút của nó đối với các nhà đầu tư. Top 1,2 và 3 lần lượt là Yearn.Finance, Wrapped Bitcoin và Bitcoin BEP2.

YFI được xem là loại tiền điện tử, sử dụng và hoạt động dựa trên nền tảng Ethereum. Wrapped Bitcoin là phiên bản token hóa của Bitcoin hoạt động trên nền tảng blockchain của Ethereum. BEP2 là một chuẩn của các token chạy trên Binance Chain.

### 1.1.2. Tác động từ bài tweet của Elon Musk

Elon Musk tỷ phú Mỹ gốc Nam Phi, ông có công rất lớn giúp đồng Bitcoin "thăng hoa" và trở nên nổi bật nhất trong số các đồng tiền điện tử hiện nay. Chỉ mới đầu năm 2021, ông còn chỉ đạo Tesla mua vào 1,5 tỉ USD Bitcoin cho quỹ tài sản của công ty.

Ngày 24-3, ông Musk tuyên bố Tesla sẽ chấp nhận cho khách mua xe điện thanh toán bằng Bitcoin. Tuyên bố đó đã đẩy giá Bitcoin từ mức khoảng 54.000 USD/Bitcoin lên 57.000 USD/Bitcoin.

Tuy nhiên nếu Elon Musk có thể đẩy giá Bitcoin lên, ông lại cũng có thể "dìm" giá nó xuống. Chuyện đó đã xảy ra hôm 13-5, khi ông chủ công ty Tesla tuyên bố sẽ dừng chấp nhận thanh toán bằng Bitcoin vì cho rằng việc "đào" Bitcoin cũng như các giao dịch đồng tiền này đã làm tăng nhiên liệu hóa thạch gây ô nhiễm môi trường.


*Hình 3: Bài tweet ngày 13/05/2021 của Elon Musk*

Tuyên bố của ông làm Bitcoin mất giá hơn 12%, "thổi bay" hơn 360 tỷ USD giá trị vốn hóa của đồng tiền này. Và không chỉ Bitcoin, các đồng tiền điện tử khác cũng chịu "vạ lây" đáng kể. Các đồng tiền khác cũng giảm: Ethereum giảm 10%, trong khi XRP giảm 7%. Trong vòng 24 giờ sau tuyên bố của ông Musk, thống kê của trang Stockhead cho thấy chỉ 15 trong số 100 đồng tiền điện tử có thể tìm lại sắc xanh.

Chỉ một câu nói của một tỷ phú cũng đã làm rung chuyển thị trường Bitcoin nói riêng và thị trường tiền điện tử nói chung. Từ đó dẫn đến người dùng đã có những phản ứng khác nhau đối với câu nói này.

Từ những lý do trên, nhóm quyết định chọn đề tài **"Phân tích sự phản ứng của người dùng Twitter đối với tuyên bố của Elon Musk về Bitcoin ngày 13/05/2021 trên Twitter"**

### 1.2. Mục tiêu đề tài

- Nghiên cứu tổng quan về kiến thức khai phá dữ liệu web và tìm hiểu cách thức và thực hiện phân tích dữ liệu trên mạng xã hội.
- Sử dụng ngôn ngữ Python và những kiến thức đã được học để thu thập các tweets.
- Phân tích dữ liệu thu thập được về phản ứng của người xem đối với Avengers thông qua nội dung các tweets.
- Đưa ra những đánh giá, nhận xét khách quan nhất về tuyên bố của Elon Musk đối với Bitcoin: tác động đến thị trường Bitcoin, thị trường tiền điện tử, sự chuyển dịch dòng tiền Bitcoin,...
- Phân tích dữ liệu thu thập được về phản ứng của người dùng twitter đối với tuyên bố của Elon Musk thông qua nội dung các tweets.
- Làm tài liệu hỗ trợ cho những phân tích về thị trường tiền điện tử về sau.

### 1.3. Đối tượng phân tích

- Các bài đăng trên twitter (tweets) có từ khóa #bitcoin, #elonmusk,...
- Số lượng những account quan tâm đến tuyên bố của Elon Musk có tương tác thông qua số lượng retweets
- Tỷ giá của các đồng tiền điện tử khác như dogecoin, ethereum, XRP,...

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VỀ KHAI PHÁ DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

## 2.1. Tổng quan về khai phá dữ liệu

### 2.1.1. Khái niệm khai phá dữ liệu

Khái niệm khai phá dữ liệu: Data Mining là quá trình khai phá, trích xuất, khai thác và sử dụng những dữ liệu có giá trị tiềm ẩn từ bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu (CSDL), kho dữ liệu, trung tâm dữ liệu… lớn hơn là Big Data dựa trên kĩ thuật như mạng nơ ron, lý thuyết tập thô, tập mờ, biểu diễn tri thức… Đây là một công đoạn trong hoạt động "làm sạch" dữ liệu.

### 2.1.2. Quy trình khai phá dữ liệu

- Quy trình khai phá dữ liệu bao gồm 3 bước:

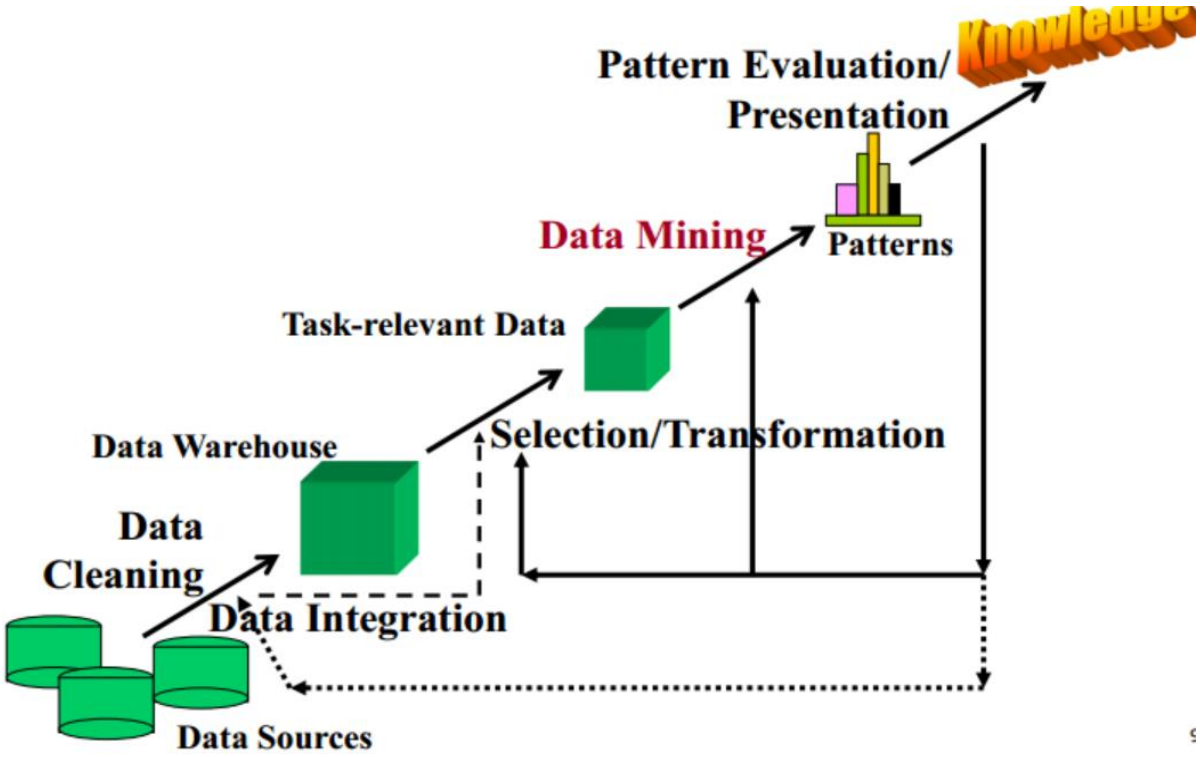
  + Khai phá

  + Trích xuất

  + Khai thác và sử dụng dữ liệu có giá trị trong các CSDL đang có (kể cả Big Data).

- Đây là công đoạn đưa ra dữ liệu đã được "làm sạch".
- Dựa trên kĩ thuật như mạng nơ ron, lí thuyết tập thô, tập mờ, biểu diễn tri thức…
- Quá trình chọn lọc dữ liệu của Data Mining dựa trên các phương pháp: Phân loại (Classification), Phân nhóm (Clustering), Tổng hợp (Summarization), Mô hình ràng buộc (Dependency modeling), Hồi quy (Regression), Dò tìm biến đổi và độ lệch (Change and Deviation Dectection).

- Đây là quá trình khám phá tri thức:




*Hình 4: Quy trình khám phá tri thức*

- Đi từ một chuỗi quy trình: Data Cleaning ➔ Data Integration ➔ Data Selection ➔ Data Transformation ➔ Data Mining ➔ Pattern Evaluation ➔ Knowlegde Presentation.

