Đại học Kinh tế – Luật

Đại học Quốc gia Thành phố Hồ Chí Minh

-----📖-----

Báo cáo đồ án môn học Phân tích dữ liệu web:

PHÂN TÍCH DỮ LIỆU VỀ MỨC ĐỘ QUAN TÂM TIỀN MÃ HÓA TRÊN MẠNG XÃ HỘI TWITTER

Nhóm nghiên cứu: BEST

Vũ Quang Huy	K184111445
Đỗ Nguyễn Nhật Hàn	K184111444
Lê Trần Giản Đơn	K184111442
Phan Hồng Oanh	K184111457

Giảng Viên: Ths. Đặng Nhân Cách

LỜI MỞ ĐẦU

Lời đầu tiên, nhóm BEST xin gửi lời cảm ơn chân thành đến Thầy Đặng Nhân Cách - Giảng viên Hướng dẫn của nhóm. Qua những bài giảng, tài liệu tham khảo cũng như bài tập nhận được từ thầy, chúng em đã có đủ kiến thức để hoàn thành phần đồ án của môn Phân tích dữ liệu WEB.

Ngoài ra, nhóm chúng em cũng muốn gửi lời cảm ơn đến các anh/ chị/ bạn bè đã giúp đỡ cho chúng em trong quá trình thực hiện và hoàn thiện đồ án này một cách chỉnh chu nhất.

Chúng em đã rất cố gắng trong việc tìm hiểu và học tập thêm những công nghệ khác nhau, tuy nhiên với vốn kiến thức còn hạn chế, đồ án này không thể tránh khỏi những sai sót. Rất mong nhận được sự góp ý từ thầy để đồ án được phát triển hơn nữa.

Một lần nữa nhóm BEST xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

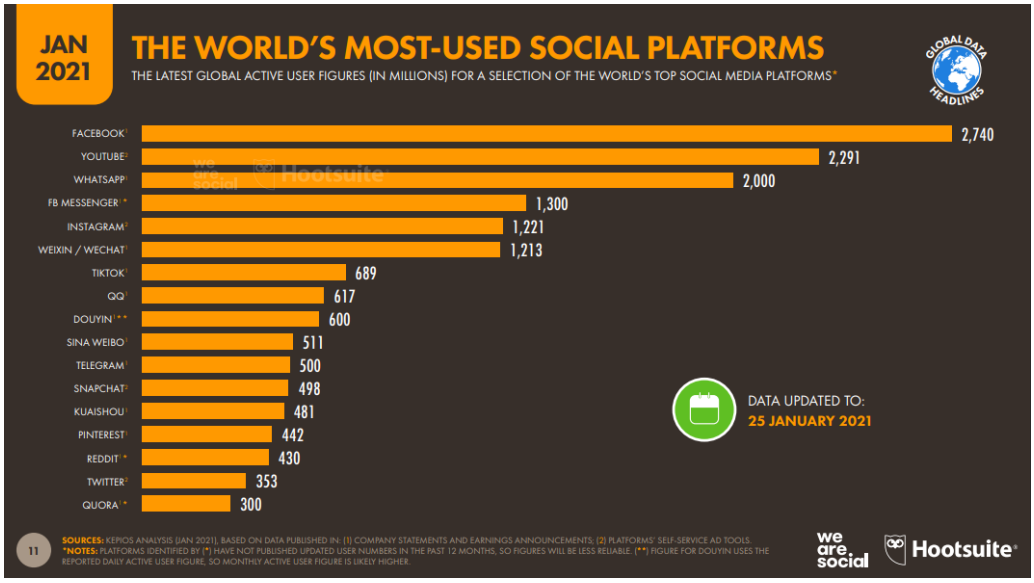
CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1. Lý do chọn đề tài

Năm 2009, tiền mã hóa bắt đầu xuất hiện và trở thành một thị trường đầy biến động trên thế giới. Sau 2 năm, tiền mã hóa dần được chú ý đến nhiều hơn, đặc biệt là với Bitcoin - đồng tiền kỹ thuật số đầu tiên và cho đến nay, đã có thêm hàng trăm đồng tiền mã hóa khác ra đời.

Thời gian gần đây, sự thay đổi mạnh mẽ về giá của các đồng tiền mã hóa bị gây ra bởi những biến động từ thị trường tài chính và tác động xã hội dẫn đến như: các sự kiện chính trị và kinh tế trên toàn thế giới; thay đổi cung-cầu thị trường, tổng số lượng tiền mã hóa và người sở hữu chúng cũng như những ảnh hưởng từ các loại tiền mã hóa khác. Với sự phát triển và những biến động giá mạnh mẽ của tiền mã hóa, nhất là sự lên ngôi của đồng Bitcoin gần đây đã thu hút sự quan tâm của giới tài chính và dư luận trên nền tảng mã hội Twitter.

Theo báo cáo Digital 2021 của We are social và Hootsuite, Twitter giữ vị trí 16 trong tổng số các nền tảng mạng xã hội được người dùng sử dụng nhiều nhất trên toàn thế giới với 353 triệu người dùng. Bên cạnh đó, những người có sức ảnh hưởng lớn trên thế giới quan tâm về tiền mã hóa thường đăng tải và bày tỏ quan điểm của mình trên mạng xã hội này, gây tác động rất nhiều đến thị trường tài chính - tiền tệ thế giới. Một số nhân vật điển hình có thể kể đến như Elson Mush, Andreas Antonopoulos, Adam Back, Nick Szabo, Josh Olszewicz, WhalePanda,... Điều này cho thấy mạng xã hội này ảnh hưởng không nhỏ đến mức độ thảo luận của người dùng với thị trường tiền mã hóa biến động này.



Hình 1. 1: Thống kê các nền tảng mạng xã hội được sử dụng nhiều nhất trên

Từ đó, nhóm nhận thấy việc tiến hành phân tích, đánh giá dữ liệu về mức độ quan tâm của người dùng về tiền điện tử trên nền tảng mạng xã hội Twitter là cần thiết. Với kết quả đạt được theo bài nghiên cứu, chúng tôi có thể áp dụng để đánh giá và đưa ra các kết luận mức độ quan tâm và dự đoán mức độ phổ biến của top 10 đồng tiền mã hóa được chú ý nhiều nhất cũng như đề xuất giải pháp giúp tăng trưởng thị trường tiền mã hóa trong tương lai.

1.2. Mục tiêu đề tài

- Phân tích các dữ liệu liên quan về mức độ quan tâm của người dùng đến đồng tiền kỹ thuật số trên mạng xã hội Twitter.
- Đề xuất các giải pháp tăng sự tiếp cận và sử dụng tiền mã hóa của người dùng.

1.3. Phạm vi nghiên cứu

- **Không gian:** Thu thập và nghiên cứu dữ liệu từ mạng xã hội Twitter.
- **Thời gian:** Đào dữ liệu real time trong 7 ngày.
- **Lĩnh vực nghiên cứu:** Tài chính, Tiền mã hóa.

1.4. Đối tương nghiên cứu

- Đối tượng nghiên cứu: Mức độ quan tâm sử dụng của người dùng đối với các loại tiền Cryptocurrency trên mạng xã hội Twitter.
- Khách thể nghiên cứu: Thông tin nội dung của các Tweets trên mạng xã hội Twitter có liên quan đến Cryptocurrency.

1.5. Phương pháp nghiên cứu

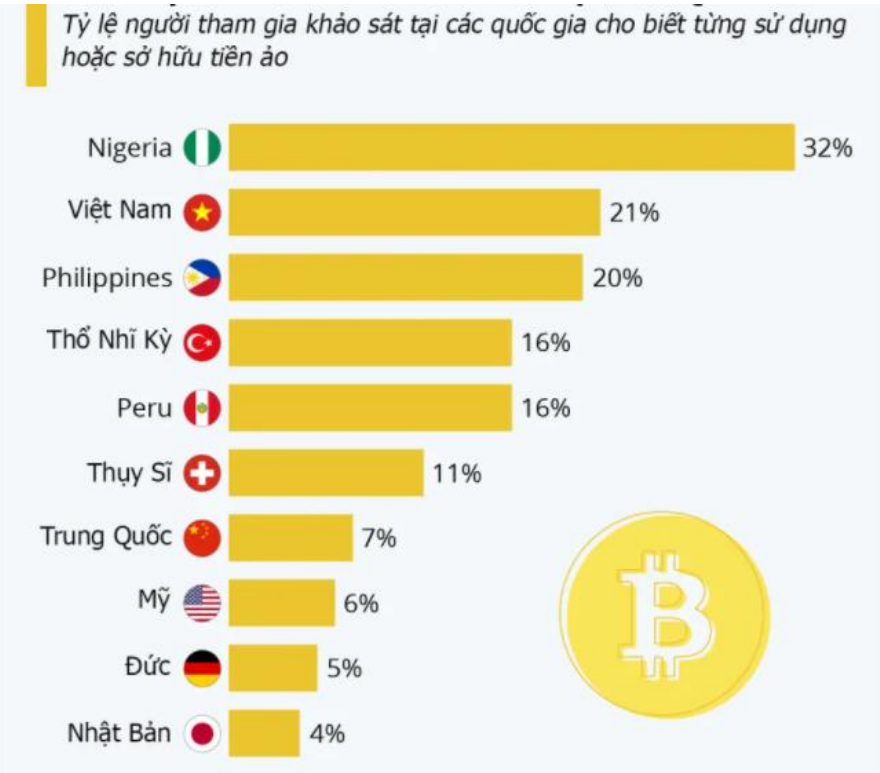
- Phương pháp phân tích, thống kê: Phân tích các hiện tượng, xu hướng và các số liệu thống kê để đưa ra những nhận định hoặc giải pháp cho vấn đề.
- Phương pháp mô tả: Từ phân tích số liệu tiến hành diễn giải dữ kiện để có cái nhìn khái quát và khách quan về các vấn đề gặp phải.
- Phương pháp thu thập số liệu: Sử dụng những thông tin đã sẵn có từ các nguồn khác nhau hoặc chủ động thu thập thông qua việc đào và khai thác dữ liệu trên các nền tảng website, mạng xã hội, ...
- Phương pháp phân loại và hệ thống hóa lý thuyết: Sắp xếp các tài liệu khoa học theo từng mặt, từng đơn vị, từng vấn đề có cùng dấu hiệu bản chất, cùng một hướng phát

triển và chuẩn bị tri thức thành một hệ thống trên cơ sở một mô hình lý thuyết làm sự hiểu biết về đối tượng tất tần tật hơn.

- Phương pháp tổng kết kinh nghiệm: Là phương pháp nghiên cứu và xem xét lại những thành quả thực tiễn trong quá khứ để rút ra tóm lại bổ ích cho thực tiễn và khoa học.

1.6. Ý nghĩa đề tài

Hiện nay, với xu hướng phát triển tất yếu của công nghệ Blockchain và sự xuất hiện, phát triển của các loại tiền mã hoá (Cryptocurrencies) được coi là xu thế của tương lai. Năm 2021 là năm mà các loại tiền mã hóa nhận được sự quan tâm đặc biệt từ giới đầu tư, thậm chí đến các nhân vật nổi tiếng có tầm ảnh hưởng trên thế giới cũng thể hiện sự quan tâm của mình đến loại tiền này - thể hiện rõ nhất trên mạng xã hội Twitter. Cơn sốt về tiền mã hóa trở nên nóng hơn bao giờ hết khi đồng Bitcoin (BTC) - đồng tiền mã hóa lớn nhất thị trường. Khởi điểm ở mức 30.000USD/BTC đầu năm 2021, bitcoin liên tiếp thiết lập kỷ lục về giá, gần nhất là đỉnh 61.500USD/BTC. Theo thống kê người tiêu dùng toàn cầu của Statista năm 2020, Việt Nam là quốc gia đứng thứ 2 thế giới về độ phổ biến của tiền mã hóa với 21% người tham gia khảo sát cho biết đã từng sử dụng hoặc sở hữu tiền mã hóa.



Hình 1. 2: Top 10 quốc gia có độ phổ biến tiền ảo cao nhất thế giới

Có thể nhận thấy xu hướng và mức độ quan tâm của người tiêu dùng Việt Nam đến loại tiền tệ này ngày càng lớn, điều đó chứng tỏ tính cấp thiết của đề tài trong bối cảnh các quốc gia lớn như Mỹ dần chấp nhận chúng trong các giao dịch thương mại và Việt Nam, trong tương lai, có thể không phải ngoại lệ. Đề tài sẽ đi sâu vào tìm hiểu và phân tích các tính chất đặc trưng của tiền mã hóa và mối quan hệ của chúng với sự quan tâm của người tiêu dùng trên thế giới và chỉ ra danh sách các khu vực quan tâm đến tiền mã hóa trên thế giới.

1.7. Kết cấu đề tài

Đề tài nghiên cứu được chia làm 4 chương:

- ***Chương 1: Tổng quan đề tài***

Trình bày về tổng quan về đề tài bao gồm bối cảnh hiện nay, lý do chọn đề tài, đối tượng, phạm vi, phương pháp nghiên cứu, mục tiêu và ý nghĩa thực tiễn cũng như tính cấp thiết của đề tài nghiên cứu.

