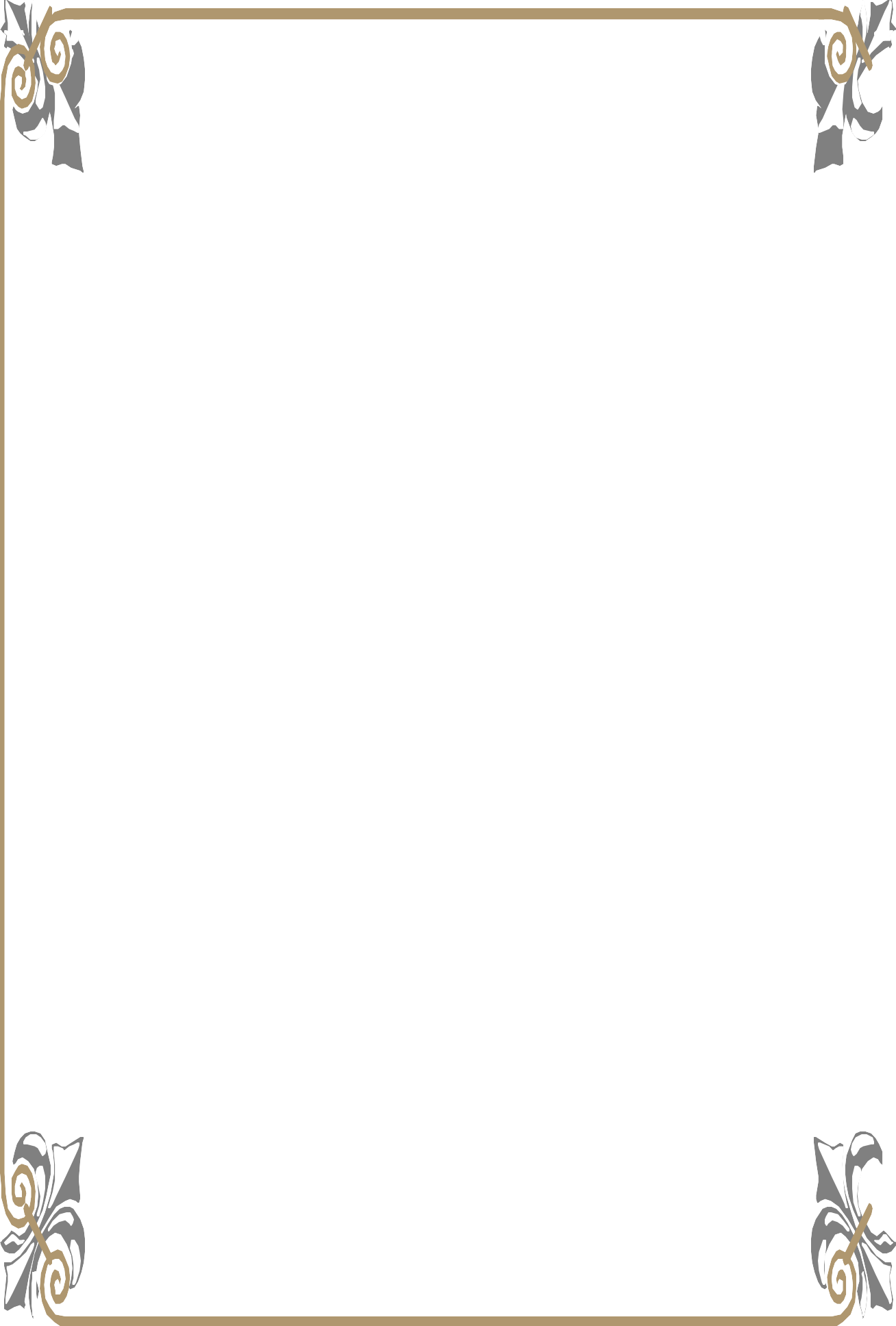
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI



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

====o0o====

**BÀI TẬP LỚN**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

***ĐỀ TÀI:***

TÓM TẮT VĂN BẢN

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **PGS.TS Lê Thanh Hương** |
| **Các thành viên:** | **Nguyễn Văn Lâm – 20151362**  **Đặng Văn Hà – 20151137**  **Phạm Huy Hoàng – 20151556** |
| **Hà Nội, 9/5/2019** |  |
|  |  |
|  |  |
|  |  |

Contents

[PHẦN 1: BÀI TOÁN 3](#_Toc9201906)

[I. Mô tả bài toán 3](#_Toc9201907)

[II. Hướng giải quyết 3](#_Toc9201908)

