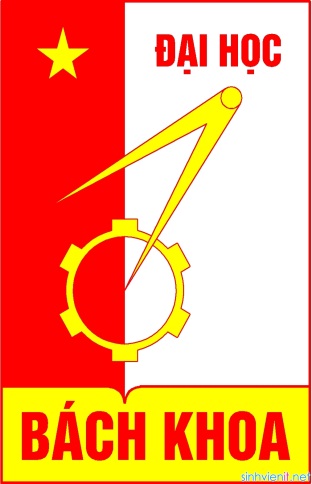
**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

-----\*\*\*-----



**BÁO CÁO BÀI TẬP LỚN**

**MÔN: KHAI PHÁ WEB**

**Đề Tài: Xác định Tin hot dựa vào dữ liệu Twitter**

***Giảng viên hướng dẫn:*** TS. Nguyễn Kiêm Hiếu

***Nhóm sinh viên thực hiện:*** Đặng Văn Hà 20151137

Phạm Huy Hoàng 20151556

Nguyễn Quốc Lâm 20152128

Nguyễn Văn Phú 20143455

Hà Nội, 05/2019

Mục Lục

[**I. Giới thiệu 4**](#_Toc9412164)

[**II. Phân tích bài toán 5**](#_Toc9412165)

[**1. Yêu cầu bài toán 5**](#_Toc9412166)

[**2. Cơ sở lý thuyết 5**](#_Toc9412167)

[**2.1 Hierarchical Clustering 5**](#_Toc9412168)

[**2.2. Bag of Words 6**](#_Toc9412169)

[**2.3. Tf-idf 7**](#_Toc9412170)

[**III. Thực hiện 9**](#_Toc9412171)

[**1. Lấy dữ liệu 9**](#_Toc9412172)

[**2. Tiền xử lý dữ liệu 9**](#_Toc9412173)

[**3. Trích chọn đặc trưng 9**](#_Toc9412174)

[**4. Phân cụm theo thứ bậc các Tweet (AHC) 10**](#_Toc9412175)

[**4.1. Tính toán khoảng cách theo cặp tweet 10**](#_Toc9412176)

[**4.2. Tính toán phân cụm thứ bậc (Hierarchy cluster) 10**](#_Toc9412177)

[**4.3. Cắt dendrogram 10**](#_Toc9412178)

[**4.4. Sắp xếp kết quả các cụm 10**](#_Toc9412179)

[**4.5. Chọn tiêu đề cho chủ đề 11**](#_Toc9412180)

[**4.6. Phân cụm lại đối với tiêu đề** 12](#_Toc9412181)

[**4.7. Lấy ra chủ đề cuối cùng** 12](#_Toc9412182)

[**IV.** **Thực nghiệm và kết quả** 13](#_Toc9412183)

[**1. Thực nghiệm** 13](#_Toc9412184)

[**2. Kết quả** 14](#_Toc9412185)

[**V.** **Tài liệu tham khảo** 16](#_Toc9412189)

1. **Giới thiệu**

Twitter là mạng xã hội nổi tiếng được dung phổ biến trên toàn thế giới. Độ ảnh hưởng và truyền bá thông tin trên mạng xã hội này rất lớn. Thậm chí những tin hay sự kiện lớn luôn được đẩy lên trước cả các báo nổi tiếng. Hiện nay, do nhận ra sự cập nhật nhanh chóng thông tin trên mạng xã hội, nên nhiều trang báo hay các trung tâm phân tích đã sử dụng dữ liệu của mạng xã hội để phân tích các xu hướng của người dùng từ đó dự đoán các xu hướng tương lai.

Đề tài này xây dựng một hệ thống xác định các xu hướng mới nhất trên mạng xã hội từ các tweet mà người dùng đăng tải trên Twitter.