- ***Chương 2: Cơ sở lý thuyết***

Trình bày cơ sở lý thuyết về khái niệm, lý thuyết về tiền mã hóa, thị trường tiền mã hóa; mạng xã hội Twitter và các công cụ hỗ trợ khai thác, phân tích dữ liệu như Python, Java, Twitter API, ...

- ***Chương 3: Khai thác và phân tích dữ liệu***

Thực hiện các bước khai thác và phân tích dữ liệu cụ thể. Sử dụng ngôn ngữ Python và twitter API để khai thác dữ liệu trên mạng xã hội Twitter thông qua các hashtag liên quan và trực quan hóa dữ liệu để đưa ra các phân tích, nhận định, dự đoán về xu hướng tiền mã hóa tại Việt Nam.

- ***Chương 4: Kết luận và kiến nghị***

Trình bày tóm tắt kết quả nghiên cứu, kết quả đạt được của nghiên cứu, những đóng góp của nghiên cứu và hướng phát triển tiếp theo của đề tài.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Tiền mã hóa – Cryptocurrency

2.1.1. Khái niệm tiền mã hóa

Tiền mã hóa thường được gọi là cryptocurrency, là một loại tiền - tài sản kỹ thuật số sử dụng mật mã để bảo mật, được thiết kế làm phương tiện trao đổi sử dụng cơ chế mã hóa để đảm bảo các giao dịch tài chính, kiểm soát việc tạo ra các đơn vị bổ sung và xác minh việc chuyển giao tài sản. Đồng tiền này sử dụng kiểm soát phi tập trung trái ngược với tiền tệ kỹ thuật số tập trung và hệ thống ngân hàng trung ương. Tiền mã hóa được phân loại như một tập con của loại tiền tệ kỹ thuật số.

Tiền mã hóa được dựa vào một lĩnh vực toán học đặc biệt có tên là mật mã học. Mật mã học là khoa học về bảo mật thông tin, và có hai thành tố thật sự quan trọng: giấu thông tin bằng cách đưa về dạng chưa mã hóa, và kiểm tra nguồn của một thông tin. Mật mã học củng cố nhiều hệ thống quanh ta. Và nó mạnh mẽ đến nỗi nhiều lúc chính phủ Hoa Kỳ liệt nó vào một loại vũ khí.

Cryptocurrency là các hệ thống cho phép thanh toán an toàn cho các giao dịch trực tuyến là token ảo, đại diện cho các mục sổ cái bên trong chính hệ thống "cryptocurrency" đề cập đến thực tế là các thuật toán mã hóa và kỹ thuật mã hóa khác nhau, chẳng hạn như mã hóa đường cong elip, cặp khóa công khai và hàm băm, được sử dụng.

Cryptocurrency dựa trên blockchain đầu tiên là Bitcoin, vẫn là loại tiền phổ biến nhất và có giá trị nhất. Ngày nay, có hàng ngàn loại cryptocurrencies thay thế với các chức năng hoặc thông số kỹ thuật khác nhau. Một số trong số này là bản sao của Bitcoin trong khi một số khác là nhánh hoặc cryptocurrency mới tách ra từ một loại đã tồn tại.

2.1.2. Các dịch vụ hướng đến thanh toán tiền mã hóa

Bên cạnh các loại chứng khoán, ngày nay các nhà đầu tư cũng quan tâm đến việc đầu tư vào bitcoin như 1 phương thức để sinh ra lợi nhuận như các loại chứng khoán. Dựa vào đặc điểm giá cả biến động của các loại tiền mã hóa này mà các nhà đầu tư thường đầu tư lướt sóng(Mua và bán ra để kiến lợi), đầu tư nắm giữ (mua và tích trữ chờ tăng giá). Ở các quốc gia phát triển và hợp thức hóa các đồng tiền mã hóa này đã có các sàn giao dịch tiền mã hóa, điển hình như: binance, poloniex, bittrex….. Theo số liệu cụ thể từ sàn giao dịch tiền mã hóa Binance- sàn giao dịch lớn nhất hiện nay, có khoảng 1200000000 khối lượng giao dịch hàng ngày được thực hiện trên Binance, 1400000 giao dịch trên giây, và

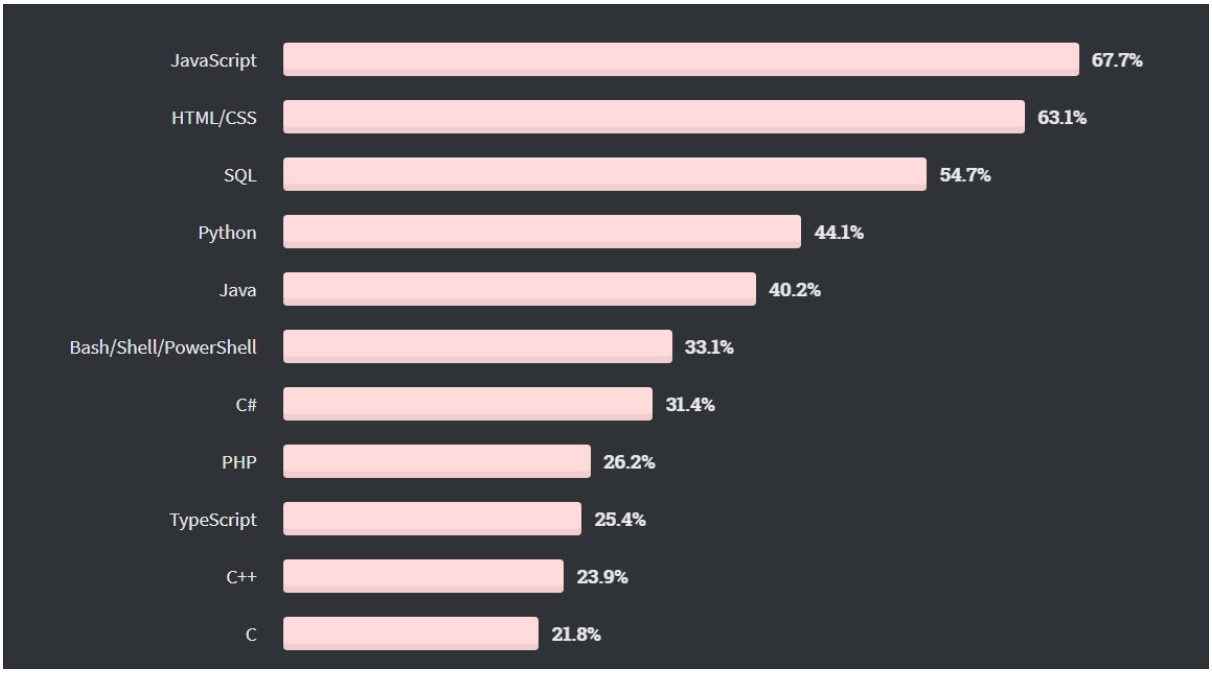
10000000 người dùng Binance, … Đây là những con số vô cùng lớn, chứng minh rằng đầu tư vào thị trường tiền mã hóa ngày càng phổ biến hiện nay.

Nhiệm vụ chính của các đồng tiền mã hóa chính là thực hiện các giao dịch cơ bản trong cuộc sống hiện nay. Khi công nghệ thanh toán điện tử càng phát triển thì việc sử dụng các đồng tiền mã hóa trong giao dịch sẽ ngày càng được phổ biến hơn. Đơn cử, các công ty lớn như Microsoft, Dell… hay các hãng hàng không AirBaltic… đến các dịch vụ giải trí khác đã chấp nhận việc thanh toán bằng các đồng tiền mã hóa. Thậm chí mới đây, sàn giao dịch Coinbase đã cho ra mắt Coinbase Card để giúp các giao dịch bằng tiền mã hóa dễ dàng hơn.

2.2. Ngôn ngữ lập trình Python

2.2.1. Giới thiệu về Python

Python là ngôn ngữ lập trình bậc cao, được ra mắt lần đầu vào năm 1991. Đây là một trong những ngôn ngữ đang có xu hướng phát triển mạnh mẽ trong những năm gần đây với sự lên ngôi của trí tuệ nhân tạo và phân tích dữ liệu. Theo thống kê của Stackoverflow khảo sát hơn 65.000 lập trình viên vòng quanh thế giới, Python đang có số lượng vượt hơn cả Java - một ngôn ngữ lập trình được dùng xây dựng các ứng dụng trong các doanh nghiệp tập đoàn lớn yêu cầu cao về hiệu suất.



Hình 1. 3: Top 10 ngôn ngữ lập trình được sử dụng nhất nhất hiện tại

Khác với một số ngôn ngữ khác, Python là một ngôn ngữ dễ đọc dễ học, sử dụng trình thông dịch (compiler) - tức bắt chước ngôn ngữ nguồn để thực hiện. Do đó, cú pháp của Python khá ngắn gọn so với các ngôn ngữ bậc cao khác như C, C++ hay C#, tuy vậy với các chương trình có ứng dụng lớn, Python sẽ chạy tương đối chậm so với các ngôn ngữ kể trên nếu không được tối ưu hóa dòng lệnh.

2.2.2. Một số đặc điểm, tính năng nổi bật của Python

- Ngôn ngữ lập trình đơn giản, dễ học: Python có cú pháp rất đơn giản, rõ ràng, dễ đọc và viết hơn rất nhiều khi so sánh với những ngôn ngữ lập trình khác như C++, Java, C#. Python tập trung vào những giải pháp chứ không phải cú pháp. Tính tự nhiên của mã giả trong Python là 1 trong các điểm mạnh nhất của ngôn ngữ này. Điều này giúp cho lập trình viên tập trung vào giải pháp giải quyết vấn đề hơn là việc tập trung vào ngôn ngữ

- Miễn phí, mã nguồn mở: Vì là mã nguồn mở, bạn không những có thể sử dụng các phần mềm, chương trình được viết trong Python mà còn có thể thay đổi mã nguồn của nó. Một trong những lý do Python là ngôn ngữ mạnh vì nó được cộng đồng thường xuyên phát triển và nâng cấp.

- Khả năng di chuyển: Các chương trình Python có thể di chuyển từ nền tảng này sang nền tảng khác và chạy nó mà không có bất kỳ thay đổi nào. Nó chạy liền mạch trên hầu hết tất cả các nền tảng như Windows, macOS, Linux.

- Khả năng mở rộng và có thể nhúng: Python cho phép người dùng có thể dễ dàng kết hợp các phần code bằng C, C++ và những ngôn ngữ khác (có thể gọi được từ C) vào code Python. Điều này sẽ cung cấp cho ứng dụng của người lập trình những tính năng tốt hơn cũng như khả năng scripting mà những ngôn ngữ lập trình khác khó có thể làm được.

- Ngôn ngữ thông dịch cấp cao: Không giống như C/C++, với Python, người lập trình không phải lo lắng những nhiệm vụ khó khăn như quản lý bộ nhớ, dọn dẹp những dữ liệu vô nghĩa... Khi chạy code Python, nó sẽ tự động chuyển đổi code sang ngôn ngữ máy tính có thể hiểu.

- Thư viện tiêu chuẩn lớn để giải quyết những tác vụ phổ biến: Python có một số lượng lớn thư viện tiêu chuẩn giúp cho công việc lập trình của bạn trở nên dễ thở hơn rất nhiều, đơn giản vì không phải tự viết tất cả code.

- Hướng đối tượng: Mọi thứ trong Python đều là hướng đối tượng. Lập trình hướng đối tượng (OOP) giúp giải quyết những vấn đề phức tạp một cách trực quan. Với OOP, người dùng có thể phân chia những vấn đề phức tạp thành những tập nhỏ hơn bằng cách tạo ra các đối tượng. Nếu so sánh với C++ hoặc Java, Python rất mạnh nhưng lại cực kỳ đơn giản để thực hiện lập trình hướng đối tượng
- Python được ứng dụng nhiều trong: Trí tuệ nhân tạo (Machine Learning, Deep Learning), lập trình nhúng, phân tích dữ liệu

2.2.3. Các thư viện Python phục vụ nghiên cứu

2.2.3.1. Numpy

Numpy (Numeric Python): là một thư viện toán học phổ biến và mạnh mẽ của Python. Cho phép làm việc hiệu quả với ma trận và mảng. Numpy là một thư viện lõi phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó. Numpy đặc biệt hữu ích khi thực hiện các hàm liên quan tới Đại Số Tuyến Tính.

Để cài đặt numpy bằng Anaconda chỉ cần gõ "conda install numpy" hoặc sử dụng tools pip "pip install numpy". Sau khi cài đặt xong, trong Python, chúng ta cần khai báo "import numpy" để có thể bắt đầu sử dụng các hàm của numpy.

2.2.3.2. Pandas

Pandas là một thư viện mã nguồn mở được xây dựng dựa trên NumPy, sử dụng thao tác và phân tích dữ liệu, được thiết kế để cho phép bạn làm việc với dữ liệu được gắn nhãn hoặc quan hệ theo cách trực quan hơn:

- Có thể xử lý tập dữ liệu khác nhau về định dạng: chuỗi thời gian, bảng không đồng nhất, ma trận dữ liệu
- Khả năng import dữ liệu từ nhiều nguồn khác nhau như CSV, DB/SQL
- Có thể xử lý vô số phép toán cho tập dữ liệu: subsetting, slicing, filtering, merging, groupBy, re-ordering, and re-shaping, ...
- Xử lý dữ liệu mất mát theo ý người dùng mong muốn: bỏ qua hoặc chuyển sang 0
- Xử lý, phân tích dữ liệu tốt như mô hình hoá và thống kê

- Tích hợp tốt với các thư viện khác của python
- Cung cấp hiệu suất tốt

Để cài đặt pandas bằng Anaconda chỉ cần gõ "conda install pandas" hoặc sử dụng tools pip "pip install pandas". Sau khi cài đặt xong, trong Python, chúng ta cần khai báo "import pandas" để có thể bắt đầu sử dụng các hàm của pandas.