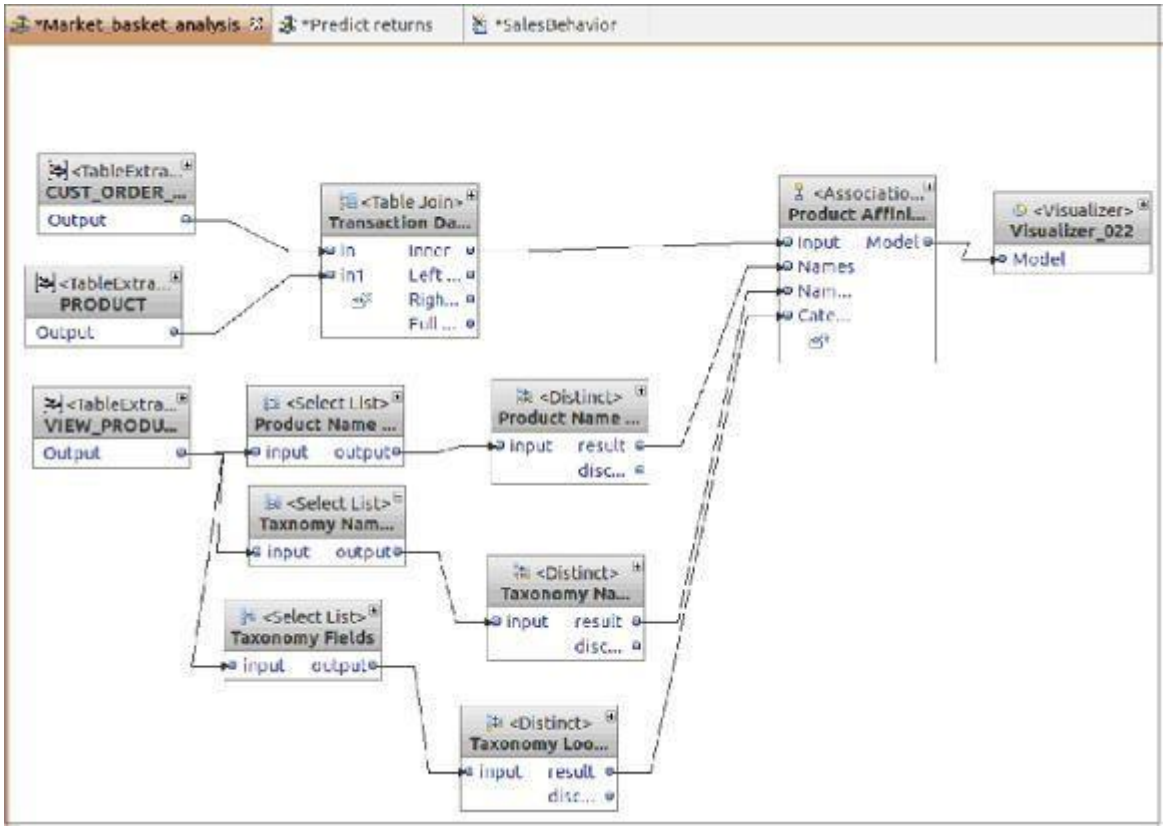
Như vậy có thể thấy Data Mining chính là một "mắt xích" trong quá trình khám phá tri thức. Nếu đang sử dụng nguồn dữ liệu lớn – Big Data thì quá trình khai phá chúng quả không dễ dàng gì bạn có thể mất nhiều thời gian hơn, nhiều nhân lực, chi phí cho hoạt động Data Mining trong Big Data.

### 2.1.3. Các kỹ thuật và phương pháp trong khai phá và xử lý dữ liệu

**** Sự kết hợp***

- Sự kết hợp (hay mối quan hệ) có lẽ là kỹ thuật khai phá dữ liệu được biết đến nhiều hơn, hầu như quen thuộc và đơn giản. Ở đây, thực hiện một sự tương quan đơn giản giữa hai hoặc nhiều mục, thường cùng kiểu để nhận biết các mẫu. Ví dụ, khi theo dõi thói quen mua hàng của người dân, có thể nhận biết rằng một khách hàng luôn mua kem khi họ mua dâu tây, nên hệ thống có thể đề xuất rằng lần tới khi họ mua dâu tây, họ cũng có thể muốn mua kem.

- Việc xây dựng các công cụ khai phá dữ liệu dựa trên sự kết hợp hay mối quan hệ có thể thực hiện đơn giản bằng các công cụ khác nhau. Ví dụ, trong InfoSphere Warehouse một trình hướng dẫn đưa ra các cấu hình của một luồng thông tin được sử dụng kết hợp bằng cách xem xét thông tin nguồn đầu vào của cơ sở dữ liệu, thông tin về cơ sở ra quyết định và thông tin đầu ra của bạn. Hình 2 cho thấy một ví dụ mẫu của cơ sở dữ liệu.




*Hình 5: Ví dụ mẫu về cơ sở dữ liệu*

### ** Sự phân loại*

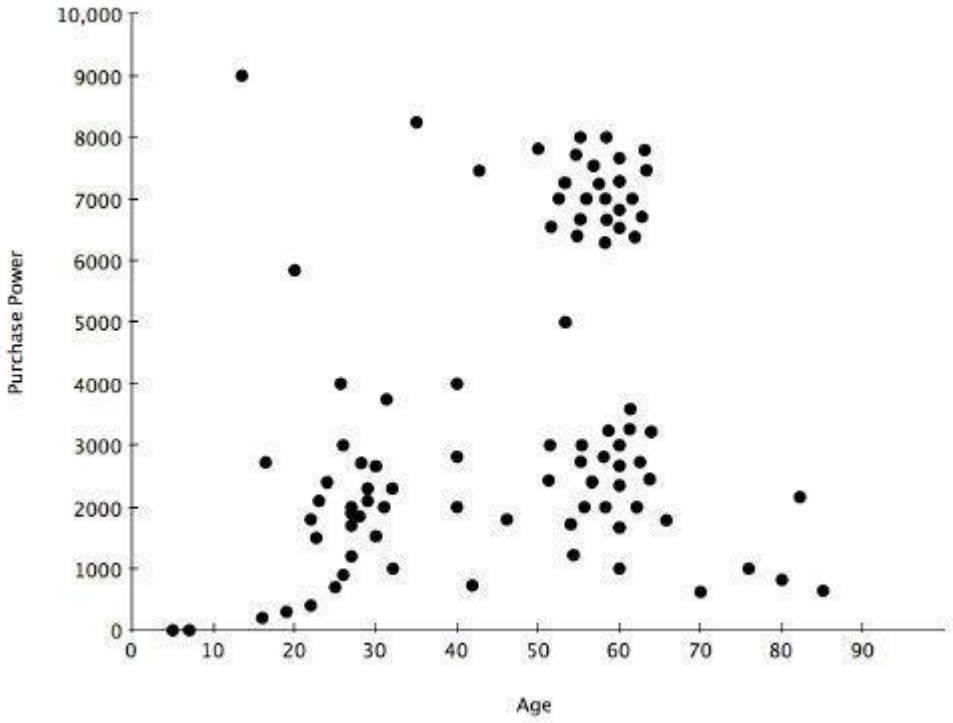
- Có thể sử dụng sự phân loại để xây dựng một ý tưởng về kiểu khách hàng, kiểu mặt hàng hoặc kiểu đối tượng bằng cách mô tả nhiều thuộc tính để nhận biết một lớp cụ thể. Ví dụ, ta có thể dễ dàng phân loại các xe ô tô thành các kiểu xe khác nhau (xe mui kín, 4x4, xe có thể bỏ mui) bằng cách xác định các thuộc tính khác nhau (số chỗ ngồi, hình dạng xe, các bánh xe điều khiển). Ta có thể áp dụng các nguyên tắc tương tự ấy cho các khách hàng, ví dụ bằng cách phân loại khách hàng theo độ tuổi và nhóm xã hội.

Hơn nữa, có thể sử dụng việc phân loại như một nguồn cấp, hoặc như là kết quả của các kỹ thuật khác. Ví dụ, có thể sử dụng các cây quyết định để xác định một cách phân loại. - Việc phân cụm sẽ cho phép người dùng sử dụng các thuộc tính chung theo các cách phân loại khác nhau để nhận biết các cụm.

**** Việc phân cụm (Clustering)***

- Bằng cách xem xét một hay nhiều thuộc tính hoặc các lớp, ta có thể nhóm các phần dữ liệu riêng lẻ với nhau để tạo thành một quan điểm cấu trúc. Ở mức đơn giản, việc phân cụm đang sử dụng một hoặc nhiều thuộc tính làm cơ sở để nhận ra một nhóm các kết quả tương quan. Việc phân cụm giúp để nhận biết các thông tin khác nhau vì nó tương quan với các ví dụ khác, nên có thể thấy ở đâu có những điểm tương đồng và các phạm vi phù hợp.

- Trong hình 3 bên dưới, là một ví dụ mẫu về dữ liệu kinh doanh so sánh tuổi của khách hàng với quy mô bán hàng. Thật hợp lý khi thấy rằng những người ở độ tuổi hai mươi (trước khi kết hôn và còn nhỏ), ở độ tuổi năm mươi và sáu mươi (khi không còn con cái ở nhà), có nhiều tiền tiêu hơn.




*Hình 6: Đồ thị phân cụm*

- Trong ví dụ này, chúng ta có thể nhận ra hai cụm, một cụm xung quanh nhóm 2.000 Đô la Mỹ/ 20-30 tuổi và một cụm ở nhóm 7.000-8.000 Đô la Mỹ/ 50-65 tuổi.

- Việc vẽ đồ thị phân cụm theo cách này là một ví dụ đơn giản về cái gọi là nhận ra sự lân cận gần nhất. Ta có thể nhận ra các khách hàng riêng lẻ bằng sự gần gũi theo nghĩa đen của họ với nhau trên đồ thị. Có nhiều khả năng là các khách hàng trong cùng một cụm cũng dùng chung các thuộc tính khác và có thể sử dụng sự mong đợi đó để giúp hướng dẫn, phân loại và nếu không thì phân tích những người khác trong tập hợp dữ liệu.

**** Dự báo***

- Dự báo là một chủ đề rộng và đi từ dự báo về lỗi của các thành phần hay máy móc đến việc nhận ra sự gian lận và thậm chí là cả dự báo về lợi nhuận của công ty nữa. Được sử dụng kết hợp với các kỹ thuật khai phá dữ liệu khác, dự báo gồm có việc phân tích các xu hướng, phân loại, so khớp mẫu và mối quan hệ. Bằng cách phân tích các sự kiện hoặc các cá thể trong quá khứ, ta có thể đưa ra một dự báo về một sự kiện.