1. **Phân tích bài toán**

## 1. Yêu cầu bài toán

Đầu vào của bài toán là thông tin từ các Tweet được thu thập từ Twitter. Do số lượng các Tweet là rất nhiều và gồm nhiều ngôn ngữ nên nhóm em quyết định thu thập dữ liệu thông qua Twiiter API tại khu New York, USA để thử nghiệm và giảm khối lượng tính toán. Các Tweet sử dụng tiếng Anh.

Thông qua việc sử dụng các phương pháp phân cụm, đánh giá, tính điểm để đưa ra nội dung của các Topic đang là xu hướng trong khoảng thời gian nhất định.

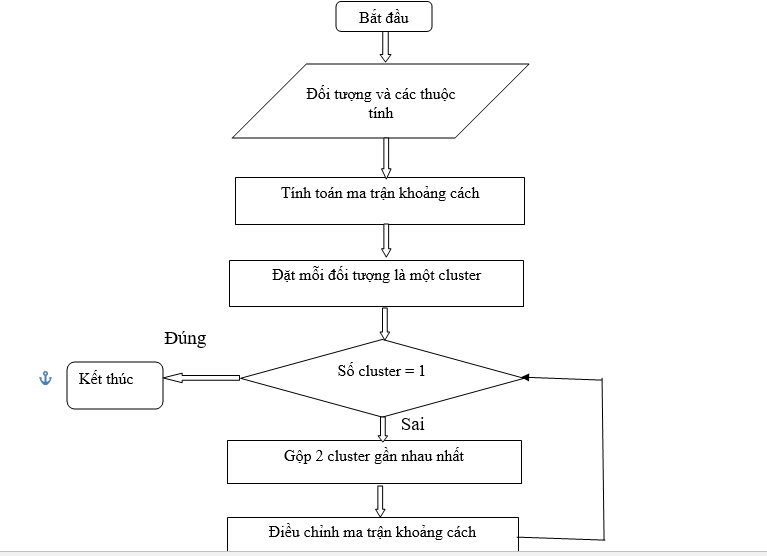
## 2. Cơ sở lý thuyết

**2.1 Hierarchical Clustering**

Hierarchical Clustering là một thuật toán học máy không giám sát được sử dụng để phân cụm các điểm dữ liệu không được gắn nhãn. Giải thuật Hierarchical Clustering thường được chia thành 2 nhóm chính:

* **Agglomerative Aapproach (bottom up approach)**: Khi bắt đầu ta coi mỗi một đối tượng cần phân cụm là một cluster và nhóm 2 đối tượng gần nhất thành một cluster. Quá trình tiếp tục cho đến khi tất cả các đối tượng được nhóm vào một cluster cuối cùng.
* **Divisive Approach (top down approach)**: Ngược lại với phương pháp trên, ban đầu chúng ta coi tất cả các đối tượng cần phân nhóm là một cluster, sau đó tiến hành phân 2 nhóm con dựa vào khoảng cách xa nhất. Công việc được thực hiện đến khi mỗi đối tượng được phân thành một cluster.

Trong bài tập này nhóm em sử dụng Thuật toán A**gglomerative Hierarchical Clustering** để phân cụm các Tweets. Thuật toán được mô tả theo sơ đồ:



Hình 1: Sơ đồ thuật toán Agglomerative Hierarchical Clustering

Trong bài toán:

* Các đối tượng là các Tweet được biểu diễn dưới dạng các vector dựa trên tập từ điển.
* Tính liên kết giữa các đối tượng: có nhiều cách để tính liên kết giữa các đối tượng như **Single Linkage, Average Group, Complete Linkage**. Trong bài toán này nhóm em sử dụng **Average Group**

Khoảng cách giữa 2 cluster được tính là khoảng cách trung bình giữa các đối tượng trong 2 cluster đó (average distance):