2.2.3.3. Beautiful Soup, Boilerpipe 3 và Regex

- BeautifulSoup là một thư viện Python dùng để lấy dữ liệu ra khỏi các file HTML và XML. Nó hoạt động cùng với các parser (trình phân tích cú pháp) cung cấp cho bạn các cách để điều hướng, tìm kiếm và chỉnh sửa trong parse tree (cây phân tích được tạo từ parser). Nhờ các parser này nó đã giúp các lập trình viên tiết kiệm được nhiều giờ làm việc.
- Boilerpipe 3 là một thư viện để loại bỏ bản soạn sẵn và trích xuất văn bản đầy đủ từ các trang HTML.
- Regular Expression (RegEx) hay còn gọi là Biểu thức chính quy là một đoạn các ký tự đặc biệt theo những khuôn mẫu (pattern) nhất định, đại diện cho chuỗi hoặc một tập các chuỗi. VD: ^a...s$ có nghĩa là bất kỳ chuỗi nào có năm chữ cái, bắt đầu bằng a và kết thúc bằng s.

2.2.3.4. NLTK và spaCy

- NLTK hay còn gọi là natural Language Toolkit một trong những thư viện Python dẫn đầu trong việc xử lý ngôn ngữ tự nhiên (NLP - Natural language processing) để phục vụ cho các mục đích phân tích hoặc các nền tảng AI. Thư viện này hỗ trợ các công cụ như: phân nhánh, đánh dấu, hay hỗ trợ phân tích sentiment (đánh giá về mức độ tích cực hay tiêu cực của các bài viết thông qua từ ngữ và điểm số trên thang điểm 1).
- spaCy là một thư viện mã nguồn mở miễn phí phục mục cho NLP với Python. Thư viện này giúp chúng ta xử lý các vấn đề liên quan đến từ ngữ, câu từ đó trả lời cho các câu hỏi: nó là gì? trong ngữ cảnh đó câu từ đó mang ý nghĩa nào? và ai đang làm gì với ai,...Với mục đích xây dựng phục vụ cho các sản phẩm thực tế nên thư viện này có thể xử lý một lượng lớn các thông tin một lúc, mạnh mẽ hơn các thư viện khác.

2.3. Ngôn ngữ lập trình Java

2.3.1. Giới thiệu về Java

Java là một ngôn ngữ lập trình có tính bảo mật cao được phát triển bởi Sun Microsystem vào năm 1995, là ngôn ngữ kế thừa trực tiếp từ C/C++ và là một ngôn ngữ lập trình hướng đối tượng dựa trên các lớp. Ngoài ra Java còn được biết đến như một Platform - là một tập hợp các chương trình giúp phát triển và chạy các chương trình được viết bằng ngôn ngữ lập trình Java.

- Platform: Bất cứ môi trường phần cứng hoặc phần mềm nào mà trong đó một chương trình chạy, thì được biết đến như là một Platform. Với môi trường runtime riêng cho mình là JRE và API, Java được gọi là Platform.
- Java Platform: Gồm có 3 thành phần chính:
    - Java Virtual Machine (Java VM): Máy ảo Java.
    - Java Application Programming Interface (Java API).
    - Java Development Kit (JDK) gồm trình biên dịch, thông dịch, trợ giúp, soạn tài liệu... và các thư viện chuẩn.

2.3.2. Một số đặc điểm, tính năng nổi bật của Java

- Phát triển ứng dụng cho các thiết bị điện tử thông minh, các ứng dụng cho doanh nghiệp với quy mô lớn, cơ sở dữ liệu, mạng, Internet, viễn thông, giải trí.
- Ứng dụng web: Tạo các trang web có nội dung động (web applet), nâng cao chức năng của server.
- Ứng dụng trong máy chủ dịch vụ tài chính.
- Công nghệ Big Data: Java được xem là có tiềm năng lớn để có thể đạt được thị phần ngày càng cao nếu như Hadoop hoặc ElasticSearch lớn mạnh.

2.3.3. Các thư viện Java phục vụ cho nghiên cứu

- Apache POI: là một thư viện mã nguồn mở Java, được cung cấp bởi Apache. Thư viện này cung cấp các API (phương thức) làm việc với các tài liệu của Microsoft như Word, Excel, Powerpoint, Visio,…
- Twitter4j, TwitterStream và TwitterStreamFactory: là 3 thư viện thu thập thông tin do Twitter API xây dựng; trong đó Twitter4j được sử dụng để thu thập các tweet

đăng tải trên Twitter rồi truyền về Kafka (hệ thống message pub/ sub phân tán). Dữ liệu từ Kafka sau đó được lưu vào Cassandra Database, đồng thời cũng được chuyển đến Spark Streaming để thực hiện việc xử lý thông tin trong thời gian thực. Tiếp theo chúng ta sử dụng Spark Batch Processing để phân tích dữ liệu trong Cassandra. Ở đây ta sử dụng Akka Scheduler để thiết lập việc chay Batch processing cứ 30 phút một lần. Kết quả phân tích của Spark Streaming (real time) và Spark Batch processing sau đó được lưu vào Cassandra. Cuối cùng, chúng ta viết một Client UI đơn giản và sử dụng Akka HTTP để tạo ra REST API nhằm giúp người dùng có thể truy xuất được thông tin cần thiết.

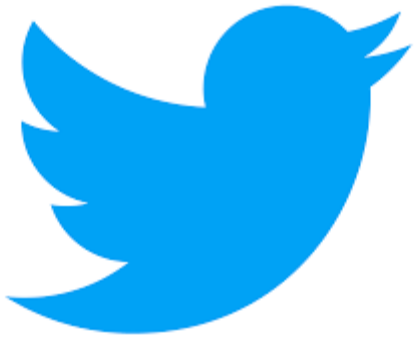
2.4. Mạng xã hội Twitter

2.4.1. Giới thiệu về Twitter

Twitter là một dịch vụ mạng xã hội trực tuyến miễn phí cho phép người sử dụng đọc, nhắn và cập nhật các mẩu tin nhỏ gọi là tweets, một dạng tiểu blog. Những mẫu tweet được giới hạn tối đa 280 ký tự được lan truyền nhanh chóng trong phạm vi nhóm bạn của người nhắn hoặc có thể được trưng rộng rãi cho mọi người. Thành lập từ năm 2006, Twitter đã trở thành một hiện tượng phổ biến toàn cầu. Những tweet có thể chỉ là dòng tin vặt cá nhân cho đến những cập nhật thời sự tại chỗ kịp thời và nhanh chóng hơn cả truyền thông chính thống. Twitter Inc. được đặt ở San Francisco và có hơn 35 công ty khắp thế giới.

Giới hạn về độ dài của tin nhắn: 280 ký tự, có tính tương thích với tin SMS (Short Message Service), mang đến cho cộng đồng mạng một hình thức tốc ký đáng chú ý, đã được sử dụng rộng rãi đối với SMS. Giới hạn về ký tự cũng giúp thúc đẩy các dịch vụ thu gọn địa chỉ website như tinyurl, bit.ly và tr.im, hoặc các dịch vụ nội dung tên miền như là Twitpic và NotePub nhằm thu thập các thông tin đa phương tiện và những đoạn dài hơn 280 ký tự. Hiện nay Twitter đã hỗ trợ người dùng đăng các Tweet dưới dạng đoạn hội thoại, đăng ảnh, video, ảnh động, và tính năng cập nhật Khoảnh khắc.

Twitter đã được tạo ra tháng 3 năm 2006 bởi Jack Dorsey , Evan Williams , Biz Stone và Noah Glass và hoạt động vào tháng 7 năm 2006. Twitter có trụ sở chính tại San Francisco và đã có hơn 25 văn phòng trên toàn thế giới. Tính đến tháng 5 năm 2015, Twitter đã có hơn 500 triệu người dùng, trong đó có hơn 302 triệu người hoạt động thường xuyên. Twitter cũng được xem như là SMS của Internet.



Hình 2. 1: Biểu tượng đặc trưng của Twitter

Hiện nay Facebook, Pinterest và Twitter là một trong những mạng xã hội phổ biến nhất trên thế giới hiện nay. Ngoài ra còn có những mạng xã hội khác như: LinkedIn, Google Plus, Tumblr, ... Mỗi nền tảng đều có ưu và nhược điểm khác nhau.

2.4.2. Twitter API cho nhà phát triển

API Twitter là một giao diện mà thông qua đó một trang web hoặc một ứng dụng có thể tương tác với Twitter. Nó cho phép truy cập vào chính các tính năng của nền tảng, chẳng hạn như đăng tweet, tweet lại và tìm tweet có chứa một từ cụ thể thông qua một trang web. Để đào dữ liệu từ Twitter, người dùng cần phải đăng ký nhận API với Twitter để được hỗ trợ tính năng này.

API Twitter người dùng nhận được từ nhà phát triển Twitter sẽ bao gồm 4 khóa: CONSUMER_KEY,CONSUMER_SECRET,OAUTH_TOKEN,OAUTH_TOKEN_SECRET. Một trong những yếu tố cơ bản trong Twitter là một tweet. API Twitter cho bạn biết những gì bạn có thể làm với các tweet: tìm kiếm các tweet, tạo một tweet, yêu thích một tweet. Nó cũng cho bạn biết làm thế nào để thực hiện những hành động này. Để tìm kiếm các tweet, bạn cần chỉ định tiêu chí tìm kiếm của mình: thuật ngữ hoặc hashtag để tìm kiếm, định vị địa lý, ngôn ngữ, v.v

CHƯƠNG 3: KHAI THÁC VÀ PHÂN TÍCH DỮ LIỆU

3.1. Quy trình phân tích

3.1.1. Đặt vấn đề

Cryptocurrency hay tiền mã hóa là một lĩnh vực được đông đảo người quan tâm và tìm hiểu. Twitter là một trong những mạng xã hội phổ biến nhất hiện nay, là lựa chọn của hàng triệu người dùng trên thế giới, vì thế nó sở hữu khối lượng thông tin cực kỳ lớn về các lĩnh vực mà người dùng chia sẻ thông qua Twitter. Vì thế việc nghiên cứu về mức độ quan tâm của người dùng đối với lĩnh vực tiền mã hóa, đặc biệt là trên nền tảng mạng xã hội Twitter là cần thiết cho việc tìm hiểu nhu cầu, mức độ phổ biến của từng loại tiền. Từ đó, ngay cả người sử dụng cũng như các nhà cung cấp nền tảng sử dụng tiền mã hóa sẽ nắm bắt được xu hướng thị trường để đưa ra các quyết định đúng đắn cho bản thân, doanh nghiệp. Nhận thấy sự cần thiết đó, chúng tôi đã tiến hành bài phân tích cơ bản về mối quan tâm của người dùng đến lĩnh vực này, đây sẽ là tài liệu tham khảo và căn cứ để các cá nhân, doanh nghiệp tiếp cận dễ dàng hơn trong thị trường này, và là kết quả để phục vụ cho các hương hướng phát triển trong tương lai.

3.1.2. Phát biểu bài toán thực nghiệm

Bài toán nghiên cứu mức độ quan tâm của người dùng đối với các loại tiền mã hóa trên nền tảng mạng xã hội Twitter

- Input: tập dữ liệu của người dùng về Cryptocurrency được crawl từ Twitter
- Output: mức độ quan tâm của người dùng như độ thảo luận của các đồng tiền mã hóa phổ biến nhất, các khu vực trên thế giới có mức độ thảo luận lớn nhất, ...

3.2. Đề xuất giải pháp

Để giải quyết vấn đề cũng như thực hiện đề tài "Phân tích dữ liệu về mức độ quan tâm tiền mã hóa trên mạng xã hội Twitter" chúng tôi đã thực hiện phân tích tập dữ liệu được crawl từ mạng xã hội Twitter, cụ thể giải pháp được thực hiện như sau:

- Thực hiện thu thập dữ liệu: đào dữ liệu realtime trong 7 ngày dựa trên tập từ khóa về các loại tiền mã hóa chính mà nhóm đưa ra trên mạng xã hội Twitter. Code được viết trên nền tảng Java chạy bằng Eclipse, dữ liệu đào về được lưu trực tiếp vào file excel dạng bảng với 4 cột: Content, Place, User Location và Hashtag. 4 thành viên phụ trách đào mỗi người 6 giờ/ ngày để đảm bảo đào hết các khoảng thời gian trong ngày.

- Thực hiện phân tích tập dữ liệu: Sử dụng ngôn ngữ Python và Google Colaboratory. Đề tài được chia thành 4 phần phân tích chính:
    - Phân tích mức độ quan tâm theo các khu vực
    - Phân tích đánh giá sentiment theo các loại tiền
    - Phân tích mức độ quan tâm các loại tiền theo cột Content của Tweets
    - Phân tích mức độ quan tâm các loại tiền theo cột Hashtags của Tweets
- Thực hiện đánh giá kết quả đã phân tích được, chỉ ra ưu nhược điểm của giải pháp và đề xuất hướng phát triển trong tương lai.