[PHẦN 2: LÝ THUYẾT VỀ CÁC KĨ THUẬT ĐÃ SỬ DỤNG 5](#_Toc9201909)

[I. Work2vec 5](#_Toc9201910)

[II. K-means clustering 9](#_Toc9201911)

[III. Hierarchical clustering (HCA) 10](#_Toc9201912)

[PHẦN 3: ỨNG DỤNG DEMO 13](#_Toc9201913)

[I. Các công nghệ sử dụng 13](#_Toc9201914)

[II. Hình ảnh demo 13](#_Toc9201915)

[*Tài liệu tham khảo* 16](#_Toc9201916)

# PHẦN 1: BÀI TOÁN

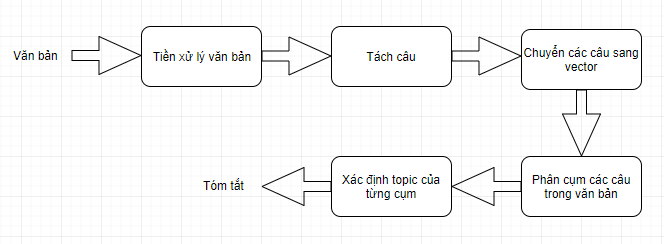
## I. Mô tả bài toán

Text summarization là một trong những bài toán ứng dụng của xử lý ngôn ngữ được kì vọng là sẽ đem lại nhiều lợi ích trong việc trính xuất thông tin từ lượng dữ liệu vô cùng lớn trên internet.

Đầu vào của bài toán: một văn bản có độ dài lớn, gồm nhiều câu

Đầu ra của bài toán : một văn bản có độ dài ngắn hơn nhiều văn bản ban đầu, nhưng vẫn nắm rõ những ý chính của văn bản cũ

## II. Hướng giải quyết



*Sơ đồ các bước chính trong xử lí bài toán*

Các bước xử lí bài toán gồm:

1. **Tiền xử lí văn bản:** Thực hiện các bước loại bỏ các chữ viết hoa và các khoảng trắng
2. **Tách câu:** Tách văn bản thành các câu tùy thuộc vào ngôn ngữ
3. **Chuyển các câu thành vector:** Tiến hành tách nhỏ các câu thành các từ sau đó chuyển sang vector (sử dụng work2vec còn việc tách từ ta dùng Vitokenizer)
4. **Phân cụm các câu trong văn bản:** Tiến hành ghép các câu có độ tương đồng với nhau vào một nhóm (số nhóm = số câu của văn bản tóm tắt) - ở đây chúng ta dùng Hierarchical clustering (hoặc Kmean) để phân cụm
5. **Xác định topic của từng cụm:** Với mỗi cụm xác định các câu mà có độ tổng quát nhất chọn làm câu nói ý chính cảu cả đoạn - với Hierarchical clustering ta xác định điểm trung bình giữa các cụm và lấy đó làm chủ đề, còn với Kmean ta xác định điểm gần tâm cụm nhất để làm chủ đề

# PHẦN 2: LÝ THUYẾT VỀ CÁC KĨ THUẬT ĐÃ SỬ DỤNG

## I. Work2vec

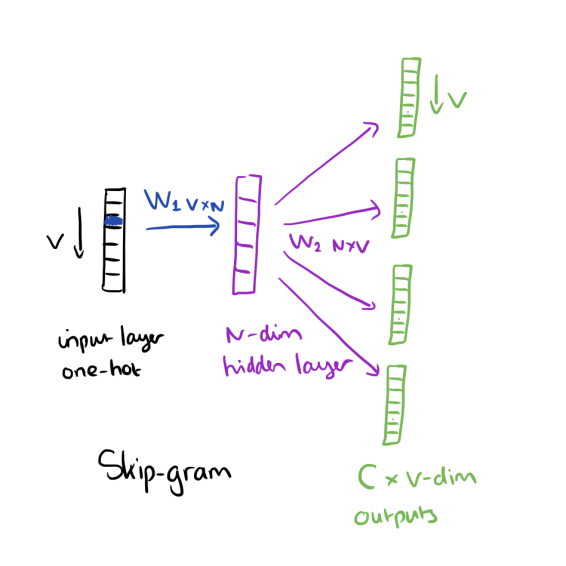
Việc mã hóa một từ và câu sang dạng số hoặc vector để dễ dàng xác định sự tương đồng và khác nhau trong các từ và câu, để áp dụng các giải thuật xử lí tầng sau. Và để giảm số chiều miêu tả một câu người ta đã nghĩ ra một mô hình Word embedding. Mục tiêu là xác định các từ có ý nghĩa tương đồng nhau trong ngôn ngữ và gắn cho chúng có khoảng cách gần nhau hơn trong Word embedding space. Vd: he – himself = she – herself. Áp dụng mô hình này khiến cho số chiều giảm xuống còn NxD thay vì NxN như ban đầu. Trong đó word2vec là một trong nhưng mô hình khá nổi tiếng của word embebding.