A

B

C

E

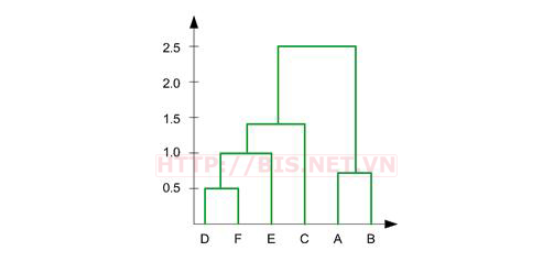
G

D1

D2

Hình 2: Khoảng cách giữa các Cluster

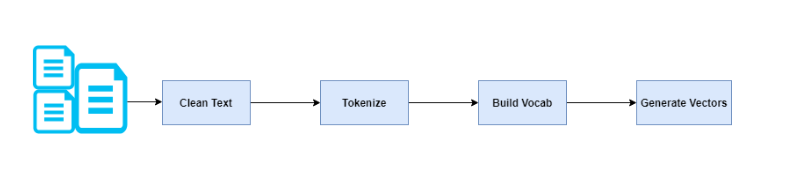
Quá trình phân cụm theo thuật toán AHC trong có thể được biểu diễn bằng biểu đồ dendrogram như sau:



Hình 3: Quá trình thực hiện phân cụm

**2.2. Bag of Words**

Trong bài toán, việc làm sao biểu diễn các Tweets (văn bản) để tính khoảng cách là một vấn đề quan trọng. Bag of Words là một thuật toán hữu dụng trong việc hỗ trợ xử lý ngôn ngữ tự nhiên. Ý tưởng của BoW là phân tích và phân nhóm dựa theo "Bag of Words"(corpus). Việc xây dựng tập từ điển sử dụng Bag of Words không quan tâm đến thứ tự từ trong câu cũng như ngữ nghĩa của từ. Phương pháp bao gồm các bước cơ bản như:



Bước 1: Clean Text

Các tập văn bản (Tweets) được chuẩn hóa, loại bỏ các ký tự đặc biệt, dấu câu, urls

Bước 2: Tokenize

Tách từ trong văn bản dựa vào khoảng trắng

Bước 3: Build Vocabulary

Xây dựng một tập từ vựng từ các từ được tách ra trong các văn bản

Bước 4: Generate Vectors

Từ tập từ điển đã có, tạo các vector tương ứng với từng văn bản (Tweets). Số chiều của vector tương ứng với độ lơn của tập từ điển

Ví dụ , với các văn bản là hai câu sau :

Câu 1: “The cat sat on the hat”

Câu 2: “The dog ate the cat and the hat”

Từ hai câu trên , vocabulary sẽ là :

{ the, cat, sat, on, hat, dog, ate, and }

Để có thể xây dựng được bag of words của 2 câu này , chúng ta đếm số lần xuất hiện của mỗi từ trong mỗi câu . Trong câu 1, “the” xuất hiện hai lần , và “cat”, “sat”, “on”, và “hat” mỗi từ xuất hiện một lần , như thế feature vector cho câu 1 sẽ là :

{ the, cat, sat, on, hat, dog, ate, and }

Câu 1: { 2, 1, 1, 1, 1, 0, 0, 0 }

Câu 2: { 3, 1, 0, 0, 1, 1, 1, 1}

Trong bài tập các trường của các vecto sẽ được đặt giá trị bằng cách tính Tf-idf cho từng từ tương ứng trong Tweet

**2.3. Tf-idf**

Tf-idf là viết tắt của term frequency – inverse document frequency, tf-idf là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản. Trong bài toán tf-idf được dùng để tính điểm của từng từ trong tập từ điển tai mỗi time\_window.

* Tf là tần số xuất hiện của một từ trong một văn bản:

Trong đó: : Số lấn xuất hiện của từ t trong văn bản d

: số lần xuất hiện nhiều nhất của một từ trong tweet\_window\_corpus.

* Idf được tính theo công thức:

Trong đó |D| : tổng số văn bản trong tập tweet\_window\_corpus

|| : số văn bản chứa từ t

Khi thực hiện trong bài toán, công thức tính tf-idf được nhóm thay đổi dựa trên công thức trên để phù hợp với việc tính toán. Phần thay đổi được trình bày cụ thể trong phần thực hiện.