3.3. Thu thập dữ liệu trên mạng xã hội Twitter

3.3.1. Mô tả dữ liệu cần thu thập

Để thu thập dữ liệu cần thiết để phục vụ cho nghiên cứu, nhóm cần phải chọn lọc ra các đặc điểm, thuộc tính thực sự cần thiết trong số rất nhiều trường dữ liệu có trong dữ liệu mà Twitter API cùng cấp. Sau cùng nhóm đã chọn ra 3 nhóm dữ liệu quan trọng và cần thiết nhất bao gồm: nội dung bài đăng, từ khóa và hashtag, vị trí.

Đối với nội dung bài đặng thu thập được, nhóm sẽ thực hiện các phương pháp lọc và làm sạch, sau đó đem đi thực hiện các phân tích sentiment để đánh giá điểm tích cực tiêu cực của từng đồng tiền mã hóa trên thị thường, sự xuất hiện của các từ, cụm từ thường xuyên trong bài tweet. Các tập dữ liệu từ khóa và hashtag sẽ hữu ích trong việc tổng kết và đánh giá mức độ quan tâm của từng loại tiền mã hóa thông qua tần suất xuất hiện của chúng. Và cuối cùng là dữ liệu về vị trí của những người dùng Twitter trên tài khoản của họ, đây là những người dùng có bài đăng Twitter bao hàm từ khóa hoặc hashtag mà nhóm thu thập. Từ đây, nhóm có thể truy xuất được các khu vực có mức độ quan tâm cao đối với nhóm tiền mã hóa này trên thế giới.

3.3.2. Quy trình thu thập dữ liệu

- Bước 1: Xây dựng model (class) Tweet trong Java

Nhóm sử dụng phương pháp hướng đối tượng trong lập trình để phân tích và xây dựng lớp đối tượng. Trong đó Tweet là khái niệm trừu tượng chỉ một bài đăng trên mạng xã hội Twitter. Một bài Tweet có rất nhiều thuộc tính đi kèm, nhưng sau khi chọn lọc và mô tả các dữ liệu cần thu thập, nhóm xây dựng cho model Tweet 4 thuộc tính chính sau:

Thuộc tính	Ý nghĩa	Kiểu dữ liệu
Text	Nội dung của bài Tweet được đăng	String
Place	Vị trí bài Tweet được đăng	String
User Location	Vị trí của người dùng cập nhật trên trang cá nhân	String
Hashtag	Danh sách các hashtag xuất hiện trong bài đăng	Arraylist String

Bảng 3. 1: Xây dựng thuộc tính cho đối tượng Tweet

*Tweet.java | WriteExcelTweet.java | App.java

```java
package nhathan.com.ExcelProject;

import java.util.ArrayList;

public class Tweet {
    private String text;
    private String place;
    private String userLocation;
    private ArrayList<String> hashtag;


    public Tweet(String text, String place, String userLocation, ArrayList<String> hashtag) {
        super();
        this.text = text;
        this.place = place;
        this.userLocation = userLocation;
        this.hashtag = hashtag;
    }
    public Tweet() {
        super();
    }


    public String getText() {
        return text;
    }
    public void setText(String text) {
        this.text = text;
    }
    public String getPlace() {
        return place;
    }
    public void setPlace(String place) {
        this.place = place;
    }
    public String getUserLocation() {
        return userLocation;
    }
    public void setUserLocation(String userLocation) {
        this.userLocation = userLocation;
    }
}
```

Hình 3. 1: Model Tweet trong Java

- Bước 2: Xây dựng model (class) WriteExcelTweet

Đây là model được xây dựng để thực hiện thao tác ghi chép dữ liệu thu thập được vào Excel với phương thức WriteToExcel do chính nhóm tự xây dựng như sau:

```
    public void writeToExcel(Tweet data, String fileName) {
        try {
//          Create top row with column heading
            String[] columnHeadings = {"Content", "Place", "User Location", "Hashtags"};
            //Style font header
            Font headerFont = workbook.createFont();
            headerFont.setBold(true);
            headerFont.setFontHeightInPoints((short)12);
            headerFont.setColor(IndexedColors.BLACK.index);
            //Style cell header
            CellStyle headerStyle = workbook.createCellStyle();
            headerStyle.setFont(headerFont);
            headerStyle.setFillPattern(FillPatternType.SOLID_FOREGROUND);
            headerStyle.setFillForegroundColor(IndexedColors.GREY_25_PERCENT.index);
            //Create header row
            Row headerRow = sh.createRow(0);
            //Iterate over column heading to create column
            for(int i=0;i<columnHeadings.length;i++) {
                Cell cell = headerRow.createCell(i);
                cell.setCellValue(columnHeadings[i]);
                cell.setCellStyle(headerStyle);
            }
            //Fill data
            Tweet tweet_data = data;
            CreationHelper creationHelper= workbook.getCreationHelper();
            CellStyle dateStyle = workbook.createCellStyle();
            dateStyle.setDataFormat(creationHelper.createDataFormat().getFormat("MM/dd/yyyy"));
            //Import data to Excel
            Row row = sh.createRow(this.rowNum);
            row.createCell(0).setCellValue(tweet_data.getText());
            row.createCell(1).setCellValue(tweet_data.getPlace());
            row.createCell(2).setCellValue(tweet_data.getUserLocation());
            row.createCell(3).setCellValue(tweet_data.getHashtag().toString());

            FileOutputStream fileOut = new FileOutputStream(fileName);
            workbook.write(fileOut);
            fileOut.close();
        //  workbook.close();
            System.out.println("Completed");

        } catch (Exception e) {
            e.printStackTrace();
```

Hình 3. 2: Hàm ghi dữ liệu vào Excel trong Java

- Bước 3: Xây dựng hàm main để chạy chương trình đào real time

Sau khi đã hoàn thành việc xây dựng các model, nhóm thực hiện xây dựng hàm để chạy chương trình trong hàm main của Java. Trong đó bước đầu để thực hiện được các thao tác đào dữ liệu nhóm cần sự hỗ trợ của các thư viện sau đây:

```
import twitter4j.HashtagEntity;
import twitter4j.StallWarning;
import twitter4j.Status;
import twitter4j.StatusDeletionNotice;
import twitter4j.StatusListener;
import twitter4j.TwitterStream;
import twitter4j.TwitterStreamFactory;
import twitter4j.conf.ConfigurationBuilder;
```

Hình 3. 3: Import thư viện hỗ trợ đào dữ liệu từ Twitter API

Tiếp theo đó ta cần cấp quyền truy cập đến Twitter API thông qua các key được cung cấp từ Twitter Developer:

```
ConfigurationBuilder cb = new ConfigurationBuilder();
  cb.setDebugEnabled(true)
    .setOAuthConsumerKey("YMwS4k9D8fgCtIhvx1bn7irKS")
    .setOAuthConsumerSecret("RPRVQzyTXcsOIQjWa2ti7On6R2HewEfgcPCYnYhkdQhfH0MjKq")
    .setOAuthAccessToken("1362444246339383300-LJOdqgDAx3sW5g0Cv3CEAq7b9ZnmlF")
    .setOAuthAccessTokenSecret("dGqqFX3z9UBNhCILTF8YZNF7Zt3R8DHi5xDamx8W1vzx8");
```

Hình 3. 4: Cấu hình và cấp quyền truy cập Twitter API

Cuối cùng nhóm xây dựng danh sách các từ khóa cần đào và sử dụng các model đã xây dựng trước đó để hoàn thành chương trình:

```
String [] lstKeyWord= {"bitcoin","ethereum","binance","dogecoin","cardano","tether","xrp",
        "polkadot","bitcoin cash","litecoin"};
WriteExcelTweet writeTweet = new WriteExcelTweet();
StatusListener listener = new StatusListener() {
    public void onStatus(Status status) {
        long end = System.currentTimeMillis();
        if(end - start >= 3600000*4) System.exit(0);
        String text = status.getText();
        if(!text.startsWith("RT") ) {
            for(String keyword: lstKeyWord) {
                // Chỉnh từ khóa ở đây
                if(text.toLowerCase().contains(keyword)) {
                    int favoriteCount = status.getFavoriteCount();
                    int retweetCount = status.getRetweetCount();
                    String place = checkPlace(status);
                    String userLocation = checkUserLocation(status);
                    ArrayList<String> hashtag = checkHashtagEntities(status);
                    Tweet newTweet = new Tweet(text, place, userLocation, hashtag);
                    // Chỉnh tên file xuất ra ở đây
                    writeTweet.writeToExcel(newTweet, "crypto_tweet4.xlsx");
                    writeTweet.increaseRowNum();
                    System.out.println(newTweet.toString());
                    System.out.println(status);
                    System.out.println("Complete add row " + writeTweet.getRowNum());
                    System.out.println(status.getHashtagEntities()[0].getText());
                }
            }
        }
    }
```

Hình 3. 5: Chương trình đào dữ liệu từ Twitter API sử dụng các model, chức năng đã xây dựng

- Bước 4: Mở kết quả thu được từ chương trình

	A	B	C	D
1	**Content**	**Place**	**User Location**	**Hashtags**
2	Every Friday, we publish our weekly #Cardano development update. So for the lowdown on		Netherlands	[Cardano]
3	@luis_adaime @elonmusk @moss_earth @MCo2token Elon, don't worry about Bitcoin		Rio de Janeiro, Brasil	[]
4	@iamtheidentity @IOHK_Charles doesn't control @Cardano you fool!		California, USA	[]
5	Another stake increase by one of our existing delegators. Thank you for supporting OZZY.			[Cardano, ada]
6	@CardanoDan Fundamentals.		Vancouver, British Columbia	[cardano]
7	Investing in Bitcoin seriously changed my life. It was the best thing I've ever done. Wow.			[]
8	@BitcoinKralice Logical Traders		İstanbul, Türkiye	[]
9	@BTCTN #Bitcoin			[Bitcoin, ETHEREUM]
10	@BTCTN #Bitcoin			[Bitcoin, ETHEREUM]
11	I checked this DEFI AMM exchange platform called @EmiSwap. They are providing			[]
12	@carmineborges11 @shellaidell5721 @sarantrena6677			[MaxxiCoin, BSC,

Hình 3. 6: Kết quả thu được từ chương trình (file excel)

3.3.3. Làm sạch dữ liệu

Sau khi hoàn tất thu thập dữ liệu, nhóm nhận thấy dữ liệu còn nhiều chỗ chưa sạch, còn tồn tại các kí tự đặc biệt, icon thậm chí là html trong nội dung đào được vì thế nhóm đã sử dụng ngôn ngữ Python trên nền tảng Google Colab để xây dựng các hàm hỗ trợ làm sạch dữ liệu sau:

```
import json
import re

def strip_emoji(text):
    regrex_pattern = re.compile(pattern = "["
                                u"\U0001F600-\U0001F64F"  # emoticons
                                u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                                u"\U0001F680-\U0001F6FF"  # transport & map symbols
                                u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                                u"\U00002500-\U00002BEF"  # chinese char
                                u"\U00002702-\U000027B0"
                                u"\U00002702-\U000027B0"
                                u"\U000024C2-\U0001F251"
                                u"\U0001f926-\U0001f937"
                                u"\U00010000-\U0010ffff"
                                u"\u2640-\u2642"
                                u"\u2600-\u2B55"
                                u"\u200d"
                                u"\u23cf"
                                u"\u23e9"
                                u"\u231a"
                                u"\ufe0f"  # dingbats
                                u"\u3030"
                                "]+", flags=re.UNICODE)
    return regrex_pattern.sub(r'',text)
```

Hình 3. 7: Hàm lọc ký tự đặc biệt và Icon

```
import os
import sys
import json
import feedparser
from bs4 import BeautifulSoup
from nltk import clean_html


def cleanHtml(html):
    if html == "":
      return ""
    return BeautifulSoup(html, 'html5lib').get_text()
```

Hình 3. 8: Hàm lọc html khỏi văn bản

Sau khi lọc phần Content của dữ liệu, nhóm tiếp tục tiến hành lọc dữ liệu User Location và lấy theo tên quốc gia của từng user location sử dụng Named Entity Recognition với NLTK,

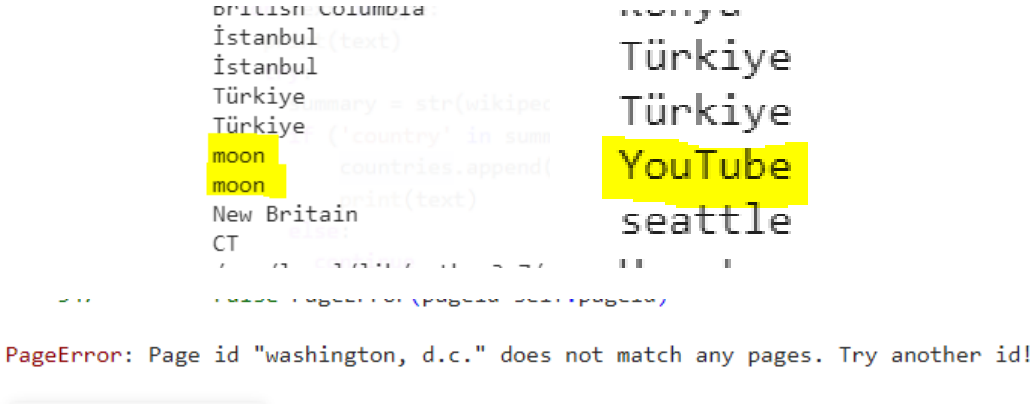
Spacy cùng với đó là wikipedia, đây là những thư viện vô cùng mạnh mẽ đã được giới thiệu ở trên:

```
import spacy
import wikipedia
spacy.cli.download("en_core_web_lg")

txt=""
for text in arr:
  if(str(text)!="nan"):
    txt+= str(text)+", "
doc = nlp(txt)
gpe=[]
for ent in doc.ents:
  if(ent.label_=="GPE"):
    gpe.append(ent.text)
# gpe
countries = []
for text in gpe:
  print(text)
  try:
    summary = str(wikipedia.summary(text, sentences=1))
    if ('country' in summary):
        countries.append(text)
        print(text)
    else:
      continue
  except wikipedia.DisambiguationError as e:
    continue
countries
```