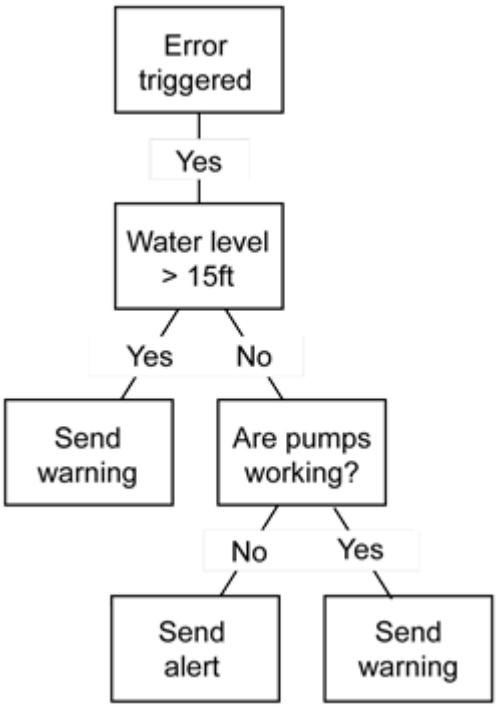
- Khi sử dụng quyền hạn thẻ tín dụng, chẳng hạn, có thể kết hợp phân tích cây quyết định của các giao dịch riêng lẻ trong quá khứ với việc phân loại và các sự so khớp mẫu lịch sử để nhận biết liệu một giao dịch có gian lận hay không. Rất có thể là việc thực hiện một sự so khớp giữa việc mua vé các chuyến bay đến Mỹ và các giao dịch tại Mỹ cho thấy giao dịch này hợp lệ.

**** Các mẫu tuần tự***

- Thường được sử dụng trên các dữ liệu dài hạn, các mẫu tuần tự là một phương pháp có ích để nhận biết các xu hướng hay các sự xuất hiện thường xuyên của các sự kiện tương tự. Ví dụ, với dữ liệu khách hàng, có thể nhận ra rằng các khách hàng cùng nhau mua một bộ sưu tập riêng lẻ về các sản phẩm tại nhiều thời điểm khác nhau trong năm. - Trong một ứng dụng giỏ hàng, ta có thể sử dụng thông tin này để tự động đề xuất rằng một số mặt hàng nào đó được thêm vào một giỏ hàng dựa trên tần suất và lịch sử mua hàng trong quá khứ của các khách hàng.

**** Các cây quyết định***

- Liên quan đến hầu hết các kỹ thuật khác (chủ yếu là phân loại và dự báo), cây quyết định có thể được sử dụng hoặc như là một phần trong các tiêu chí lựa chọn hoặc để hỗ trợ việc sử dụng và lựa chọn dữ liệu cụ thể bên trong cấu trúc tổng thể. Trong cây quyết định, bắt đầu bằng một câu hỏi đơn giản có hai câu trả lời (hoặc đôi khi có nhiều câu trả lời hơn). Mỗi câu trả lời lại dẫn đến thêm một câu hỏi nữa để giúp phân loại hay nhận biết dữ liệu sao cho có thể phân loại dữ liệu hoặc sao cho có thể thực hiện dự báo trên cơ sở mỗi câu trả lời.




*Hình 7: Mô hình cây quyết định*

- Các cây quyết định thường được sử dụng cùng với các hệ thống phân loại liên quan đến thông tin có kiểu thuộc tính và với các hệ thống dự báo, nơi các dự báo khác nhau có thể dựa trên kinh nghiệm lịch sử trong quá khứ để giúp hướng dẫn cấu trúc của cây quyết định và kết quả đầu ra.

**** Các tổ hợp***

Trong thực tế, thật hiếm khi người ta sẽ sử dụng một kỹ thuật trong số những kỹ thuật riêng biệt này. Việc phân loại và phân cụm là những kỹ thuật giống nhau. Nhờ sử dụng việc phân cụm để nhận ra các thông tin lân cận gần nhất, ta có thể tiếp tục tinh chỉnh việc phân loại của mình. Thông thường, người ta sử dụng các cây quyết định để giúp xây dựng và nhận ra các loại mà họ có thể theo dõi chúng trong một thời gian dài để nhận biết các trình tự và các mẫu.

**** Xử lý (bộ nhớ) dài hạn***

- Trong tất cả các phương pháp cốt lõi, thường có lý do để ghi lại thông tin và tìm hiểu từ thông tin. Trong một số kỹ thuật, việc này hoàn toàn rõ ràng. Ví dụ, với việc tìm hiểu các mẫu tuần tự và dự báo, việc cần thiết là xem xét lại dữ liệu từ nhiều nguồn và nhiều cá thể thông tin để xây dựng một mẫu.

- Trong một số kỹ thuật khác, quá trình này có thể rõ ràng hơn. Các cây quyết định ít khi được xây dựng một lần và không bao giờ được coi nhẹ. Khi nhận biết thông tin mới, các sự kiện và các điểm dữ liệu, có thể cần xây dựng thêm các nhánh hoặc thậm chí toàn bộ các cây mới, để đương đầu với các thông tin bổ sung.

- Người ta có thể tự động hoá một số bước của quá trình này. Ví dụ, việc xây dựng một mô hình dự báo để nhận biết sự gian lận thẻ tín dụng là xây dựng các xác suất để bạn có thể sử dụng cho giao dịch hiện tại và sau đó cập nhật mô hình đó với các giao dịch mới (đã được phê duyệt). Rồi thông tin này được ghi lại sao cho có thể đưa ra quyết định một cách nhanh chóng trong lần tới.

### 2.1.4. Vai trò và lợi ích của khai phá dữ liệu

Có rất nhiều lợi ích của việc khai thác dữ liệu. Ví dụ:

- Trong lĩnh vực tài chính ngân hàng, khai thác dữ liệu được sử dụng để tạo ra các mô hình rủi ro chính xác cho các khoản vay và thế chấp. Họ cũng rất hữu ích khi phát hiện các giao dịch gian lận.
- Trong tiếp thị, kỹ thuật khai thác dữ liệu được sử dụng để cải thiện chuyển đổi, tăng sự hài lòng của khách hàng và tạo ra các chiến dịch quảng cáo được nhắm mục tiêu. Họ thậm chí có thể được sử dụng khi phân tích nhu cầu trên thị trường và tìm ra ý tưởng cho các dòng sản phẩm hoàn toàn mới. Điều này được thực hiện bằng cách xem dữ liệu khách hàng và bán hàng lịch sử và tạo ra các mô hình dự đoán mạnh mẽ.
- Các cửa hàng bán lẻ sử dụng các thói quen/chi tiết mua sắm của khách hàng để tối ưu hóa cách bố trí các cửa hàng của họ nhằm nâng cao trải nghiệm của khách hàng và tăng lợi nhuận.
- Các cơ quan quản lý thuế sử dụng các kỹ thuật khai thác dữ liệu để phát hiện các giao dịch gian lận và khai thuế đáng ngờ hoặc các tài liệu kinh doanh khác.
- Trong sản xuất, phát hiện dữ liệu được sử dụng để cải thiện an toàn sản phẩm, khả năng sử dụng và thoải mái.
- Nhân sự: Khai phá dữ liệu từ hồ sơ của ứng viên, từ đó cung cấp cái nhìn toàn diện về ứng viên. Xác định kết quả phù hợp nhất cho từng vai trò bằng cách sử dụng phân tích dữ liệu để đánh giá trình độ, kinh nghiệm, kỹ năng, chứng chỉ và vị trí công việc đã đảm nhiệm trước đây.

## 2.2. Phân tích và khai phá dữ liệu trên mạng xã hội trên Twitter

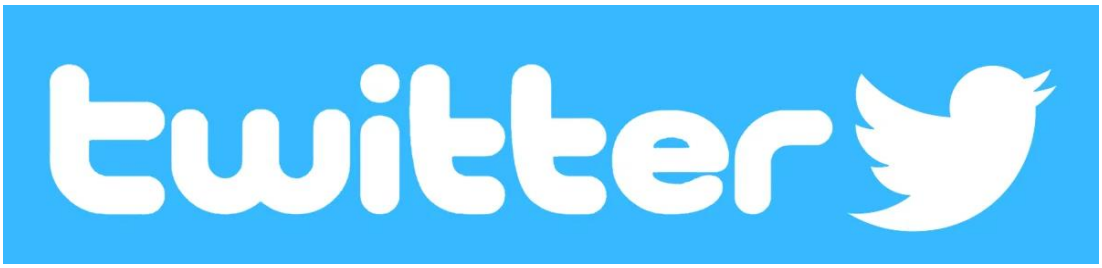
### 2.2.1. Khái niệm khai phá dữ liệu trên mạng xã hội


*Hình 8: Mạng xã hội ảo*

- Mạng xã hội hay còn được một số người gọi là mạng xã hội ảo. Đây là nơi kết nối mọi người không phân biệt quốc gia, màu da hay sở thích, nghề nghiệp… với nhau. Những người có mối quan hệ ngoài đời thực cũng có thể tiếp tục kết nối bạn bè với nhau trên mạng xã hội. Để truy cập mạng xã hội, bạn có thể thực hiện rất dễ dàng thông qua các thiết bị có kết nối internet như máy tính, điện thoại, iPad… Qua đó, Dữ liệu mạng xã hội chính là những thông tin mà người dùng mạng xã hội chia sẻ công khai, bao gồm các siêu dữ liệu như vị trí của người dùng, ngôn ngữ sử dụng, dữ liệu tiểu sử và các liên kết được chia sẻ.
- Có nhiều loại dữ liệu mạng xã hội, bao gồm các tweet từ Twitter, bài đăng trên Facebook hay ghim trên Pinterest. Facebook for Business và Twitter Ads là hai chương trình giúp nhà quảng cáo sử dụng dữ liệu mạng xã hội nhắm đến những người dùng có khả năng quan tâm đến quảng cáo của họ.
- Về cơ bản, khai phá dữ liệu là xử lý dữ liệu là nhận biết các mẫu và các xu hướng trong thông tin đó để người nghiên cứu có thể quyết định hoặc đánh giá. Các nguyên tắc khai phá dữ liệu đã được dùng nhiều năm, nhưng với sự ra đời của big data (dữ liệu lớn), nó lại càng phổ biến hơn.
- Chính vì thế, Khai phá dữ liệu mạng xã hội là một trong những vấn đề nổi bật và được giới khoa học quan tâm nhất hiện nay. Các đề tài nghiên cứu về dữ liệu mạng xã hội để ứng dụng vào nhiều lĩnh vực khác nhau như: tư vấn sản phẩm, dịch vụ tài chính, sự kiện xã hội, bầu cử chính trị, dịch vụ y tế, ...Với sự phát triển nhanh chóng

về số người sử dụng trên toàn thế giới, mạng xã hội trực tuyến như một mô hình thu nhỏ của thế giới thực. Do đó, mạng xã hội trở thành nơi cất giữ thông tin và các mối quan hệ giữa các cá nhân, doanh nghiệp, … Những thông tin này tạo thành "đám mây tri thức".Việc tìm hiểu và khai thác hiệu quả những thông tin trên mạng xã hội sẽ tạo tiền đề cho các ứng dụng khác như:tiếp thị trực tuyến,hệ thống tìm kiếm thông tin, hệ thống tư vấn, an ninh trên mạng xã hội điều tra tội phạm, dự đoán sự phát triển của mạng xã hội, ... Vì vậy, việc khai thác những thông tin trên mạng xã hội để áp dụng vào thực tiễn ngày càng trở nên quan trọng hơn bao giờ hết.