1. **Thực hiện**

## 1. Lấy dữ liệu

Dữ liệu được lấy thông qua Twitter API. Để truy cập Twitter Streaming API, cần có các thông tin từ Twitter: API key, API secret, Access Token và Access Token secret. Sau đó, sử dụng thư viện trong Python là *Tweepy* để kết nối đến Twitter Streaming API và lấy về dữ liệu. Khi lấy dữ liệu về, Twitter Streaming API có hỗ trợ lọc những tweet theo khu vực địa lý bằng hàm *filter()*. Ở đây nhóm em sẽ lọc những tweet trong lấy được trong khu vực New York bằng cách thêm tham số cho hàm filter(locations=[-74,40,-73,41]), với [-74,40] là kinh độ và vĩ độ của New York. Sau khi chạy sẽ lấy được dữ liệu định dang JSON. Ví dụ một Tweet lấy về có những thông tin như sau:



## 2. Tiền xử lý dữ liệu

Do các Tweet lấy về sẽ gồm rất nhiều trường nên sẽ chỉ lấy những trường quan trọng với mỗi Tweet tương ứng. Mỗi Tweet sẽ lấy về [tweet\_unixtime, tweet\_gmttime, tweet\_id, text, hashtags, users, urls, media\_urls, nfollowers, nfriends]. Ví dụ như Tweet ở trên sẽ lấy được:

[1557640922, 'Sun May 12 13:02:02 +0000 2019', 1127559557335650305, '@homiesafe It the truth!!!', [], ['homiesafe'], [], [], 740, 1077]

Đối với mỗi tweet, nhóm tiền xử lý văn bản như sau. Đầu tiên là chuẩn hóa văn bản lấy được từ tweet, loại bỏ urls, user mentions và hashtag, cũng như các số và dấu câu. Tiếp theo, tách từ dựa vào khoảng trắng và loại bỏ từ dừng (stopwords).

## 3. Trích chọn đặc trưng

Do tập dữ liệu ở đây là dạng text nên nhóm em sử dụng phương pháp Bag-of-words để chuẩn hóa về dạng vecto. Mỗi tweet sẽ được thể hiện dưới dạng túi của các từ trong nó: tweet\_bag. Trong mỗi time\_window, đối với mỗi tweet, sẽ lưu vào tweet\_bag gồm user mentions, hashtag và những từ trong tweet sau khi đã được làm sạch và tách từ. Ví dụ tweet\_bag:

*@user,#hashtag1,#hashtag2,content,of,tweet*

Sau đó, sẽ kiểm tra trong mỗi tweet\_bag và loại bỏ những tweet mà có trên 3 user mentions hoặc trên 3 hashtags hoặc ít hơn 4 từ trong văn bản. Lý do là vì những tweet có nhiều users mentions hoặc hashtags, không giống như nói về một tin tức và có thể là các tweet spam. Những tweet\_bag đã được lọc ở trên sẽ được lưu vào một cửa sổ ngữ liệu tweet (tweet\_window\_corpus). Lưu ý là những tweet\_bag của time\_window đang xét sẽ được lưu lại trong tweet\_window\_corpus của cửa sổ trước (tối đa 4 time\_window trước - mỗi time\_window được đặt là 15 phút). tweet\_window\_corpus sẽ được khởi tạo lại sau mỗi 4 time\_window (tương đương với 1 giờ). Việc này liên quan đến việc tính điểm cho các từ trong từ điển sau này, vì những topic có thể trải dài trong một khoảng thời gian, những tweet ở time\_window trước có thể nói về cùng chủ đề của tweet tại time\_window hiện tại.