Hình 3. 9: Hàm lọc location từ văn bản

Sau khi chạy một lượt qua tập dữ liệu, nhóm nhận thấy có một số vấn đề khi dùng code như sau:

- Sau khi thu được các entity GPE (Country, state,...) thì không thể tách Country ra một cách độc lập phải phụ thuộc vào các thư viện khác cụ thể ở đây là wikipedia.
- Kết quả thu được sau khi search trên thư viện wikipedia chưa thực sự đúng và đầy đủ, ngoài ra còn một số khó khăn như một số nước người dùng nhập vào viết tắt sẽ gây hiểu nhầm cho công cụ wikipedia không thể nhận ra đó là một nước.
- Ngoài ra, trong bộ dữ liệu có một số người dùng cần phải có UTF-8 nhưng thư viện này lại không hỗ trợ mạnh về mảng này.
- Dữ liệu thu được không hoàn toàn sạch, và đạt yêu cầu của nhóm, một số kết quả thu được như sau:



Hình 3. 10: Một số trường dữ liệu do lỗi nhập liệu từ User

Do đó, nhóm quyết định thực hiện bước làm sạch dữ liệu lần 2 bằng cách lọc thủ công các quốc gia thông qua excel để đảm bảo được chất lượng đầu ra được đúng nhất.

3.4. Phân tích và đánh giá dữ liệu

3.4.1. Đọc file dữ liệu đã thu thập

- Import các thư viện liên quan:

```
[88] import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     from collections import Counter
     from wordcloud import WordCloud
     import seaborn as sns
     import re
     sns.set()
     #!pip install feedparser
     #!pip install nltk
     from nltk.corpus import stopwords
     import nltk
     nltk.download('stopwords')
     nltk.download('punkt')
     from nltk.tokenize import word_tokenize
     from gensim.parsing.preprocessing import remove_stopwords
     import os
     import sys
     import feedparser
     from bs4 import BeautifulSoup
     from nltk import clean_html
```

Hình 3. 11: Các thư viện Python hỗ trợ nghiên cứu dữ liệu

- Đọc file dữ liệu được đào bằng ngôn ngữ Java trên Eclipse:

```
[51] #đọc file excel
     # Tải file có tên FinalProject_RawData.xlsx lên Google Colab
     data = pd.read_excel("FinalProject_RawData.xlsx")
```

```
data
```

	Content	Place	User Location	Hashtags
0	Every Friday, we publish our weekly #Cardano d...	NaN	Netherlands	[Cardano]
1	@luis_adaime @elonmusk @moss_earth @MCo2token ...	NaN	Rio de Janeiro, Brasil	[]
2	@iamtheidentity @IOHK_Charles doesn't control ...	NaN	California, USA	[]
3	Another stake increase by one of our existing ...	NaN	NaN	[Cardano, ada]
4	@CardanoDan Fundamentals.\n Technology.\n Phil...	NaN	Vancouver, British Columbia	[cardano]
...	...	...	...	...
6213	Shout out to the homie @NettieBella Go check o...	NaN	US	[Crypto, XRPtheStandard, XRP, 0Doubt]
6214	Free Ethereum - Earn $65 free eth in 5 minutes...	NaN	NaN	[ethereum, freeeth, geteth]
6215	Free Bitcoin Mining site, friends! Don't miss....	NaN	Hoshiarpur, India	[Bitcoin, free, cryptocurrency, Bitcoins]
6216	Owning #Bitcoin has RULES! What is the first r...	NaN	United States	[Bitcoin]
6217	@WOLF_Financial Bought some $pussy because it ...	NaN	NaN	[doge, dogecoin, dogekiller]

6218 rows × 4 columns

Hình 3. 12: Đọc file dữ liệu excel

- Trích xuất một số thông tin tổng quan về file dữ liệu được sử dụng trong nghiên cứu:

```
[5]  # Thông tin về số dòng dữ liệu đào được
     data.shape
```

```
(6218, 4)
```

```
[6]  # Thông tin về cột, null, non-null, phân loại dữ liệu
     data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6218 entries, 0 to 6217
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Content        6218 non-null   object
 1   Place          51 non-null     object
 2   User Location  3084 non-null   object
 3   Hashtags       6218 non-null   object
dtypes: object(4)
memory usage: 194.4+ KB
```

Hình 3. 13: Thông tin của file dữ liệu

Tập dữ liệu sử dụng trong nghiên cứu có 4 cột với số lượng 6218 dòng dữ liệu từ Twitter. Trong đó:

- Cột Content có 6218 dòng có dữ liệu với kiểu Object, chiếm 100% trên tổng số
- Cột Place chỉ có 51 dòng có dữ liệu với kiểu Object, chiếm khoảng 0.8% trên tổng 6218 dòng dữ liệu

- Cột User Location có 3084 dòng dữ liệu với kiểu Object, chiếm gần 50% trên tổng số dòng
- Cột Hashtags có kiểu dữ liệu là Object

3.4.2. Đánh giá mức độ quan tâm theo khu vực

3.4.2.1. Mô tả

Dựa vào mức độ quan tâm của khu vực chúng ta sẽ có thể nhận định được những nước nào có được sự quan tâm của tiền mã hóa. Từ đó đưa ra nhận định tại sao các nước này lại có sự quan tâm nhiều và bàn luận nhiều, trong khi có những nước chưa hiện diện trong phân tích. Trong mục này, những nhận định phân tích sẽ dựa trên trường "User Location" của dữ liệu đã thu thập.

3.4.2.2. Quy trình xử lý

- Từ tập dữ liệu đã được lọc chúng ta tiến hành lấy mẫu là những bài đăng có kèm dữ liệu về "User Location".
- Từ mẫu thử chúng ta tiến hành đếm các bài post dựa trên vị trí, và gom cụm theo biến User Location.
- Tiến hành phân tích, nhận định.

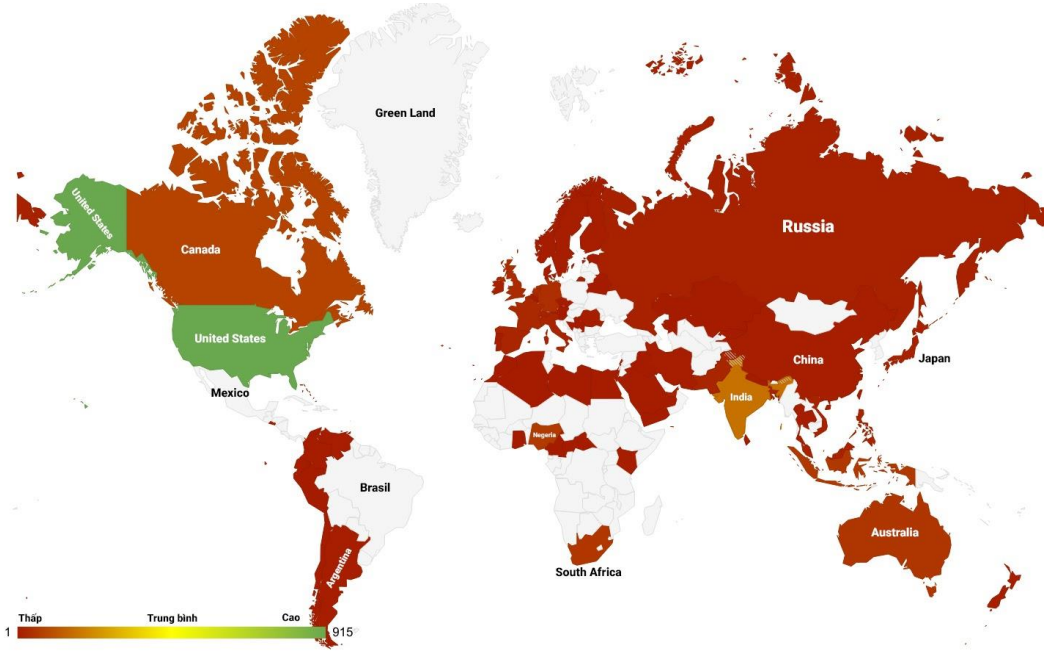
3.4.2.3. Đánh giá kết quả

Sau khi bỏ các dòng dữ liệu không có "User Location" chúng ta sẽ có được 2210 dòng dữ liệu gồm những post vừa có "Content" và "User Location".

	User Location	Post
1	United States	915
2	India	173
3	England	146
4	Canada	82
5	Nigeria	64
6	Indonesia	59
7	South Africa	56
8	Australia	54
9	Türkiye	52
10	Germany	40
11	United Kingdom	32
12	Netherlands	31
13	France	27
14	Brasil	23
15	Spain	22

Hình 3. 14: Thống kê số lượng bài Tweet theo khu vực

Từ những dữ liệu đã thu thập, nhóm sẽ đưa ra các phân tích theo khu vực để cho thấy sự tương quan giữa khu vực người dùng đang sinh sống với những đồng tiền mã hóa, so sánh với mức độ sở hữu tiền điện tử ở từng khu vực và có được biểu đồ như sau:



Hình 3. 15: Biểu đồ trực quan hóa vị trí địa lý trên số lượng bài Tweet

Theo biểu đồ trên chúng ta có thể thấy khu vực có sự bàn luận nhiều nhất đối những đồng tiền số là United States (Hợp chủng quốc Hoa Kỳ) tiếp theo đó là India (Ấn Độ) và Nước Anh (England). Trong khi đó, nước có tỉ lệ bàn luận ít nhất trong tập dữ liệu thu được khi không tính các nước không thu thập được dữ liệu là những nước như Áo, Iraq, Lybia,...

Kết quả này có một sự tương quan nhất định với số người sở hữu những đồng tiền mã hóa ở các nước:

Country	Number of crypto owners	Percentage of the population
USA	27,491,810	8.31%
Russia	17,379,175	11.91%
Nigeria	13,016,341	6.31%
Vietnam	5,961,684	6.12%
Ukraine	5,565,881	12.73%
Kenya	4,580,760	8.52%
South Africa	4,215,944	7.11%
Bangladesh	3,742,571	2.27%
Thailand	3,629,713	5.20%
United Kingdom	3,360,591	4.95%

Hình 3. 16: Thống kê số người sở hữu tiền mã hóa trong 2021 (Nguồn: Triple A)

Có thể thấy, những nước có số lượng tham gia vào thị trường tiền ảo lớn sẽ có sự quan tâm tương đối và thể hiện sự quan tâm đó trên mạng xã hội, có sự tương tác lẫn nhau. Điển hình có thể kể đến những tin tức trong thời gian gần đây (tính đến 19/05/2021) sự rời bỏ của các nhà đầu tư lớn hay còn gọi là "Cá voi" đang có tác động rất lớn vào giá trị của các đồng tiền này, điểm đáng chú ý ở đây là những "cá voi" sẽ đăng tải các thông tin phát ngôn cá nhân về tiền ảo và trong những ngày kế tiếp sau đó giá trị của các đồng tiền này liên tục tăng trong thời gian dài hoặc giảm ngay trong thời gian ngắn sau đó.

3.4.3. Đánh giá điểm sentiment theo từng loại coin

3.4.3.1. Mô tả

Sau khi thu được file dữ liệu ban đầu, ta tiến hành làm sạch dữ liệu trước khi đưa vào phân tích. Sau khi làm sạch, nhóm sẽ phân tích đánh giá sentiment theo từng loại tiền mã hóa đã chọn nghiên cứu. Phân tích nội dung của từng tweet của 10 loại tiền phổ biến nhất dựa vào dữ liệu text của cột "Content" của dữ liệu. Điểm đánh giá sentiment sẽ được phân tích dựa trên nội dung câu từ được viết trong từng dòng tweet ở cột "Content".