### 2.2.2. Tổng quan về twitter




- Twitter là một trang mạng xã hội cho phép người sử dụng có thể viết và đọc nội dung có độ dài giới hạn và tải hình ảnh lên. Nếu bạn là người hay nhắn tin điện thoại thì bạn sẽ biết rõ giới hạn 160 ký tự của tin nhắn SMS. Twitter cũng gần giống như thế thậm chí số ký tự cho phép còn ít hơn chỉ có 140 ký tự mà thôi.
- Biểu tượng: Twitter Con chim màu xanh là biểu tượng đặc trưng của Twitter hiện nay. Facebook, Twitter và Pinterest là một trong những mạng xã hội phổ biến nhất hiện nay.
- Xuất xứ: Twitter đã được tạo ra vào tháng 3 năm 2006 bởi Jack Dorsey, Evan Williams, Biz Stone và Noah Glass, nó bắt đầu hoạt động vào tháng 7 năm 2006. Twitter đặt trụ sở chính tại San Francisco và đã có hơn 25 văn phòng trên toàn thế giới. Twitter đã có hơn 500 triệu người dùng, trong đó có hơn 302 triệu người hoạt động thường xuyên (năm 2015). Twitter cũng được xem như là SMS trên mạng Internet.

### 2.2.3. Thu thập dữ liệu từ twitter

- Twitter là một nền tảng truyền thông xã hội, nơi 328 triệu người dùng hoạt động hàng tháng trên tiểu blog (chia sẻ cập nhật 280 ký tự) với những người theo dõi của họ. Nó là sự kết hợp giữa nhắn tin tức thì và viết blog hoặc nhắn tin xã hội, nhưng nó cũng rất quan trọng đối với báo cáo tin tức, quảng bá sự kiện, tiếp thị và kinh doanh.

- Tweet thường có dạng dạng blog thời gian thực, mang tính xã hội cao cho phép người dùng đăng các cập nhật trạng thái ngắn, xuất hiện trên các dòng thời gian. Nó có thể bao gồm một hoặc nhiều thực thể trong 140 ký tự nội dung cập nhật trạng thái của người dùng trong thế giới thực. Tweet cần thiết để sử dụng hiệu quả API của twitter.

- Ngoài nội dung văn bản của chính một tweet, các tweet đi kèm với hai phần siêu dữ liệu bổ sung - hashtag, thẻ bắt đầu bằng #, URL và phương tiện có thể được liên kết với một tweet và địa điểm là các vị trí trong thế giới thực.

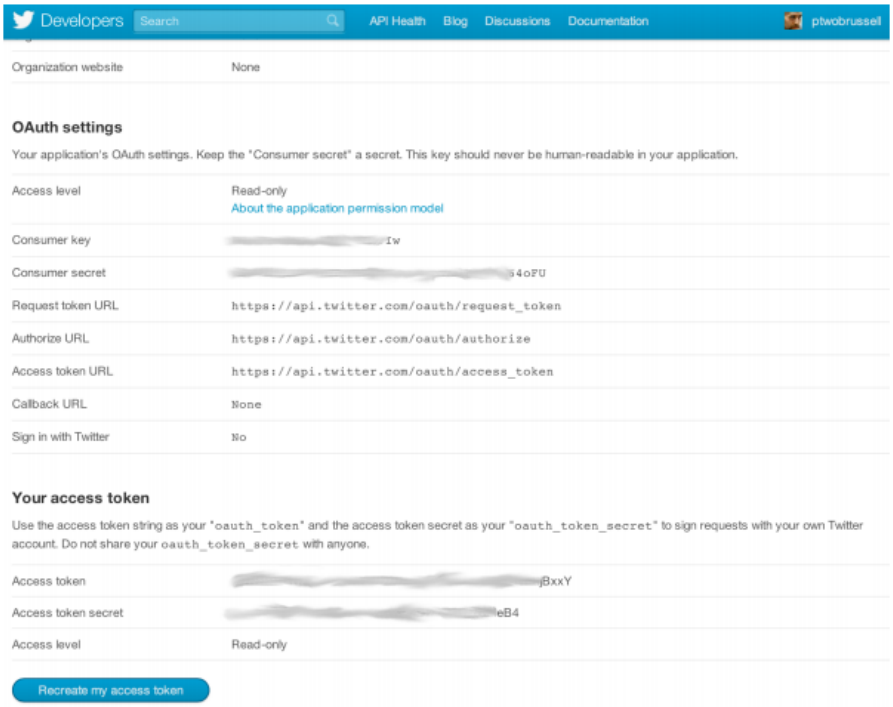

*Hình 9: Tweets trong Twitter*

- Mức độ bùng nổ công khai của tất cả các tweet đã được biết là đạt đỉnh điểm hàng trăm nghìn tweet mỗi phút trong các sự kiện có sự quan tâm đặc biệt rộng rãi, chẳng hạn như các cuộc tranh luận tổng thống.

- Quy trình phân tích dữ liệu twitter:

+ Nếu bạn chưa có tài khoản Twitter, điều đầu tiên bạn cần làm là tạo một tài khoản. Sau đó, truy cập apps.twitter.com để tạo một ứng dụng cho phép bạn thu thập dữ liệu Twitter. Ứng dụng bạn tạo sẽ kết nối với giao diện chương trình ứng dụng Twitter (API). Có hai API mà bạn có thể sử dụng để thu thập các tweet. Nếu bạn muốn thu thập một lần các tweet, thì bạn sẽ sử dụng API REST. Nếu bạn muốn thu thập liên tục các tweet trong một khoảng thời gian cụ thể, bạn sẽ sử dụng API phát trực tuyến.

+ Trước khi có thể thực hiện bất kỳ yêu cầu API nào đối với Twitter, bạn cần tạo một ứng dụng tại https://dev.twitter.com/apps. Tạo một ứng dụng là cách tiêu chuẩn để những người tham gia phát triển có được quyền truy cập API và để Twitter giám sát và tương tác với các nhà phát triển biểu mẫu cố định bên thứ ba khi cần thiết.




*Hình 10: Tạo một ứng dụng twitter mới*

```
import twitter

# Go to https://developer.twitter.com/en/apps to create an app and get values
# for these credentials, which you'll need to provide in place of these
# empty string values that are defined as placeholders.
# See https://developer.twitter.com/en/docs/basics/authentication/overview/oauth
# for more information on Twitter's OAuth implementation.

CONSUMER_KEY = ''
CONSUMER_SECRET = ''
OAUTH_TOKEN = ''
OAUTH_TOKEN_SECRET = ''

auth = twitter.oauth.OAuth(OAUTH_TOKEN, OAUTH_TOKEN_SECRET,
                           CONSUMER_KEY, CONSUMER_SECRET)

twitter_api = twitter.Twitter(auth=auth)

# Nothing to see by displaying twitter_api except that it's now a
# defined variable

print(twitter_api)
```

*Hình 11: Cấp phép ứng dụng truy cập dữ liệu tài khoản Twitter*

- Bước tiếp theo là mở python và sẵn sàng viết mã.

Tóm lại, Với kết nối API được ủy quyền tại chỗ, bây giờ bạn có thể đưa ra một yêu cầu. Phân tích Twitter về các chủ đề hiện đang thịnh hành trên toàn thế giới hoặc các vấn đề bạn đang quan tâm.

### 2.2.4. Các công cụ sử dụng để khai phá dữ liệu twitter

- Sử dụng truy xuất dữ liệu trên Twitter qua các API dạng webservice, cho phép truy cập và sử dụng toàn bộ dữ liệu. Hiện nay đã có rất nhiều công cụ có thể sử dụng cho mục đích khai phá dữ liệu, có thể kể đến như ANGOSSKnowledgeSTUDIO, SQL Server, R, Clementine, Python.
- Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Nó dễ dàng để tìm hiểu và đang nổi lên như một trong những ngôn ngữ lập trình nhập môn tốt nhất cho người lần đầu tiếp xúc với ngôn ngữ lập trình. Python hoàn toàn tạo kiểu động và sử dụng cơ chế cấp phát bộ nhớ tự động. Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng. Nhóm sẽ sử dụng trực tiếp trên Google Colab, là môi trường đắc lực cho việc chạy ngôn ngữ Python để thực hiện đề tài.

# CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU, KIỂM ĐỊNH VÀ ĐÁNH GIÁ

## 3.1. Nguồn dữ liệu

Sau khi tìm hiểu về đề tài nhóm đã tiến hành thu thập dữ liệu từ thu thập từ mạng xã hội Twitter, đây là trang mạng xã hội phổ biến ở có hầu hết ở các quốc gia và có xu hướng cập nhật các trend hot nhanh nhất thế giới. Dữ liệu sẽ được lấy từ các tweets với hashtag là #bitcoin sau đó triển khai lọc lại thông tin cần thiết và chỉ lưu lấy các tweets có nội dung là tiếng anh để dễ dàng dùng thư viện có sẵn để phân tích.