Bước tiếp theo là lọc từ vựng. Đối với mỗi time\_window, từ tweet\_window\_corpus, tạo một ma trận các vecto đặc trưng(sử dụng tfidf) đối với từ điển (ma trận là ma trận thưa), trong đó từ điển đã được loại bỏ user mention (nhưng vẫn giữ hashtags), và lưu vào tập từ điển chỉ gồm unigram và bi-gram.

## 4. Phân cụm theo thứ bậc các Tweet (AHC)

**4.1. Tính toán khoảng cách theo cặp tweet**

Tính toán khoảng cách giữa các cặp tweet dựa vào ma trận vector. Khoảng cách sẽ được tính bằng độ tương đồng euclidean. Nhóm sử dung thư viên sklearn và scipy trong python để tính toán khoảng cách cũng như ma trân vector của tập từ điển.

**4.2. Tính toán phân cụm thứ bậc (Hierarchy cluster)**

Để tính toán phân cum thứ cấp, nhóm em sử dụng thư viện fastcluster hỗ trợ cho phân cụm thứ cấp(hierarchical clustering). Ý tưởng ở đây là phân cụm những tweet trong cùng một chủ đề (topic) sẽ được phân cùng một cụm, và sau đó sẽ xem xét mỗi cụm cho một chủ đề phát hiện được.

**4.3. Cắt dendrogram**

Sau khi tính toán phân cụm, nhóm sẽ cắt kết quả trên cây thứ cấp tại ngưỡng khoảng cách là 0.5\*(khoảng cách lớn nhất giữa các cặp vecto). Lý do chọn ngưỡng này là do có thể ngưỡng cao thì kết qủa sẽ được các cụm sẽ lỏng lẻo hơn, có thể các chủ đề khác nhau sẽ trong cùng một cum, và ngược lại với ngưỡng thấp hơn. Nhóm thấy ngưỡng 0.5 hoạt động khá tốt.

**4.4.** **Sắp xếp kết quả các cụm**

Bước đầu sẽ là gán điểm và xếp hạng các cụm bởi kích thước của nó, chỉ lấy những cụm mà có số tweet trong cụm lớn hơn một mức nào đó (ở đây nhóm chọn là 3). Sau đó sẽ tính điểm của từng cụm bằng cách lấy tổng điểm của các từ trong cụm đó và chia trung bình.

Về cách tính điểm cho các từ trong từ điển ứng với mỗi time\_window. Điểm sẽ đươc tính dựa theo TF-IDF (Term Frequency - Inverse Document Frequency). Trước tiên về TF sẽ tính như sau:

(1)

Trong đó, : số lần xuất hiện của từ trong tập tweet\_window\_corpus.

: số lần xuất hiện nhiều nhất của một từ trong tweet\_window\_corpus.

: tổng số lần xuất hiện của từ trong tweet\_window\_corpus trong 3 time\_window trước đó.

Việc tính TF của từ trong time\_window hiện tại sẽ phụ thuộc vào các time\_window trước đó. Điều này thể hiện ở mẫu số của công thức trên. Mẫu số được tính là trung bình tổng của tần xuất xuất hiện của từ đó trong *t* time\_window trước, đồng thời lấy logarit để giảm trọng số của tần xuất xuất hiện của từ tại time\_window trước so với hiện tại. Việc lấy mẫu số này là giảm TF của từ đó trong time\_window hiện tại, vì những từ đã xuất hiện time\_window trước rồi thì từ đó có thể không còn quá nổi bật trong time\_window hiện tại.

Tiếp theo là tính IDF:

(2)

Trong đó, |D| : tổng số văn bản trong tập tweet\_window\_corpus

|| : số văn bản chứa từ t

Kết hợp công thức (1) và (2) tính được điểm của từng từ trong tweet\_window\_corpus sẽ là tích của TF và IDF:

(3)

Bên cạnh đó, những tweet có cùng hashtag thì có khả năng cao nói về một chủ đề hơn, do đó, những từ hashtag sẽ được tăng thêm điểm: *wtfidf\*boost*. Nhóm em chọn boost = 1.25 để việc tăng này không quá vượt trội và ảnh hưởng quá nhiều đến điểm của cả tập tweet\_window\_corpus.