3.4.3.2. Quy trình xử lý

- Import bộ thư viện cần dùng: nltk, numpy

```
[23] # pip install nltk
     import nltk
     nltk.download('vader_lexicon')

     import numpy as np
     from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
/usr/local/lib/python3.7/dist-packages/nltk/twitter/__init__.py:20: UserWarning: The twython library has not been installed. Some functionality f
  warnings.warn("The twython library has not been installed. "
```

Hình 3. 17: Thêm bộ thư viện nltk vào sử dụng

- Làm sạch dữ liệu trước khi đưa vào phân tích. Dữ liệu sẽ được loại bỏ những ký tự không cần thiết.

```
[24] #Làm sạch dữ liệu (loại bỏ những thứ không cần thiết, chuyển thành chữ thường ...)
     arrSen=[]
     for i in data['Content']:
       i=i.lower()
       i = re.sub(r"http\S+", " ", i)
       i = re.sub(r'[^a-zA-Z0-9\n\.]', " ", i)
       i = re.sub(r'(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))*\))+(?:\(([^\s()<>]+|(
       arrSen.append(i)
     #arrSen
```

Hình 3. 18: Làm sạch dữ liệu sử dụng regular expression

- Tiến hành phân tích sentiment của tổng nội dung các tweet thu được

```
[25] compound=[]
     neg=[]
     neu=[]
     pos=[]
     analyzer = SentimentIntensityAnalyzer()
     for i in arrSen:
       compound.append(analyzer.polarity_scores(i)['compound'])
       neg.append(analyzer.polarity_scores(i)['neg'])
       neu.append(analyzer.polarity_scores(i)['neu'])
       pos.append(analyzer.polarity_scores(i)['pos'])
     dataSen= pd.DataFrame({"Content": arrSen,"compound":compound,"neg": neg,"neu": neu,"pos":pos})
     dataSen
```

	Content	compound	neg	neu	pos
0	every friday we publish our weekly cardano d...	-0.2648	0.098	0.902	0.000
1	luis adaime elonmusk moss earth mco2token ...	-0.4404	0.146	0.854	0.000

Hình 3. 19: Kết quả thu được (Bảng điểm sentiment) từ nội dung Tweet

- Tiếp theo, thực hiện tính tổng số compound cho mỗi loại tiền mã hóa

```
# tính toán tổng số Compound cho mỗi loại tiền
lstSen=[(["bitcoin", "btc"],0.0,0),(["eth","ethereum"],0.0,0),(["bnb","binance"],0.0,0),(["doge","dogecoin"],0.0,0),(["ada","cardano"],0.0,0),
    (["usdt","tether"],0.0,0),(["xrp"],0.0,0),(["dot","polkadot"],0.0,0),(["bch","bitcoin cash","bitcoincash"],0.0,0),(["ltc","litecoin"],0.0,0)]

for i in range(dataSen.shape[0]):
  num=dataSen.iloc[i]['Content']
  for j in range(len(lstSen)):
    lstSen2=list(lstSen[j])
    for k in lstSen[j][0]:
      if k in str(num):
        lstSen2[1]+=dataSen.iloc[i]['compound']
        lstSen2[2]+= 1
      break
    lstSen[j]=tuple(lstSen2)

lstSen
```

Hình 3. 20: Tính tổng điểm compound cho từng loại coin

```
[(['bitcoin', 'btc'], 321.9163000000002, 2857),
 (['eth', 'ethereum'], 105.94779999999967, 752),
 (['bnb', 'binance'], 29.168900000000008, 119),
 (['doge', 'dogecoin'], 185.56790000000015, 1384),
 (['ada', 'cardano'], 33.11670000000001, 199),
 (['usdt', 'tether'], 12.1357, 76),
 (['xrp'], 106.06889999999989, 574),
 (['dot', 'polkadot'], 18.81299999999999, 138),
 (['bch', 'bitcoin cash', 'bitcoincash'], 3.7156999999999996, 24),
 (['ltc', 'litecoin'], 10.425800000000002, 79)]
```

Hình 3. 21: Kết quả thu được (Tổng điểm compound) của từng loại coin

- Thể hiện giá trị compound theo mỗi loại tiền mã hóa dưới dạng bảng:

```
#bảng
obj=[]
comp=[]
for i in lstSen:
  obj.append(i[0])
  comp.append(i[1]/i[2])
table= pd.DataFrame({"Tiền mã hóa":obj,"Compound":comp})
table
```

	Tiền mã hóa	Compound
0	[bitcoin, btc]	0.112676
1	[eth, ethereum]	0.140888
2	[bnb, binance]	0.245117
3	[doge, dogecoin]	0.134081
4	[ada, cardano]	0.166416
5	[usdt, tether]	0.159680
6	[xrp]	0.184789
7	[dot, polkadot]	0.136326
8	[bch, bitcoin cash, bitcoincash]	0.154821
9	[ltc, litecoin]	0.131972

Hình 3. 22: Kết quả giá trị compound theo mỗi loại tiền mã hóa dưới dạng bảng

- Tổng quan đánh giá của người dùng đối với thị trường tiền mã hóa

```
# Tổng quan đánh giá của người dùng đối với thị trường tiền mã hóa
DanhGia=[0,0,0,0,0,0,0,0]
for i in range(dataSen.shape[0]):
  num=dataSen.iloc[i]['compound']
  if (num <= -0.75):
    DanhGia[0]+=1
  elif (num <=-0.5):
    DanhGia[1]+=1
  elif (num <=-0.25):
    DanhGia[2]+=1
  elif (num <=0):
    DanhGia[3]+=1
  elif (num <=0.25):
    DanhGia[4]+=1
  elif (num <=0.5):
    DanhGia[5]+=1
  elif (num <=0.75):
    DanhGia[6]+=1
  else :
    DanhGia[7]+=1
label=["-1.0 : -0.75","-0.75 : -0.5","-0.5 : -0.25","-0.25 : 0","0 : 0.25","0.25 : 0.5","0.5 : 0.75","0.75 : 1"]
table2=pd.DataFrame({"Khoảng Compound":label,"Số lượng":DanhGia})
table2
```

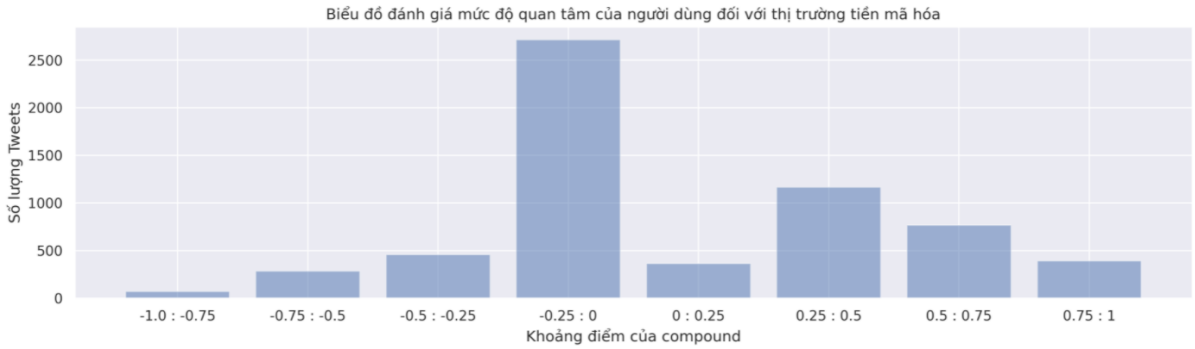
Hình 3. 23: Thực hiện đánh giá của người dùng đối với thị trường tiền mã hóa

	Khoảng Compound	Số lượng
0	-1.0 : -0.75	70
1	-0.75 : -0.5	286
2	-0.5 : -0.25	457
3	-0.25 : 0	2715
4	0 : 0.25	365
5	0.25 : 0.5	1167
6	0.5 : 0.75	766
7	0.75 : 1	392

Hình 3. 24: Kết quả thực hiện đánh giá của người dùng đối với thị trường tiền mã hóa

- Vẽ biểu đồ đánh giá mức độ quan tâm của người dùng đối với thị trường tiền mã hóa

```
# Biểu đồ đánh giá mức độ quan tâm của người dùng đối với thị trường tiền mã hóa
import matplotlib.pyplot as plt
x = table2['Khoảng Compound']
y = table2['Số lượng']
plt.figure(figsize=(16,4), dpi=300)
plt.bar(x, y, align='center', alpha=0.5)
plt.plot()

plt.xlabel("Khoảng điểm của compound")
plt.ylabel("Số lượng Tweets")
plt.title("Biểu đồ đánh giá mức độ quan tâm của người dùng đối với thị trường tiền mã hóa")
plt.show()
```



Hình 3. 25: Biểu đồ đánh giá mức độ quan tâm của người dùng đối với

3.4.3.3. Đánh giá kết quả

Dựa vào kết quả thu được ở cột dữ liệu "Content", nhóm nghiên cứu đã phân tích đánh giá về sentiment theo từng loại tiền với kết quả như sau: Bitcoin là tiền mã hóa có tổng số compound nhiều nhất (khoảng 321.92) với số tweet có liên quan là 2857 tweet. Theo sau đồng Bitcoin lần lượt là: dogecoin với điểm compound là 185.6 và số lượt tweet có liên quan là 1384. Tiếp theo là các đồng: xrp, ethereum, cardano, binance, polkadot, tether, litecoin, bitcoincash.

Nhóm thực hiện so sánh sentiment của các đồng tiền mã hóa dựa vào kết quả so sánh như sau:

- Positive sentiment: (compound score >= 0.05)
- Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
- Negative sentiment: (compound score <= -0.05)

Vì điểm compound của riêng đồng tiền mã hóa đều > 0.05. Suy ra, tất cả các đồng tiền mã hóa được nghiên cứu đều có điểm sentiment tích cực, cao nhất là đồng binance với compound là 0.245117

Nhìn chung về đánh giá tổng quan của người dùng đối với thị trường tiền mã hóa, khoảng compound có mức độ sentiment tích cực từ 0.25 đến 1 đạt số lượng 2325/ 6218 tweet. Trong đó, khoảng điểm compound từ -0.25:0 chiếm số lượng 2715, theo sau đó là khoảng 0.25:0.5 chiếm 1167.

3.4.4. Đánh giá mức độ quan tâm dựa trên tần suất lập của coin trong nội dung Tweet

3.4.4.1. Mô tả

Sau khi thu được file dữ liệu ban đầu, ta tiến hành phân tích mức độ nhắc đến của 10 loại tiền phổ biến nhất dựa vào dữ liệu text của cột “Content” của dữ liệu. Mức độ quan tâm sẽ được phân tích dựa trên số lần lặp lại của từ khóa và các từ khóa liên quan

Danh sách 10 loại tiền mã hóa phổ biến nhất hiện nay:

"Bitcoin", "Ethereum", "Binance", "Dogecoin", "Cardano", "Tether", "Xrp", "Polkadot", "Bitcoin cash", "Litecoin"

Ứng với mỗi loại tiền mã hóa, chúng tôi quy định từ khóa để phân tích tập dữ liệu chỉ bao gồm tên loại tiền và từ viết tắt của tên, danh sách được thể hiện như sau:

Bảng 3. 2: Bảng từ khóa cho từng loại coin

STT	Tiền mã hóa	Từ khóa
1	Bitcoin	`"btc","bitcoin"`
2	Ethereum	`"eth","ethereum"`
3	Binance	`"bnb","binance"`
4	Dogecoin	`"doge","dogecoin"`
5	Cardano	`"ada","cardano"`
6	Tether	`"usdt","tether"`
7	Xrp	`"xrp"`
8	Polkadot	`"dot","polkadot"`
9	Bitcoin cash	`"bch","bitcoin cash","bitcoincash"`
10	Litecoin	`"ltc","litecoin"`

3.4.4.2. Quy trình xử lý

- Bảng thống kê mức độ quan tâm theo các loại tiền mã hóa:
    - Sử dụng cột "Content" để tiến hành phân tích:

```
[ ]  #data['Content']
     dataND= pd.DataFrame({"Content": data['Content']})
     dataND
```

	Content
0	Every Friday, we publish our weekly #Cardano d...
1	@luis_adaime @elonmusk @moss_earth @MCo2token ...
2	@iamtheidentity @IOHK_Charles doesn't control ...
3	Another stake increase by one of our existing ...
4	@CardanoDan Fundamentals.\n Technology.\n Phil...
...	...
6213	Shout out to the homie @NettieBella Go check o...
6214	Free Ethereum - Earn $65 free eth in 5 minutes...
6215	Free Bitcoin Mining site, friends! Don't miss....
6216	Owning #Bitcoin has RULES! What is the first r...
6217	@WOLF_Financial Bought some $pussy because it ...