## 3.2. Thu thập các tweets, retweets có từ khóa là Bitcoin

Một trong những mục phổ biến giữa các nhóm chủ đề thịnh hành là thẻ bằng đầu bằng "#". Sử dụng từ khóa bitcoin làm cơ sở truy vấn tìm kiếm để tìm và nạp một số tweet để phân tích. Trước tiên, dùng lệnh có sẵn thu thập những tweets có từ khóa liên quan đến bitcoin cụ thể là 200 tweets để phân tích. Sử dụng GET statuses/sample trả lại độ dài của các tweet.

```
import json

# Set this variable to a trending topic,
# or anything else for that matter. The example query below
# was a trending topic when this content was being developed
# and is used throughout the remainder of this chapter.

q = '#bitcoin'

count = 100

# Import unquote to prevent url encoding errors in next_results
from urllib.parse import unquote

# See https://dev.twitter.com/rest/reference/get/search/tweets

search_results = twitter_api.search.tweets(q=q, count=count)

statuses = search_results['statuses']


# Iterate through 5 more batches of results by following the cursor
for _ in range(200):
    print('Length of statuses', len(statuses))
    try:
        next_results = search_results['search_metadata']['next_results']
    except KeyError as e: # No more results when next_results doesn't exist
        break
```

*Hình 13: Câu lệnh lấy tweet có liên quan tới từ khóa bitcoin (1)*

```
    # Create a dictionary from next_results, which has the following form:
    # ?max_id=847960489447628799&q=%23RIPSelena&count=100&include_entities=1
    kwargs = dict([ kv.split('=') for kv in unquote(next_results[1:]).split("&") ])

    search_results = twitter_api.search.tweets(**kwargs)
    statuses += search_results['statuses']

# Show one sample search result by slicing the list...
print(json.dumps(statuses[0], indent=1))
```

*Hình 12: Câu lệnh lấy tweet có liên quan tới từ khóa bitcoin (2)*

Trích xuất các thực thể trong bao gồm: date, text, favorite count, retweet count. Trong đó:

- ['favourite_count']: thể hiện sự "sự thú vị" của một tweet có sẵn
- t ['retweet_count']: Trả về số lần nó được đánh dấu trang hoặc đã tweet lại

tương ứng. Nếu một tweet đã được tweet lại, trường t ['retweet_status'] sẽ cung cấp chi tiết quan trọng về chính tweet gốc và tác giả của nó.

```
[4]  import os
     blog_posts = []
     for i in range(200):
       if statuses[i]['lang'] == 'en':
         blog_posts.append({'date': statuses[i]['created_at'], 'text'
                            : statuses[i]['text'], 'favorite_count': statuses[i]['favorite_count'], 'retweet_count': statuses[i]['retweet_count']})

     out_file = os.path.join('drive/My Drive/bitcoin.json')
     f = open(out_file, 'w+')
     f.write(json.dumps(blog_posts, indent=1))
     f.close()
```

*Hình 14: Câu lệnh lưu trữ vào file json*


```
[
 {
  "date": "Tue May 18 01:20:11 +0000 2021",
  "text": "Analyse de prix 5/17: BTC, ETH, BNB, ADA, DOGE, XRP, DOT, BCH, LTC, UNI https://t.co/tTdzCFDpPR | #Bitcoin",
  "favorite_count": 0,
  "retweet_count": 0
 },
 {
  "date": "Tue May 18 01:20:10 +0000 2021",
  "text": "RT @ComPro01: ComPro.Finace @Airdrop is live! \n$copr Reward for Tasks: Up to 150 \n #Airdrop #Bitcoin #Ethereum
#bsc #HECOchain #copr\n\ud83d\udd36 Join\u2026",
  "favorite_count": 0,
  "retweet_count": 608
 },
```


*Hình 15: Định dạng trong file bitcoin.json*

Ở đây, trong file bitcoin.json thể hiện rất rõ ràng và trực quan các thuộc tính trong từng tweet.

### 3.3. Phân tích tổng quan nội dung của các tweet thu được

#### 3.3.1. Thống kê 20 words (các từ xuất hiện nhiều), screen name (có thể hiểu như thông tin cá nhân của người dùng) và hashtag phổ biến nhất

- Trích xuất các thực thể tweet, tiến hành tách các thực thể và văn bản của một số tweet thành một cấu trúc dữ liệu thuận tiện để kiểm tra thêm.
- Trích xuất bao gồm: text, screen name và thẻ bắt đầu bằng # từ các tweet được thu thập. Trong khi chỉ trích xuất 5 mục đầu tiên của mỗi danh sách chưa được xếp hạng để có cảm nhận về dữ liệu, bây giờ tiến hành xem xét kỹ hơn những gì có trong dữ liệu bằng cách tính toán phân phối tần suất và xem 10 mục hàng đầu trong mỗi danh sách “tính toán phân bố tần số dưới dạng danh sách các thuật ngữ được xếp hạng”.

Kết luận: qua số liệu thu thập trên ta thấy các cụm từ: RT, Elonmusk, bitcoin, cryptocurrency và hashtag #bitcoin xuất hiện với tần suất cao.

```
[5]  status_texts = [ status['text']
                      for status in statuses ]

     screen_names = [ user_mention['screen_name']
                      for status in statuses
                          for user_mention in status['entities']['user_mentions'] ]

     hashtags = [ hashtag['text']
                  for status in statuses
                      for hashtag in status['entities']['hashtags'] ]

     # Compute a collection of all words from all tweets
     words = [ w
               for t in status_texts
                   for w in t.split() ]

     from collections import Counter

     for item in [words, screen_names, hashtags]:
         c = Counter(item)
         print(c.most_common()[:20]) # top 10
         print()
```

```
[('RT', 11729), ('#Bitcoin', 7829), ('to', 6436), ('the', 5785), ('and', 3813), ('is', 34

[('elonmusk', 1039), ('flurbnb', 971), ('michael_saylor', 700), ('PeterSchiff', 499), ('D

[('Bitcoin', 8476), ('bitcoin', 3025), ('cryptocurrency', 1234), ('dogecoin', 1188), ('BT
```

*Hình 16: câu lệnh thống kê 20 word, screen name, hashtag phổ biến nhất*

Đưa các dữ liệu thống kê trên vào bảng cho dễ nhìn, dễ kiểm soát, dùng lệnh cho thuận tiện. Kết quả của phân bố tần suất là một bản đồ của các cặp khóa/giá trị tương ứng với các cụm từ và tần suất của chúng. Điều này giúp việc xem xét kết quả dễ dàng hơn một chút bằng cách tạo ra một định dạng bảng.

```
[6]  from prettytable import PrettyTable

     for label, data in (('Word', words),
                         ('Screen Name', screen_names),
                         ('Hashtag', hashtags)):
         pt = PrettyTable(field_names=[label, 'Count'])
         c = Counter(data)
         [ pt.add_row(kv) for kv in c.most_common()[:100] ]
         pt.align[label], pt.align['Count'] = 'l', 'r' # Set column alignment
         print(pt)
```

*Hình 17: Code đưa 20 word, Screen name, hashtag vào bảng*

| Word | Count |
|---|---|
| RT | 3499 |
| #Bitcoin | 2442 |
| the | 1689 |
| to | 1460 |
| in | 1111 |
| a | 1091 |
| is | 1043 |
| of | 940 |
| and | 938 |
| #bitcoin | 857 |
| you | 828 |
| for | 749 |
| - | 702 |
| I | 624 |
| on | 510 |
| be | 407 |
| do | 405 |
| this | 398 |
| that | 373 |
| Bitcoin | 360 |

*Hình 18: Bảng kết quả thống kê 20 word phổ biến nhất*

| Screen Name | Count |
|---|---|
| cryptokenspace | 168 |
| elonmusk | 150 |
| DocumentingBTC | 140 |
| flurbnb | 132 |
| BTC_Archive | 119 |
| maxkeiser | 105 |
| hackkkeedd | 100 |
| michael_saylor | 93 |
| AltcoinDailyio | 89 |
| AirdropStario | 86 |
| TheCryptoLark | 72 |
| Hut8Mining | 67 |
| HudCrypto | 62 |
| SquawkCNBC | 58 |
| SenLummis | 55 |
| BigwinOfficials | 54 |
| SJosephBurns | 52 |
| airdropinspect | 46 |
| Nathan_Combs_ | 45 |
| BitcoinMagazine | 44 |

*Hình 19: Bảng kết quả thống kê 20 Screen Name phổ biến nhất*

| Hashtag | Count |
|---|---|
| Bitcoin | 2741 |
| bitcoin | 1045 |
| cryptocurrency | 326 |
| BSC | 256 |
| Binance | 244 |
| BTC | 239 |
| Crypto | 234 |
| btc | 224 |
| BNB | 211 |
| crypto | 201 |
| ethereum | 193 |
| cryptocurrencies | 138 |
| dogecoin | 134 |
| cryptocurrencyNews | 102 |
| NFA | 100 |
| blockchain | 99 |
| Ethereum | 74 |
| ETH | 71 |
| DeFi | 57 |
| twitter | 50 |

*Hình 20: Bảng kết quả thống kê 20 Hashtag phổ biến nhất*

- Sau khi thống kê, ta tiến hành thể hiện một số giá trị của các thuộc tính có dữ liệu dạng số liệu. Từ đó, bắt đầu phân tích các số liệu nhận được và đưa ra kết luận.