Word embebding là một model học không giám sát được training từ large corpus. Số chiều của word2vec là NxD (N là số văn bản, D là số chiều word embebding). Word2vec thì có 2 mô hình được áp dụng là Skip-gram, Continuous Bag of Words (CBOW)

1. Skip - gram

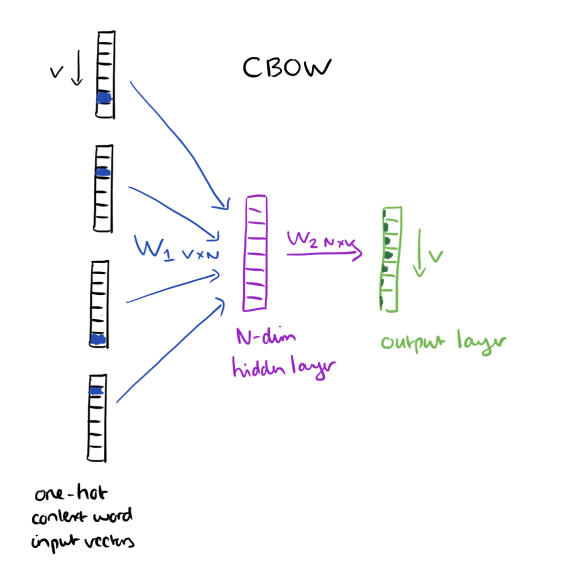
Mô hình dự đoán từ xung quang từ cho trước.

VD: với câu “I have a dog”. Với input là “have” thì output sẽ là “I” và “a”

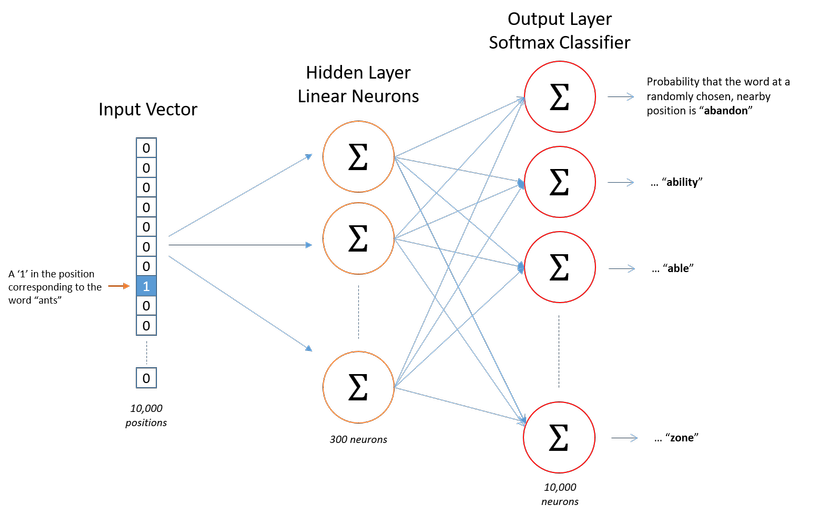


1. Continuous Bag of Words (CBOW)

CBOW thì ngược lại với skip-gram. Input là các câu, ouput là từ gần với các câu này.



Mô hình train của skip-gram như sau (của CBOW thì ngược lại)



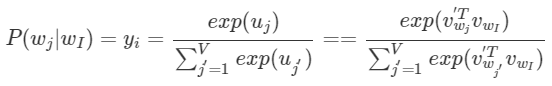
* Input là one-hot-vector mỗi word sẽ có dạng x1,x2,..xv trong đó V là số vocabulary, là một vector trong đó mỗi word sẽ có giá trị 1 tương đương với index trong vocabulary và còn lại sẽ là 0.
* Weight matrix giữa input và hidden layer là matrix W(có dimention VxN) có active function là linear, weight giữa hidden và out put là W′ (có dimention là NxV) active function của out put là soft max.
* Mỗi row của W là vector N chiều đại diện cho vw là mỗi word trong input layer.Mỗi row của W là w. Lưu ý là input là 1 one hot vector (sẽ có dạng 000100) chỉ có 1 phần tử bằng 1 nên.