6218 rows × 1 columns

Hình 3. 26: Dữ liệu để phân tích - cột "Content"

- Làm sạch Text và loại bỏ Stop words trong nội dung của các Tweets:

```
# làm sạch text và Stopword trong nội dung tweet:
chuoi=""
for i in arrSen:
  chuoi+= i+ " ";

filtered_sentence = remove_stopwords(chuoi)
print(filtered_sentence)
```

```
friday publish weekly cardano development update. lowdown iohk s dev team bee luis adaime elonmusk moss earth mco2token elon t worry bitcoin emis
```

Hình 3. 27: Làm sạch dữ liệu - cột "Content"

- Đếm số lượt quan tâm của người dùng với các loại tiền mã hóa theo mô tả

```
[22] # đếm số lần quan tâm của người dùng đối với các loại tiền ảo dựa trên sự lặp lại của các từ khóa
     lstND=[(["btc","bitcoin"],0),(["eth","ethereum"],0),(["bnb","binance"],0),(["doge","dogecoin"],0),(["ada","cardano"],0),
          (["usdt","tether"],0),(["xrp"],0),(["dot","polkadot"],0),(["bch","bitcoin cash","bitcoincash"],0),(["ltc","litecoin"],0)]
     for i in range(len(lstND)):
       dem=0
       for j in lstND[i][0]:
         dem += filtered_sentence.count(j)
         lstDemo=list(lstND[i])
         lstDemo[1]=dem
         lstND[i]=lstDemo
     lstND

     [[['btc', 'bitcoin'], 3836],
      [['eth', 'ethereum'], 1416],
      [['bnb', 'binance'], 1865],
      [['doge', 'dogecoin'], 3420],
      [['ada', 'cardano'], 550],
      [['usdt', 'tether'], 174],
      [['xrp'], 650],
      [['dot', 'polkadot'], 212],
      [['bch', 'bitcoin cash', 'bitcoincash'], 83],
      [['ltc', 'litecoin'], 142]]
```

Hình 3. 28: Phân tích số lượt quan tâm - loại tiền mã hóa

- Thể hiện dữ liệu quan tâm của người dùng với các loại tiền mã hóa dưới dạng bảng:

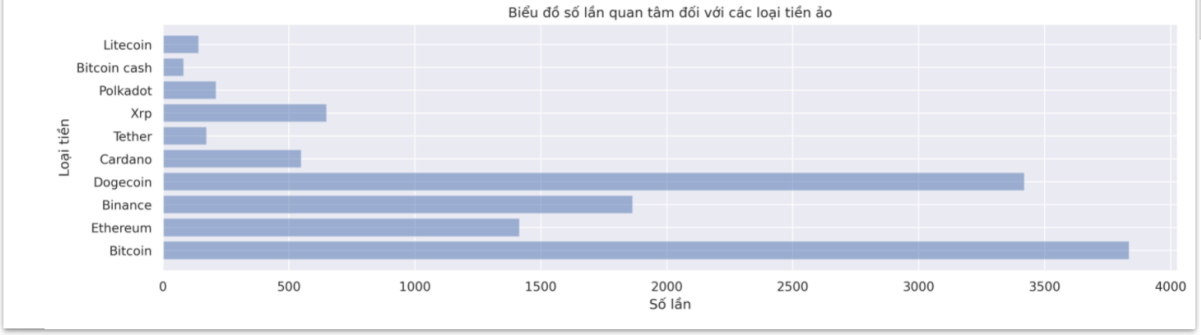
```
[55] NumType= pd.DataFrame(lstND,columns=["Keyword","Total"])
     NumType.sort_values('Total')
```

	Keyword	Total
8	[bch, bitcoin cash, bitcoincash]	83
9	[ltc, litecoin]	142
5	[usdt, tether]	174
7	[dot, polkadot]	212
4	[ada, cardano]	550
6	[xrp]	650
1	[eth, ethereum]	1416
2	[bnb, binance]	1865
3	[doge, dogecoin]	3420
0	[btc, bitcoin]	3836

Hình 3. 29: Số lần xuất hiện của từng loại coin

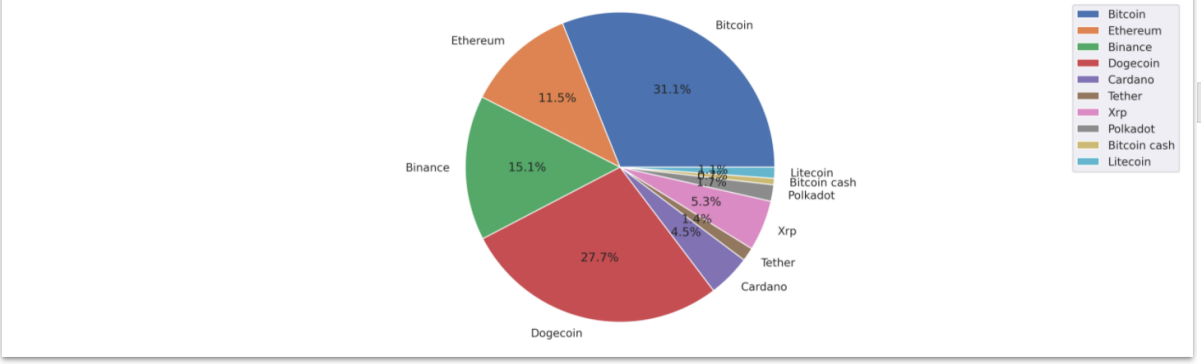
- Biểu đồ thống kê mức độ quan tâm theo các loại tiền mã hóa

```
# Bar chart
objects = ("Bitcoin","Ethereum","Binance","Dogecoin","Cardano","Tether","Xrp",
           "Polkadot","Bitcoin cash","Litecoin")
y_pos = np.arange(len(objects))
val = NumType['Total']
plt.figure(figsize=(16,4), dpi=300)
plt.barh(y_pos, val, align='center', alpha=0.5)
plt.yticks(y_pos, objects)
plt.xlabel('Số lần')
plt.ylabel('Loại tiền')
plt.title('Biểu đồ số lần quan tâm đối với các loại tiền ảo')

plt.show()
```



Hình 3. 30: Bar chart thể hiện mức độ quan tâm đến các loại tiền

```
import matplotlib.pyplot as plt

labels = objects
sizes = NumType['Total']
plt.figure(figsize=(16,5), dpi=300)
plt.pie(sizes, labels=labels,autopct='%1.1f%%')
plt.legend()
plt.axis('equal')
plt.tight_layout()
plt.show()
```

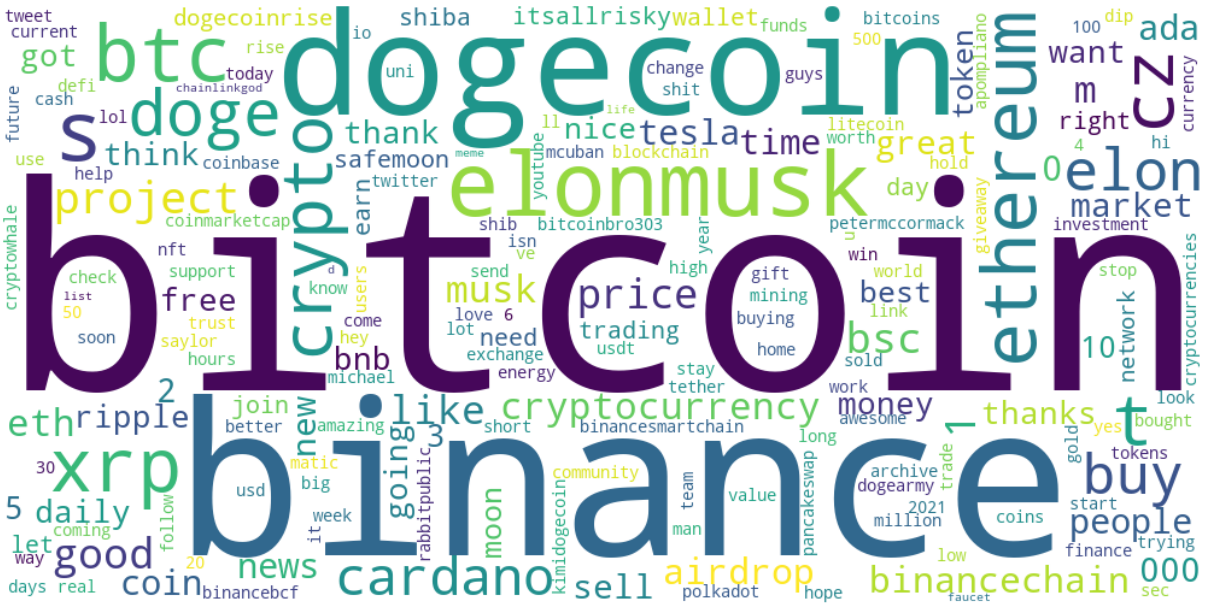


Hình 3. 31: Pie chart thể hiện mức độ quan tâm đến các loại tiền

- Vẽ Wordcloud thể hiện mức độ quan tâm theo các loại tiền mã hóa

```
# Vẽ wordcloud các cụm từ xuất hiện nhiều trong nội dung Twitter
arrContent=[]
for i in filtered_sentence.split('.'): # chuỗi sau khi làm sạch và xóa Stopwords
  for j in i.split(' '):
    arrContent.append(j.strip())
word_could_dict=Counter(arrContent)
wordcloud = WordCloud(background_color="white", width = 1000, height = 500).generate_from_frequencies(word_could_dict)
wordcloud.to_file('4_3_wordcloud.png')
plt.figure(figsize=(12,12))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
plt.close()
```

Hình 3. 32: Vẽ wordcloud các từ được xuất hiện nhiều trong nội dung Tweets



Hình 3. 33: Kết quả vẽ wordcloud các từ được xuất hiện nhiều trong nội dung Tweets

3.4.4.3. Đánh giá kết quả

Dựa vào các kết quả thu được sau khi phân tích cột "Content" của tập dữ liệu, ta nhận thấy trong thị trường Cryptocurrency, Bitcoin chiếm vị trí vô cùng quan trọng, nhận được sự quan tâm đông đảo của người dùng Twitter. Mức độ thảo luận đứng sau đồng Bitcoin (số lượt quan tâm 3836) theo thứ tự gồm có Dogecoin (3420 lượt), Binance (1865 lượt), Ethereum (1416 lượt), ... và thấp nhất là đồng Bitcoin cash chỉ với 83 lượt

➔ Bitcoin Cash (ký hiệu: BCH) là một đồng tiền điện tử ra đời vào 01/08/2017, sau khi chia tách (hard fork) khỏi Bitcoin nhưng có một số tính năng cải tiến hơn. Mặc dù có nhiều điểm giống với "người anh" Bitcoin nhưng thực sự nó là một đồng tiền độc lập với blockchain riêng, mà mức độ thảo luận của đồng Bitcoin cash này là thấp nhất so với đồng Bitcoin có độ thảo luận lớn nhất. Điều này có thể do hầu hết tính năng của Bitcoin (ví, dịch vụ thanh toán, v.v.) không hỗ trợ Bitcoin Cash và coi đây chỉ là một altcoin khác, không

phải là "Bitcoin mới". Điều này có nghĩa là việc sử dụng nó bị giới hạn trong đầu cơ trên các sàn giao dịch chứng khoán nên mức độ thảo luận còn thấp

Dựa vào kết quả thu được từ Word Cloud: trong thị trường tiền mã hóa, mức độ thảo luận của người dùng nhiều nhất ở các từ khóa như: bitcoin, binance, dogecoin, Elon Musk, crypto, xrp, ...

Từ khóa "Elon Musk" là một trong những từ khóa được nhắc đến nhiều trong thị trường tiền mã hóa. Trong khoảng thời gian chúng tôi tiến hành crawl dữ liệu từ Twitter, tỷ phú Elon Musk, tổng giám đốc công ty Tesla có tuyên bố trên nền tảng Twitter sẽ quay lưng với đồng điện tử Bitcoin khi xác nhận Tesla không nhận thanh toán khi mua xe bằng đồng tiền mã hóa này. Đây là biến cố khiến cho mức độ thảo luận của từ khóa này tăng lên đột biến trong khoàng thời gian này.

3.4.5. Đánh giá mức độ quan tâm dựa trên tần suất lập Hashtags trong nội dung Tweet

3.4.5.1. Mô tả

Dựa vào tập dữ liệu thu được, ta tiến hành phân tích theo cột "Hashtags" của bảng để chọn ra những Hashtags nào về vấn đề tiền mã hóa được phổ biến nhất trong cộng đồng người quan tâm.