- Dựa vào bảng trên, chúng ta rút ra được một số kết luận như sau:

+ Screen name Elonmusk đứng cao trong danh sách các thực thể cho mẫu dữ liệu tương ứng với 1039, điều đó có nghĩa là phần lớn các tweet liên quan đến bitcoin được ảnh hưởng bởi Elonmusk - người sáng lập Tesla. Bên cạnh, số lượng từ khóa tesla chiếm 155. Qua đó, nhận thấy khách hàng có xu hướng quan tâm nhiều đến các bài viết liên quan đến Elon Musk, bitcoin hiện nay

+ Tại đây, đối với 20 Word từ xuất hiện nhiều nhất chính là RT (Retweet), đơn thuần là trong 200 tweets mà lệnh thu được thì kết quả về số retweet rất cao khoảng gần 11729 RT, gấp hơn 59 lần số tweet. Có thể thấy các tweet về Bitcoin đang nhận

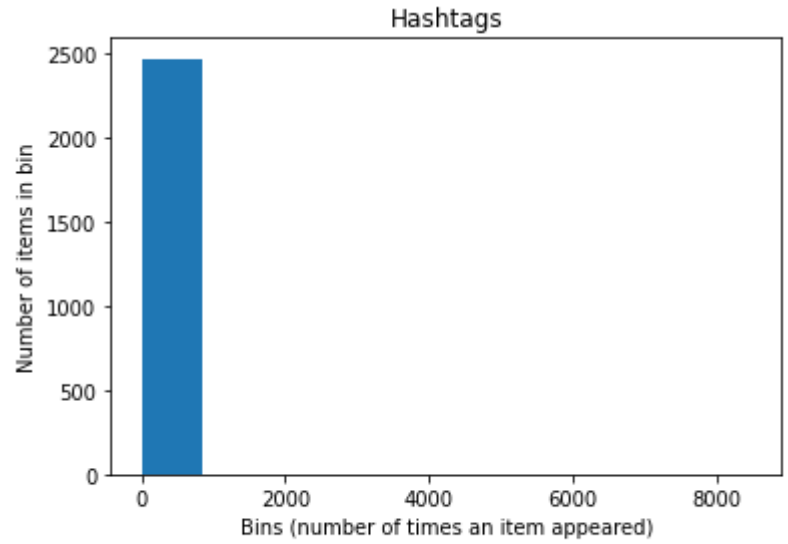
được phần lớn sự quan tâm của người dùng Twitter trên thế giới. Đồng thời hashtag "#RT" là một mã thông báo rất phổ biến, ngụ ý rằng có là một số lượng retweet đáng kể.

### 3.3.2. Vẽ biểu đồ cho các Word và Screen name

- Nhìn lại dữ liệu dạng bảng tương ứng và xem xét rằng có một số lượng lớn các từ, tên màn hình hoặc thẻ bắt đầu bằng "#" có tần suất thấp và xuất hiện ít lần trong văn bản. Tuy nhiên, khi chúng ta kết hợp tất cả các thuật ngữ có tần suất thấp và gộp chung lại với nhau thành một phạm vi phạm vi "có tần suất từ 1 đến 1000" thì tỉ lệ chúng rất cao. Từ đó dữ liệu được tiến hành trực quan hóa thành dạng bảng và đặt lên cho các cột.

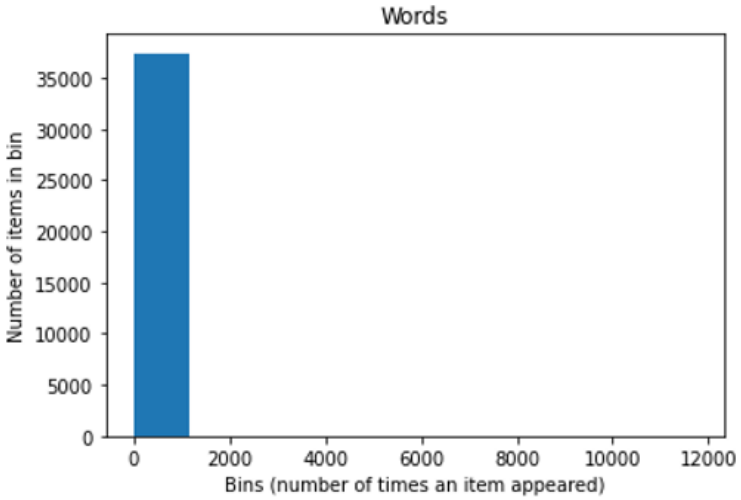
```
[15] for label, data in (('Words', words),
                         ('Screen Names', screen_names),
                         ('Hashtags', hashtags)):

         # Build a frequency map for each set of data
         # and plot the values
         c = Counter(data)
         plt.hist(list(c.values()))

         # Add a title and y-label ...
         plt.title(label)
         plt.ylabel("Number of items in bin")
         plt.xlabel("Bins (number of times an item appeared)")

         # ... and display as a new figure
         plt.figure()
```

*Hình 21: Code vẽ biểu đồ*

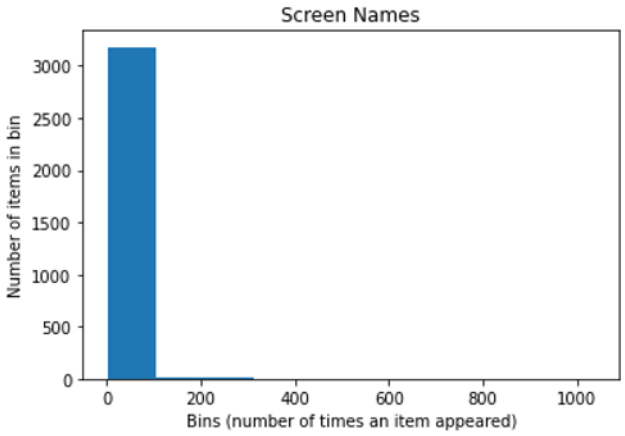



<Figure size 432x288 with 0 Axes>

*Hình 23: Biểu đồ hashtags*




*Hình 22: Biểu đồ word*




*Hình 24: Biểu đồ screen name*

- Từ dữ liệu thu được, trực quan hóa dữ liệu thành biểu đồ cột, ta thấy:

+ Phạm vi tần suất từ nằm trong khoảng [1,1000] và number of item in bin [0,35000].

+ Phạm vi tần suất Screen name nằm trong khoảng [1,100] và number of item in bin [0,3000].

+ Phạm vi tần suất Hashtag nằm trong khoảng [1,1000] và number of item in bin [0,2500]. Thống kê, chỉ có 3 hashtag (bitcoin, cryptocurrency, Bitcoin) có tần suất cao

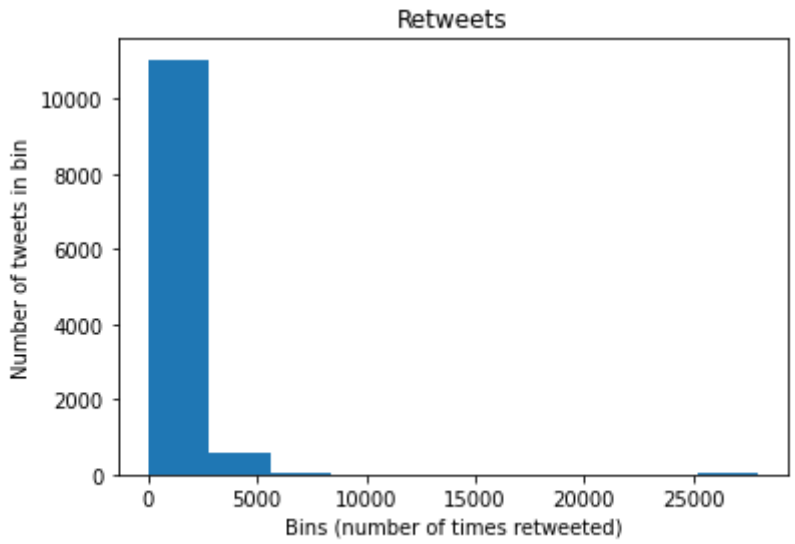
Qua đó, tổng số các word, screen name, hashtags có tần suất thấp trong các tweet chiếm phần lớn văn bản.

### 3.3.3. Vẽ biểu đồ về số lượng các Retweet

Chúng ta thấy rằng có một số tweet được chọn lọc và tweet lại với tần suất cao hơn nhiều so với phần lớn các tweet chỉ được tweet lại một lần và chiếm phần lớn khối lượng được đưa ra bởi lớn nhất hình chữ nhật màu xanh ở phía bên trái của biểu đồ

```
[16] # Using underscores while unpacking values in
     # a tuple is idiomatic for discarding them

     counts = [count for count, _, _, _ in retweets]

     plt.hist(counts)
     plt.title('Retweets')
     plt.xlabel('Bins (number of times retweeted)')
     plt.ylabel('Number of tweets in bin')
```

```
Text(0, 0.5, 'Number of tweets in bin')
```




*Hình 25: Code và biểu đồ cho Retweets*

Từ biểu đồ ta thấy rằng những bài tweet có lượng chia sẻ lại từ 1 đến 2500 lần chiếm số lượng lớn (trên 10000 bài), từ 2500 đến khoảng hơn 5000 lần có khoảng 200 bài. Từ 6000 lần chia sẻ hầu như rất ít.

### 3.3.4. Tìm 20 tweets có độ phổ biến cao nhất

Lượt retweet phổ biến nhất là chỉ cần lặp lại mỗi lần cập nhật trạng thái và lưu trữ số lượt retweet.

```
[11] retweets = [
                 # Store out a tuple of these three values ...
                 (status['retweet_count'],
                  status['retweeted_status']['user']['screen_name'],
                  status['retweeted_status']['id'],
                  status['text'])

                 # ... for each status ...
                 for status in statuses

                 # ... so long as the status meets this condition.
                     if 'retweeted_status' in status.keys()
                ]

    # Slice off the first 5 from the sorted results and display each item in the tuple

    pt = PrettyTable(field_names=['Count', 'Screen Name', 'Tweet ID', 'Text'])
    [ pt.add_row(row) for row in sorted(retweets, reverse=True)[:20] ]
    pt.max_width['Text'] = 50
    pt.align= 'l'
    print(pt)
```

```
+-------+----------------+---------------------+----------------------------------------------------+
| Count | Screen Name    | Tweet ID            | Text                                               |
+-------+----------------+---------------------+----------------------------------------------------+
| 27995 | wppenergycoin  | 1393258901764775947 | RT @wppenergycoin: WPP Token 🚀                    |
|       |                |                     | #cryptocurrency #crypto #cryptocurrencies          |
|       |                |                     | #cryptonews #cryptotrading #cryptocurrencynews     |
|       |                |                     | #cryptotrade  #cry…                                |
| 27995 | wppenergycoin  | 1393258901764775947 | RT @wppenergycoin: WPP Token 🚀                    |
```

*Hình 26: Code và kết quả tìm 20 tweets phổ biến*

Kết quả thu được:

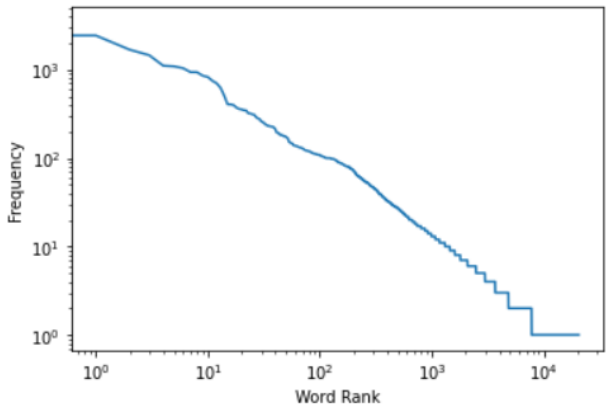
- Cryptocurrencies (tiền điện tử) đứng đầu danh sách. Và screen name wppenergycoin có tần suất reweet là 27995. Tweet ID: 1393258901764775947
- Tiếp đến là Bitboy_Crypto với tần suất retweet là 23758 Tweet ID: 137510433473680987. Sau đó là airdropinspect

Qua đó nhận thấy các chủ đề liên quan đến tiền điện tử đang là một xu hướng hot và được nhiều quan tâm theo dõi.