Sau khi tính điểm cho từng từ, điểm của mỗi cụm được gán bằng tổng điểm của các từ trong cụm đó và chuẩn hóa bởi kích thước của cụm(chia cho số tweet trong cụm):

Trong đó, : tổng điểm của từ trong cụm i

: số tweet trong cụm i

Bước chuẩn hóa này giúp cho những chủ đề bớt bị phân mảnh đi, cũng như những cụm nhỏ với những từ nổi bật có thể có điểm cao hơn.

**4.5. Chọn tiêu đề cho chủ đề**

Nhóm em chọn tweet có điểm cao nhất(lấy tổng điểm của các từ trong tweet) trong mỗi cụm cho tiêu đề chủ đề phát hiện được. Việc lựa chọn như này có thể bao phủ với một số chủ đề, nhưng một số chủ đề vẫn có thể bị phân mảnh, một vài tiêu đề thu được có thể nói về cùng một chủ đề. Do đó, bước tiếp theo sẽ xử lý đối với việc phân mảnh chủ đề này.

**4.6.** **Phân cụm lại đối với tiêu đề**

Bước phân cụm này chỉ gồm các tiêu đề của những cụm đã được phân cụm. Tương tự như phân cụm cấp bậc ở trên, tiêu đề sẽ được tách từ, biến đổi thành ma trận vecto, sử dụng unigrams cho tập từ điển và không có giới hạn gì đối với từ. Tiến hành phân cụm. Sau đó sẽ xếp hạng kết quả các cụm tiêu đề với điểm cao nhất của mỗi cụm, nếu các tiêu đề không phân cụm thì xếp hạng giữ nguyên như bước trước.

**4.7.** **Lấy ra chủ đề cuối cùng**

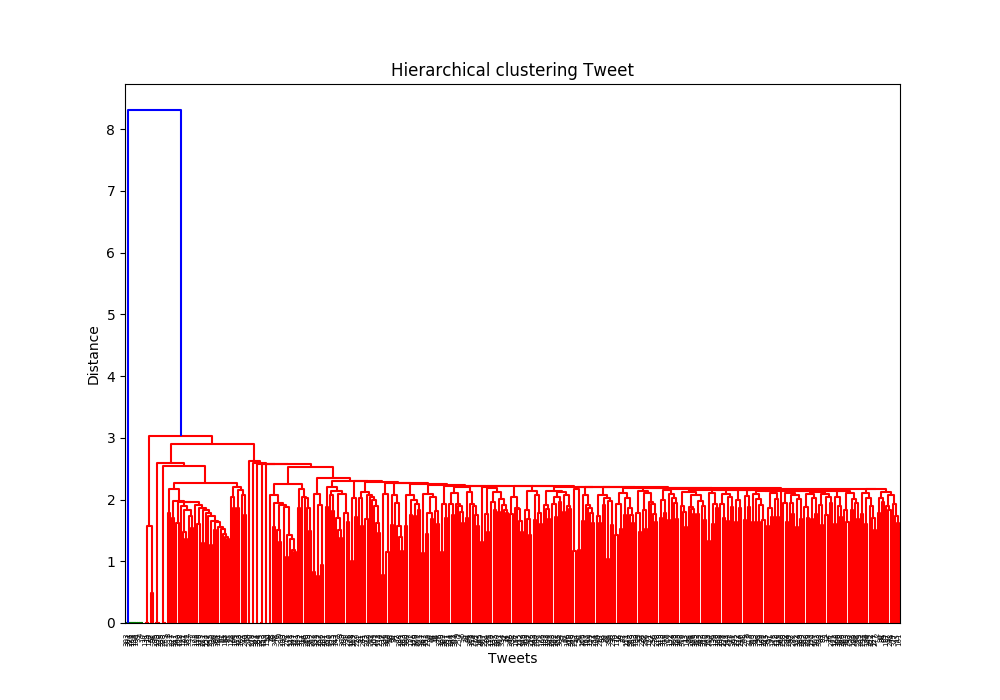
Sau khi bước phân cụm và xếp hạng cuối cùng, chọn tiêu đề với tweet đầu được đăng lên. Tiêu đề được lấy ra là gồm những keywords trong những cụm đó.