3.4.5.2. Quy trình xử lý

- Bảng thống kê mức độ quan tâm theo Hashtags:
    - Tiến hành sử dụng dữ liệu từ cột "Hashtags" để phân tích dữ liệu:

```
[29] #data['Hashtags']
     dataHT= pd.DataFrame({"Hashtags": data['Hashtags']})
     dataHT
```

	Hashtags
0	[Cardano]
1	[]
2	[]
3	[Cardano, ada]
4	[cardano]
...	...
6213	[Crypto, XRPtheStandard, XRP, 0Doubt]
6214	[ethereum, freeeth, geteth]
6215	[Bitcoin, free, cryptocurrency, Bitcoins]
6216	[Bitcoin]
6217	[doge, dogecoin, dogekiller]

6218 rows × 1 columns

Hình 3. 34: Dữ liệu phân tích - cột "Hashtags"

- Tiến hành thống kê số lượt nhắc đến của mỗi Hashtag

```
[56] txt2=""
     txt3=""
     for i in dataHT['Hashtags']:
       txt2+= i +" "
     for k in txt2:
       txt3 += re.sub(r"[^a-zA-Z0-9]+",' ',k)
     arr= txt3.lower().split(" ")
     arrHT=[]
     for i in arr:
       if(i!=""):
         arrHT.append(i)

     # tổng số lượt nhắc đến của Hashtag
     dem=Counter(arrHT).most_common()[:20]
     dem
```

```
[('bitcoin', 969),
 ('dogecoin', 439),
 ('xrp', 208),
 ('ethereum', 201),
 ('binance', 189),
 ('btc', 182),
 ('bsc', 138),
 ('crypto', 138),
 ('doge', 137),
 ('cryptocurrency', 134),
 ('ripple', 110),
 ('cardano', 101),
 ('airdrop', 73),
 ('binancesmartchain', 66),
 ('bnb', 63),
```

Hình 3. 35: Thống kê số lượt nhắc đến của mỗi Hashtag

- Biểu diễn dưới dạng bảng số lượt hiển thị của Hashtags

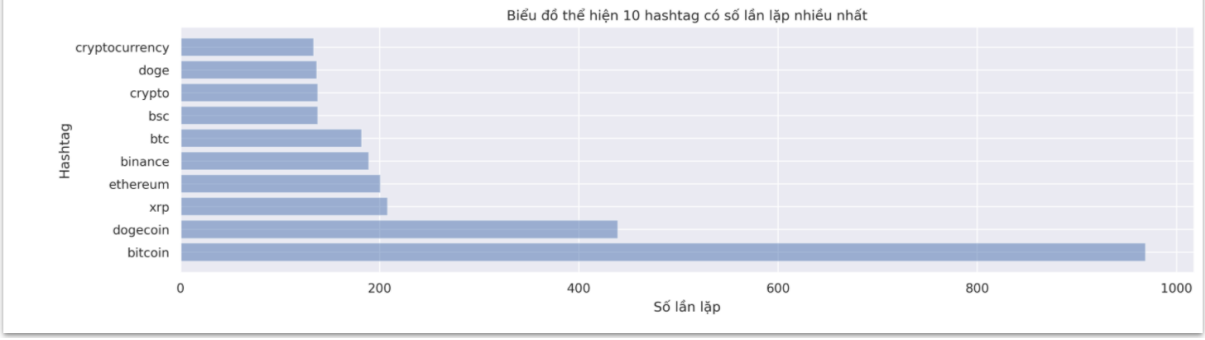
```
a=[]
b=[]
for i in dem:
  a.append(i[0])
  b.append(i[1])
dataHT2= pd.DataFrame({"Hashtags": a,"Total":b})
dataHT2.sort_values('Total')[10:] #10 hashtags có lượt nhắc đến cao nhất
```

	Hashtags	Total
9	cryptocurrency	134
8	doge	137
7	crypto	138
6	bsc	138
5	btc	182
4	binance	189
3	ethereum	201
2	xrp	208
1	dogecoin	439
0	bitcoin	969

Hình 3. 36: Thống kê số lượt nhắc đến của mỗi Hashtag dạng bảng

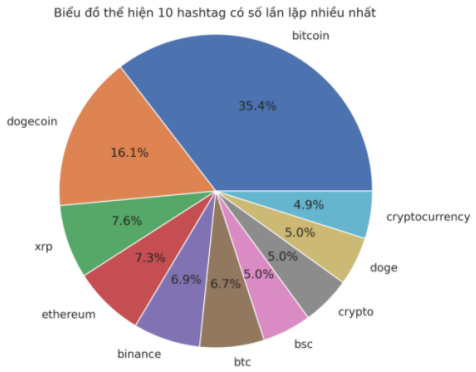
- Biểu đồ thống kê mức độ quan tâm theo Hashtags

```
# Bar chart
objects = dataHT2['Hashtags'][:10]
y_pos = np.arange(len(objects))
val = dataHT2['Total'][:10]
plt.figure(figsize=(16,4), dpi=300)
plt.barh(y_pos, val, align='center', alpha=0.5)
plt.yticks(y_pos, objects)
plt.xlabel('Số lần lặp')
plt.ylabel('Hashtag')
plt.title('Biểu đồ thể hiện 10 hashtag có số lần lặp nhiều nhất')

plt.show()
```



Hình 3. 37: Bar chart thể hiện mức độ quan tâm theo Hashtags

```
import matplotlib.pyplot as plt
labels = objects
sizes = dataHT2['Total'][:10]
plt.figure(figsize=(16,5), dpi=300)
plt.pie(sizes, labels=labels,autopct='%1.1f%%')
plt.legend()
plt.axis('equal')
plt.tight_layout()
plt.title('Biểu đồ thể hiện 10 hashtag có số lần lặp nhiều nhất')
plt.show()
```

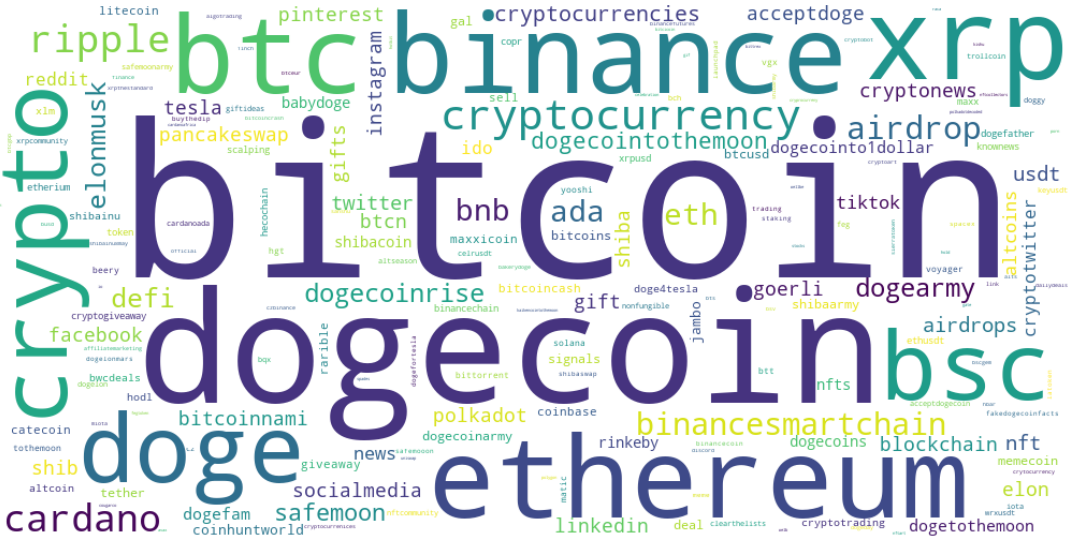


Hình 3. 38: Pie chart thể hiện mức độ quan tâm theo Hashtags

- Vẽ Wordcloud thể hiện mức độ quan tâm theo Hashtags

```
word_could_dict=Counter(arrHT)
wordcloud = WordCloud(background_color="white", width = 1000, height = 500).generate_from_frequencies(word_could_dict)
wordcloud.to_file('4_6_wordcloud.png')
plt.figure(figsize=(12,12))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
plt.close()
```

Hình 3. 39: Vẽ wordcloud số lần xuất hiện của các Hashtags



Hình 3. 40: Kết quả vẽ wordcloud số lần xuất hiện của các Hashtags

3.4.5.3. Đánh giá kết quả

Dựa vào các kết quả thu được sau khi phân tích cột "Hashtags" của tập dữ liệu, ta nhận thấy trong thị trường Cryptocurrency, mức độ thảo luận về Bitcoin vẫn chiếm vị trí vô cùng quan trọng, nhận được sự quan tâm đông đảo của người dùng Twitter để gắn Hashtag vào bài đăng trên Twitter. Tổng lượt nhắc đến của Hashtag "bitcoin" cao nhất là 969 lần, đứng sau là "dogecoin" với 439 lần, "xrp" với 208 lần,...

Thông qua các kết quả thu được, chúng ta có thể đưa ra kết luận về Bitcoin chính là đồng tiền mã hóa được người dùng thảo luận nhiều nhất trên mạng xã hội Twitter, đứng thứ 2 ở mức độ thảo luận của cả Content và Hashtag là đồng Dogecoin.

CHƯƠNG 4: KẾT LUẬN VÀ KIẾN NGHỊ

4.1. Đánh giá kết quả nghiên cứu

4.1.1. Kết quả đạt được

Từ những phân tích của bài nghiên cứu trên, có thể thấy được sự tương quan giữa tiền mã hóa và mạng xã hội. Sự tác động của Mạng xã hội Twitter đặc biệt là các nhà đầu tư lớn đã tác động lớn đến giá trị của các đồng tiền khiến cho những ngày sau đó mức độ quan tâm bàn luận trên mạng xã hội cũng nhiều và cao hơn.

Những nơi có lượng sở hữu tiền kỹ thuật số cao thường có mức bàn luận, quan tâm cao hơn các khu vực khác điển hình là Hoa Kỳ.

4.1.2. Ưu điểm

- Quá trình làm việc:
    - Hoàn thành công việc đúng tiến độ và mục tiêu đặt ra.
    - Mỗi thành viên đều tích cực làm việc để hoàn thành tốt nhiệm vụ của mình, thường xuyên hỗ trợ, góp ý để đồ án hoàn thành một cách tốt nhất.
    - Mọi người luôn biết phân chia công việc và san sẻ công việc trong trường hợp khó khăn của các thành viên.
    - Các thành viên đều tích cực đóng góp ý kiến trong các buổi họp, kiểm tra chéo báo cáo công việc của nhau để cùng nhau hoàn thiện đồ án.
    - Họp nhóm thường xuyên để có sự tương tác qua lại, tất cả thành viên đều hiểu rõ từng bước, từng phần của đồ án.
- Kết quả nghiên cứu:
    - Giải quyết được các yêu cầu đề ra của bài nghiên cứu và đem lại kết quả khả quan để phân tích.
    - Quy trình thực hiện bài toán và kết quả thu được có tính ứng dụng cao trong lĩnh vực tiền mã hóa.
    - Tận dụng được các kiến thức, công cụ đã học vào trong đồ án để khai thác tối đa dữ liệu nghiên cứu và hiệu quả thu được.

4.1.3. Nhược điểm

- Quá trình nghiên cứu: Mỗi thành viên đều có công việc/ kế hoạch học tập ngoài thời gian học tại trường nên gây khó khăn cho việc sắp xếp thời gian/ địa điểm họp. Đề tài triển khai chậm hơn dự kiến nhưng vẫn đảm bảo tiến độ công việc.
- Kết quả nghiên cứu:
    - Thời gian chạy code để thu được tập dữ liệu realtime còn ít => file dữ liệu vẫn chưa đạt được số lượng đủ lớn để đánh giá tổng quan hơn.
    - Chỉ tập trung vào dữ liệu tweet ban đầu nên những thuộc tính như retweet và favorite phát sinh sau này sẽ không được đưa vào file dữ liệu đã đào được để đánh giá và phân tích.
    - Tập dữ liệu đào được về tiền mã hóa trên mạng xã hội Twitter chịu sự ảnh hưởng của các biến cố, ở đây đề cập đến sự kiện về tỷ phú Elon Musk.
    - Chưa có phân tích theo mốc thời gian để đưa ra những sự biến động của đồng tiền trong từng quãng thời gian.
    - Dữ liệu đào được chưa được sạch, và chưa có thư viện để giải quyết triệt để, vẫn còn phải có hướng lọc thủ công.

4.2. Phương hướng phát triển đề tài

- Phát triển chức năng cập nhập realtime tập dữ liệu đã đào được về 2 thuộc tính retweet count và favorite của từng tweet để phục vụ tốt hơn việc phân tích và đánh giá đề tài
- Phát triển thêm tập từ khóa đào dữ liệu ban đầu, và tập từ khóa để phân tích từng loại tiền (Chương 3- Mục 4.3) để thu được kết quả chính xác hơn.
- Phát triển kế hoạch đào dữ liệu trong khoảng thời gian dài hơn để thu thập được nhiều dữ liệu hơn kết quả hiện tại.
- Dùng các phân tích như kiểm định các biến định tính Chi Bình Phương để thấy được sự tương quan giữa các biến định tính.
- Thu thập thêm dữ liệu theo thời gian bài đăng để đo lường mức độ quan tâm theo thời gian.

BÁO CÁO QUÁ TRÌNH LÀM VIỆC NHÓM

STT	MSSV	Họ và tên	Mức độ đóng góp
1	K184111445	Vũ Quang Huy	100%
2	K184111442	Lê Trần Giản Đơn	100%
3	K184111444	Đỗ Nguyễn Nhật Hàn	100%
4	K184111457	Phan Hồng Oanh	100%

TÀI LIỆU THAM KHẢO

[1] Joanna Jablonsk, 2021, Natural Language Processing With Python's NLTK Package

https://realpython.com/nltk-nlp-python/

[2] Prem Prakash, 2020, Extend Named Entity Recogniser (NER) to label new entities with spaCy

https://towardsdatascience.com/extend-named-entity-recogniser-ner-to-label-new-entities-with-spacy-339ee5979044

[3] Hua Shi, 2020, Data Visualization: How To Plot A Map with Geopandas in Python?

https://melaniesoek0120.medium.com/data-visualization-how-to-plot-a-map-with-geopandas-in-python-73b10dcd4b4b

[4] nicolaskruchten, 2020, Plotly Python Open Source Graphing Library Maps

https://plotly.com/python/maps/

[5] George Pipis, 2002, How To Run Sentimen George Pipist Analysis In Python Using VADER

https://predictivehacks.com/how-to-run-sentiment-analysis-in-python-using-vader/

[6] Google chart for developer

https://developers.google.com/chart

[7] Federal Trade Commission, 2021, What To Know About Cryptocurrency and Scams

https://www.consumer.ftc.gov/articles/what-know-about-cryptocurrency-and-scams

[8] Nam Ha Minh, 2019, How to Write Excel Files in Java using Apache POI

https://www.codejava.net/coding/how-to-write-excel-files-in-java-using-apache-poi

PHỤ LỤC 1: SOURCE CODE ĐỒ ÁN

Source code github: https://github.com/handnn18411c/Crypto_Mining.git

Google Drive: 1. Document - Google Drive