### 3.3.5. Vẽ biểu đồ tần số

```
#Trực quan quan hóa dữ liệu tần suất với biểu đồ
#giá trị y¬axis trên biểu đồ tương ứng với số lần một từ xuất hiện.
#giá trị x¬axis sẽ tương ứng với chỉ mục của bộ giá trị
# trong khoảng 5 chỉ mục của bộ giá trọ word rank thì số lần từ xuất hiện cao nhất xấp xỉ 10^4
import matplotlib.pyplot as plt
%matplotlib inline
word_counts = sorted(Counter(words).values(), reverse=True)

plt.loglog(word_counts)
plt.ylabel("Frequency")
plt.xlabel("Word Rank")
```

```
Text(0.5, 0, 'Word Rank')
```




*Hình 27: Code và biểu đồ tần số*

- Trực quan quan hóa dữ liệu tần suất với biểu đồ. Trong đó:

    + Trục hoành là số lượng từ

    + Trục tung là frequency (tần suất)

Ta thấy có 100 từ có tần suất xuất hiện gần 10000 lần.

### 3.3.6. Trích xuất các thực thể với NLTK

- Đoạn code này sẽ trích xuất các thực thể trong mỗi tweet bao gồm: hashtags, user mentions, links, stock tickers (symbols).

```
[23] import nltk
     import json
     import nltk
     nltk.download('averaged_perceptron_tagger')
     import nltk
     nltk.download('punkt')

     BLOG_DATA = "drive/My Drive/Elonmusk.json"

     blog_data = json.loads(open(BLOG_DATA).read())
     #tagger = nltk.data.load(_POS_TAGGER)
     for post in blog_data:
         #text=nltk.word_tokenize("Elon Musk says Tesla will no longer accept bitcoin ")
         sentences = nltk.tokenize.sent_tokenize(post['text'])
         tokens = [nltk.tokenize.word_tokenize(s) for s in sentences]
         pos_tagged_tokens = [nltk.pos_tag(t) for t in tokens]
         #  pos_tagged_tokens[[('Elonmusk', 'tesla'), ('bitcoin', 'tesla'), ('elonmusk', 'bitcoin'), ('elonmusk'),  ('.', '.')]]
         # Flatten the list since we're not using sentence structure
         # and sentences are guaranteed to be separated by a special
         # POS tuple such as ('.', '.')

         pos_tagged_tokens = [token for sent in pos_tagged_tokens for token in sent]

         all_entity_chunks = []
         previous_pos = None
         current_entity_chunk = []
         for (token, pos) in pos_tagged_tokens:
```

*Hình 28: Trích xuất các thực thể NLTK (1)*

```
[23]         if pos == previous_pos and pos.startswith('NN'):
                 current_entity_chunk.append(token)
             elif pos.startswith('NN'):

                 if current_entity_chunk != []:

                     # Note that current_entity_chunk could be a duplicate when appended,
                     # so frequency analysis again becomes a consideration

                     all_entity_chunks.append((' '.join(current_entity_chunk), pos))
                 current_entity_chunk = [token]

             previous_pos = pos

         # Store the chunks as an index for the document
         # and account for frequency while we're at it...

         post['entities'] = {}
         for c in all_entity_chunks:
             post['entities'][c] = post['entities'].get(c, 0) + 1

         # For example, we could display just the title-cased entities

         print(post['text'])
         print('-' * len(post['text']))
         proper_nouns = []
         for (entity, pos) in post['entities']:
             if entity.istitle():
                 print('\t{0} ({1})'.format(entity, post['entities'][(entity, pos)]))
         print()
```

*Hình 29: Trích xuất các thực thể NLTK (2)*

```
[ ]  RT @ Croesus_BTC; Bitcoin
     Bitcoin; stock; sports


     RT @Avatrode: So who benefits from the crypto #Tesla conflicts?
     #Bitcoin no
     #Tesla no
     @elonMusk no
     #dogecoin no
     Brokers yes
     Fossil fuel yes…
     ----------------------------------------------------------------------------------------------------------------------------------------------------------
     RT @ Avatrode; benefits; crypto; Tesla
     Bitcoin; Tesla; @; Brokers; Fossil

     Excellent point, Peter

     #Bitcoin is still in its infancy stages of life https://t.co/84K0aIDlBq
     ----------------------------------------------------------------------------------------------------
     point; Peter; Bitcoin; infancy; stages; life https

     Let's try something new @stoolpresidente #dogecoin #bitcoin #Ethereum #Litecoin #SAFEMOON #ShibaCoin let's come tog… https://t.co/vOVGUQFvWQ
     ----------------------------------------------------------------------------------------------------------------------------------------------------------
     something; @; stoolpresidente; Ethereum; Litecoin; SAFEMOON; ShibaCoin; let; tog… https

     RT @flurbnb: $60 to one person in 9 hours
```

*Hình 30: Trích xuất các thực thể NLTK (3)*

### 3.3.7. Tính toán sự đa dạng từ vựng trong các tweets

- Tính toán đa dạng từ vựng của Tweet: Một phép đo nâng cao hơn một chút liên quan đến việc tính toán các tần số đơn giản và có thể được áp dụng cho văn bản không có cấu trúc là một số liệu được gọi là đa dạng từ vựng.

- Qua phân tích: Sự đa dạng về từ vựng của các từ trong văn bản của các tweet là khoảng 0,11, mỗi cập nhật trạng thái mang khoảng 11% thông tin duy nhất. Tuy nhiên sự đa dạng về screen names còn cao hơn, Quan sát này cũng có ý nghĩa vì nhiều câu trả lời cho câu hỏi sẽ là tên hiển thị và hầu hết mọi người sẽ cung cấp các câu trả lời giống nhau. Sự đa dạng về mặt từ vựng của các thẻ bắt đầu bằng "#" là cực kỳ thấp, với giá trị khoảng 0,078, ngụ ý rằng rất ít giá trị ngoài thẻ bắt đầu bằng #bitcoin xuất hiện nhiều lần trong kết quả. Số lượng từ trung bình trên mỗi tweet là cao với giá trị 17, bản chất của hashtag, được thiết kế để thu hút các phản hồi dài.

```
[21] def lexical_diversity(tokens):
       return len(set(tokens))/len(tokens)
     def average_words(statuses):
       total_words= sum([len(s.split()) for s in statuses])
       return total_words/len(statuses)
     print(lexical_diversity(words))
     print(lexical_diversity(screen_names))
     print(lexical_diversity(hashtags))
     print(average_words(status_texts))

     0.11713754832897487
     0.16861777521129717
     0.07836910961514035
     17.782944874005675
```

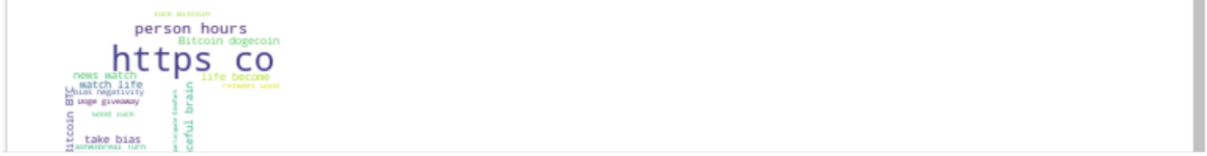
*Hình 31: Tính toán sự đa dạng từ vựng trong các tweet*

## 3.4. Visualize từ khóa từ tweets bằng hình ảnh

```
from collections import Counter
import matplotlib.pyplot as plt
from wordcloud import WordCloud
# Start with one review:
#text = df.description[0]
#text='With the wine dataset, you can group by country and look at
#either the summary statistics for all countries points and price or
#select the'
tW=''
for item in [words]:
  for i in item:
      tW=tW+' '+i
      tW=tW.replace('#','')

tSn=''
for item in [screen_names]:
  for i in item:
      tSn=tSn+' '+i
      tSn=tSn.replace('#','')

tHt=''
for item in [hashtags]:
  for i in item:
      tHt=tHt+' '+i
      tHt=tHt.replace('#','')


from wordcloud import WordCloud, STOPWORDS
from PIL import Image
import urllib
import requests
import numpy as np
```

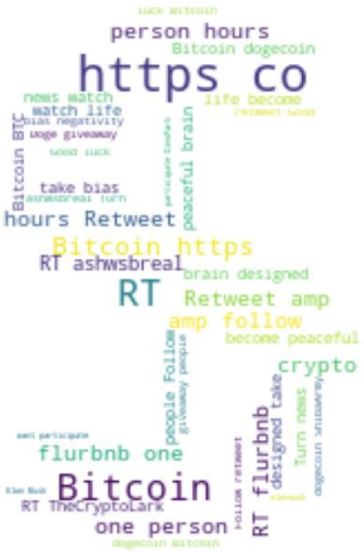
*Hình 32: Code để visualize trong Python (1)*

```
import urllib
import requests
import numpy as np
import matplotlib.pyplot as plt

mask = np.array(Image.open(requests.get('http://www.clker.com/cliparts/6/b/c/9/1194994651180363085 7money.svg.med.png', stream=True).raw))

# This function takes in your text and your mask and generates a wordcloud.
def generate_wordcloud(words, mask):
    word_cloud = WordCloud(width = 600, height = 600, background_color='white', stopwords=STOPWORDS, mask=mask).generate(words)
    plt.figure(figsize=(10,8),facecolor = 'white', edgecolor='blue')
    plt.imshow(word_cloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()

#Run the following to generate your wordcloud
generate_wordcloud(tW, mask)
generate_wordcloud(tSn, mask)
generate_wordcloud(tHt, mask)
```




*Hình 33: Code để visualize trong Python (2)*

- Hình trên là code để visualize trong Python, nhóm thực hiện cho cả Word, Screen name và hashtag, với biểu tượng là dollar như các hình bên dưới. Công cụ để visualize này rất đẹp mắt và dễ nhìn cũng như rất sinh động và trực quan. Các word như Bitcoin, screen name elonmusk, hashtag Bitcoin thể hiện rõ trên visualize. Cho thấy sự quan tâm của khán giả rất lớn đối với chủ đề này.