1. **Thực nghiệm và kết quả**

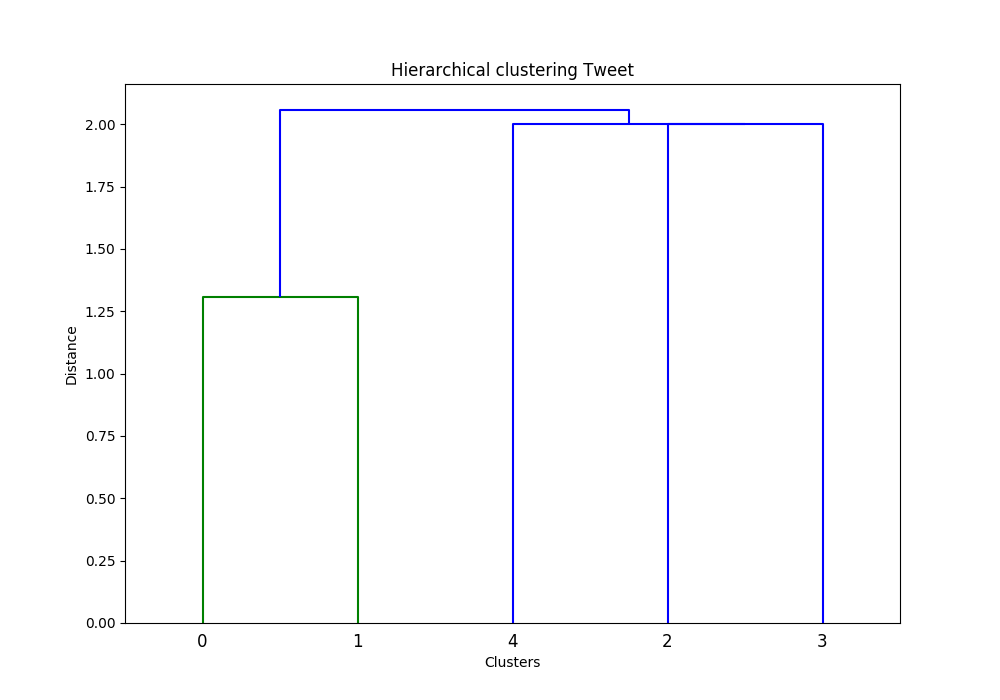
**1. Thực nghiệm**

Việc thực nghiệm được thử với dữ liệu là các Tweet được lấy vào ngày 12/5 tại khu vực New York, USA. Việc tính toán được thực hiện trên các time\_window, mỗi time\_window kéo dài 15 phút, sau 1 giờ tương đương 4 time\_window, các tweet\_window\_corpus sẽ được khởi tạo lại.

Biều đồ việc phân cụm các Tweet trong time\_window đầu tiên:

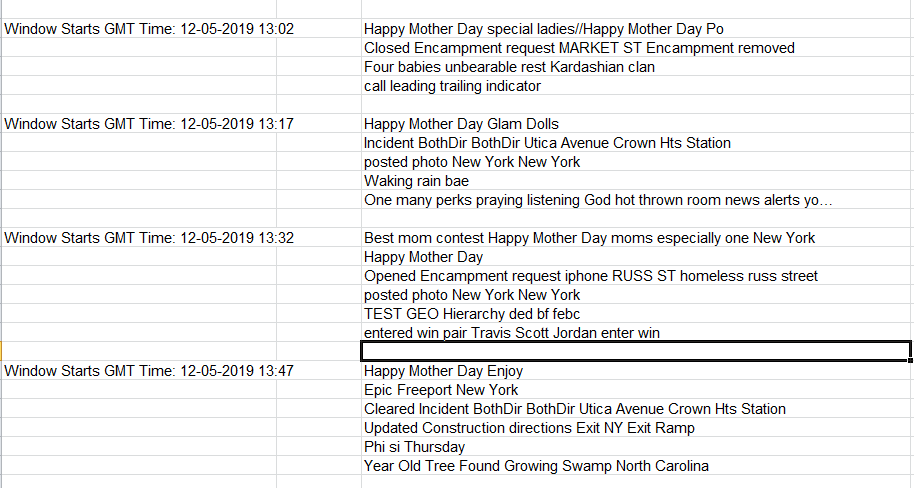


Biểu đồ phân cụm đưa ra các Topic chủ đề :



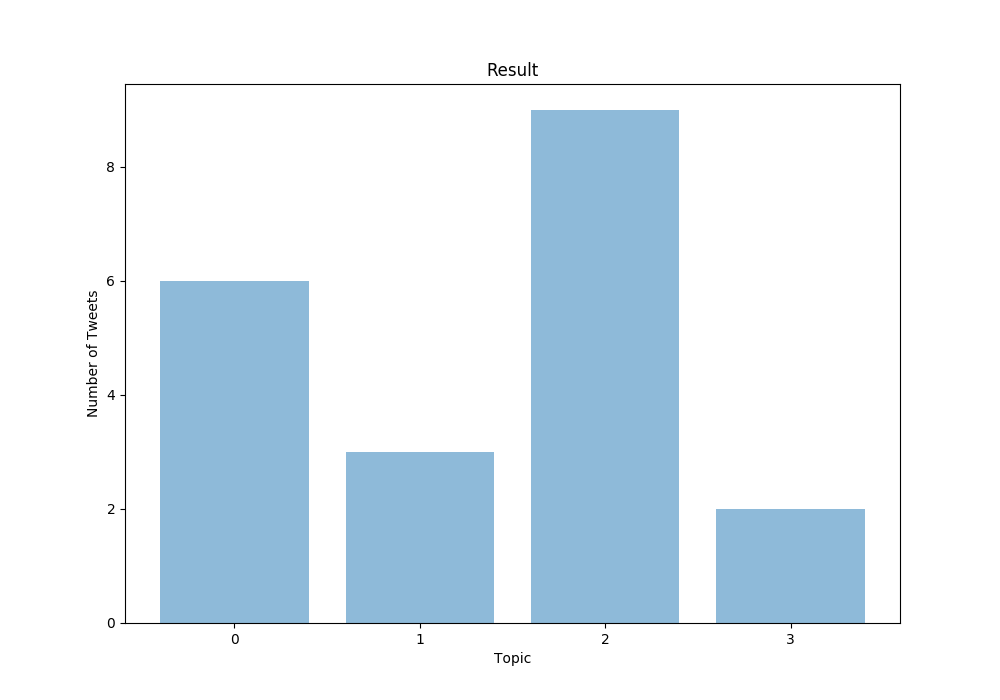
**2. Kết quả**

Chương trình chạy trên giao diện dòng lệnh. Sau khi xử lí nội dung các các hot topic được đưa ra trên màn hình:



Hình: Kết quả chạy chương trình

Biểu đồ số lượng các Tweet trong các hot topic



Hình: Biểu đồ số lượng các Tweet trong các Topic

1. **Tài liệu tham khảo**

* Giáo trình Khai phá Web trường Đại học Bách Khoa Hà Nội
* <http://ceur-ws.org/Vol-1150/ifrim.pdf>
* <https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tu-nhien/index.html>
* <http://www.cit.ctu.edu.vn/~dtnghi/dataminingR/clustering.pdf>