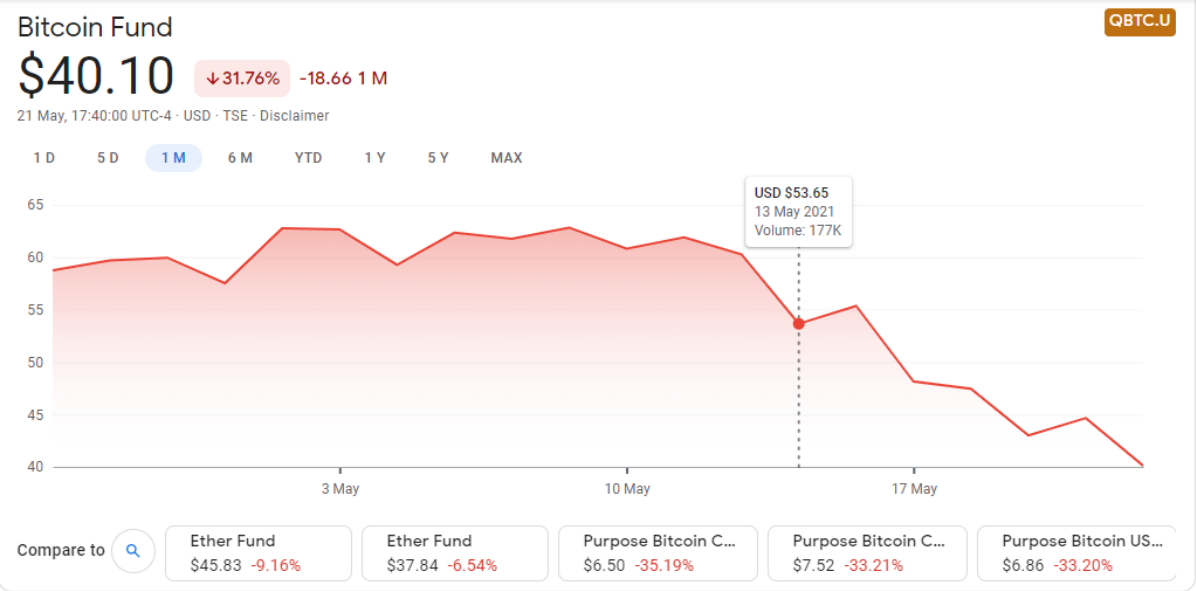

*Hình 34: Visualize Word*


*Hình 35: Visualize Screen Name*


*Hình 36: Visualize Hashtag*

## 3.5. Phân tích thay đổi tỷ giá Bitcoin

- Trước đó, vào ngày 18/02/2021 Tesla đã công bố việc mua 1,5 tỷ đô la bitcoin trong nỗ lực "đa dạng hóa hơn nữa và tối đa hóa" lợi tức tiền mặt của mình. Tại thời điểm này, giá bitcoin đã tăng trưởng đạt kỷ lục 59.04 US nhờ tesla.Trong khoảng thời gian này, Elon Musk đã gây kích thích cho các thị trường tiền kỹ thuật số bằng cách thể hiện sự quan tâm và chấp nhận đồng tiền này và tỷ giá Bitcoin có chiều hướng tăng đều trong giai đoạn này.
- Tuy nhiên, hình hình đổi chiều sau khi vị tỷ phú đăng tweet vào ngày 13/05/2021 cho biết Tesla dừng chấp nhận Bitcoin làm phương tiện thanh toán vì quan ngại môi trường. Chỉ sau một ngày giá Bitcoin giảm 11% khoảng từ 62.28USD xuống còn 53.65 USD. Theo những phân tích của nhóm cũng thấy cho thấy rằng người dùng quan tâm rất nhiều đến tuyên bố của Elon Musk, vị tỷ phú đã tác động không nhỏ đến việc rớt giá của đồng tiền này.
- Những ngày sau tuyên bố, tỷ giá Bitcoin liên tục giảm đều và không có xu hướng gia tăng. Qua đó, thấy được tầm ảnh hưởng lớn của Elon Musk đối với sự biến động tỷ giá Bitcoin trên toàn thế giới.




*Hình 37: Tỉ giá Bitcoin ngày 21/05/2021*

# CHƯƠNG 4: TỔNG KẾT

## 4.1. Kết quả đạt được

- Để thực hiện đồ án "**Phân tích sự phản ứng của người dùng Twitter đối với tuyên bố của Elon Musk về Bitcoin ngày 13/05/2021**" nhóm đã thực hiện các công việc chính là thu thập 200 tweets từ mạng xã hội Twitter, sau đó tiến hành xử lý phân tích và đưa ra các nhận định. Trong quá trình xử lý phân tích, nhóm đã sử dụng ngôn ngữ Python.

- Các kết quả đạt được:

+ Tìm hiểu các lý thuyết liên quan đến khai phá dữ liệu và đi sâu hơn về khai phá dữ liệu trên mạng xã hội.

+ Áp dụng được những kiến thức đã học về các dòng lệnh khai phá dữ liệu trên Python trong quyển sách Mining-the-Social-Web-3rd-Edition-master. Điều chỉnh và xây dựng được công cụ lấy dữ liệu tự động cho đề tài của nhóm.

+ Tự viết được các lệnh đơn giản trong Python.

+ Phân tích được mức độ quan tâm của người dùng Twitter đối với tuyên bố của Elon Musk.

## 4.2. Hạn chế đồ án

- Mặc dù đã cố gắng hết sức nhưng do vấn đề về kinh nghiệm, kiến thức và thời gian nên đồ án vẫn có một số hạn chế nhất định như sau:

- Do có sự hạn chế về công cụ, kiến thức và kỹ năng nên kết quả khi phân tích tiêu cực, tích cực vẫn chưa có độ chính xác cao.

- Các nhận xét, kết luận chỉ mang tính tham khảo tổng quát.

- Dữ liệu thu thập chưa được nhiều và đầy đủ.

- Chưa sử dụng được nhiều công cụ để phân tích hiệu quả hơn.

### 4.3. Phương hướng phát triển

- Sử dụng kết quả nghiên cứu làm tiền đề cho những nghiên cứu tương tự hoặc mở rộng sang những nghiên cứu khác nhưng có liên quan đến đề tài.
- Tìm ra phương pháp thực hiện nâng cao hiệu quả nghiên cứu về bitcoin ở Việt Nam.
- So sánh những nghiên cứu khoa học, những bài phân tích cùng vấn đề bitcoin để tìm cách khắc phục các vấn đề gặp phải của đồ án.
- Thu thập nguồn dữ liệu đầy đủ từ nhiều trang web liên quan đến thị trường tiền ảo hơn nữa để mang tính chuyên môn cao hơn.
- Kết hợp thêm nhiều công cụ để phân tích dữ liệu chính xác hơn.

# DANH MỤC THAM KHẢO

[1] Phệ, A. B. (2021, January 11). *Mạng xã hội là gì? Những MXH phổ biến ở Việt Nam hiện nay*.

https://canhrau.com/mang-xa-hoi-la-gi/

[2] Vietnambiz. (2019, October 18). Dữ liệu mạng xã hội (Social Data) là gì? Lợi ích và hạn chế của dữ liệu mạng xã hội Vietnambiz.
https://vietnambiz.vn/du-lieu-mang-xa-hoi-social-data-la-gi-loi-ich-va-han-che-cua-du-lieu-mang-xa-hoi-20191017120946614.htm

[3] phuongttt, B. (2019, July 9). *Mạng xã hội Twitter là gì? Những điều bạn cần biết khi sử dụng Twitter*.
https://webdoctor.vn/mang-xa-hoi-twitter-la-gi-nhung-dieu-ban-can-biet-khi-su-dung-twitter/

[4] Best, R. de. (2021, May 19). Bitcoin market cap 2013-2021. Statista.
https://www.statista.com/statistics/377382/bitcoin-market-capitalization/

[5] Best, R. de. (2021, May 20). *Most valuable cryptocurrency 2021*. Statista.
https://www.statista.com/statistics/655492/most-valuable-virtual-currencies-globally/

[6] *Data Mining là gì? Lợi ích khai phá dữ liệu với công nghệ 4.0*. RenovaCloud. (n.d.).

https://renovacloud.com/data-mining-la-gi-loi-ich-khai-pha-du-lieu-voi-cong-nghe-4-0/

# PHÂN CÔNG CÔNG VIỆC

| Thành viên | Công việc |
|---|---|
| **Trần Trí Tín** | Chương 2 (2.1), chương 3 (3.4), Leader |
| **Huỳnh Trí Vĩ** | Chương 1, chương 4, tổng hợp word, thuyết trình |
| **Trần Minh Nghĩa** | Chương 2 (2.2.1, 2.2.2), chương 3 (3.1, 3.2, 3.3.1, 3.3.2, 3.3.3), PPT |
| **Lê Thị Kiều Ly** | Chương 2 (2.2.3, 2.2,4), chương 3 (3.3.4, 3.3.5, 3.3.6, 3.3.7, 3.3.8, 3.5) |

**◈◈◈HẾT◈◈◈**