* Từ hidden layer đến output là matrix W′=w′i,j . Ta tính score ui cho mỗi word trong vocabulary.



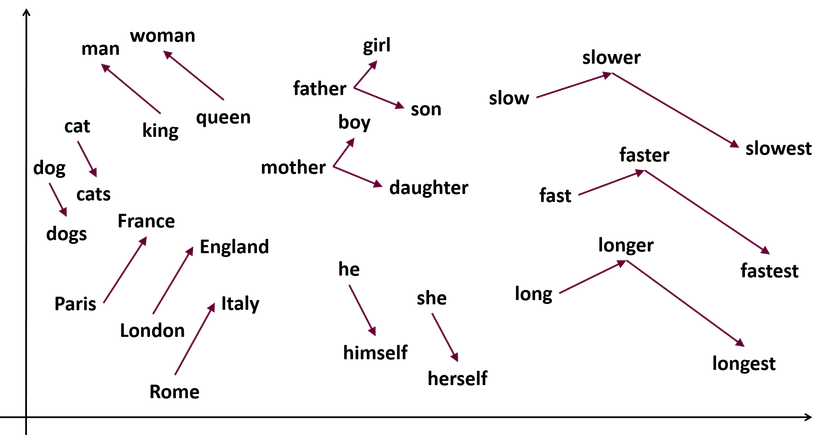
* Trong đó vwj là vector colum j trong W′. Tiếp đó ta sử dụng soft max funtion.



*Trong đó vw và vw′ là 2 vector đại diện cho word w đên từ matrix W và W′.*

Ta sẽ thu được tập các vector mô tả các từ vựng, mà vẫn lưu ngữ nghĩa

VD: như hình dưới he -> himself và she -> herself là 2 vector bằng nhau (cùng độ dài và hướng) -> chúng có độ tương đồng về ngữ nghĩa,



## II. K-means clustering

K-means clustering là thuật toán phân cụm không giám sát và được sử dụng phổ biến nhất hiện nay.

Ý tưởng của thuật toán là tạo ngẫu nhiên k tâm cụm, rồi tìm các điểm gần tâm cụm nào thì sẽ được đánh số thuộc vào cụm đó, rồi chúng ta tính trung bình các điểm trong cụm để suy ra vị trí mới của tâm cụm. Thuật toán cứ tiếp diễn cho đến khi tâm cụm ngừng dịch chuyển hoặc độ dịch chuyển không quá lớn (hoặc tổng khoảng cách các điểm trong cụm đến tâm cụm là nhỏ nhất).

Phân tích toán học:

* Cho một tập dữ liệu (x1, x2, … , x3), với mỗi quan sát là các vector d chiều, mục đích của k-means đẩy n vector trên vào k (k<=n) S = {S1,S2, …, Sk}
* Ta có phải đi tối ưu tổng khoảng cách từ các điểm trong cụm đến các tâm cụm là nhỏ nhất





## III. Hierarchical clustering (HCA)

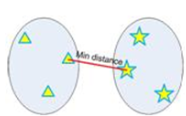
Hierarchical clustering là thuật toán phân cụm không giám sát, gồm có 2 loại chính là:

* Agglomerative: Nghĩa là phân cụm từ dưới lên, những thành phần tương đồng sẽ kết nối thành 1 cụm to hơn
* Divisive: Nghĩa là phân cụm từ trên xuống, tất cả xuất phát từ cùng 1 cụm rồi chia nhỏ ra

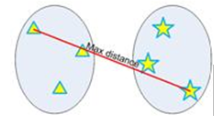
Ý tưởng của thuật toán là trong mỗi vòng lặp, chúng ta tìm các cặp có độ tương đồng gần nhau nhất (khoảng cách gần nhau nhất) gom lại thành một cụm. Cứ tiếp diễn cho đến khi tất cả các điểm được gom lại thành một cụm. Muốn chi số cụm bao nhiêu ta chỉ cần dừng lại ở bước mà ta đã có đủ số cụm

Việc xác định các cụm hay các điểm có gần nhau hay không được được giải quyết theo vài cách:

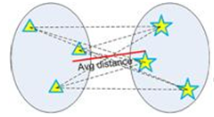
* **Single Linkage:** Khoảng cách giữa 2 cụm là khoảng cách giữa 2 điểm gần nhau nhất trong cụm đó



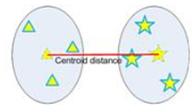
* **Complete Linkage:** Khoảng các giữa 2 cụm là khoảng cách giữa 2 điểm xa nhất

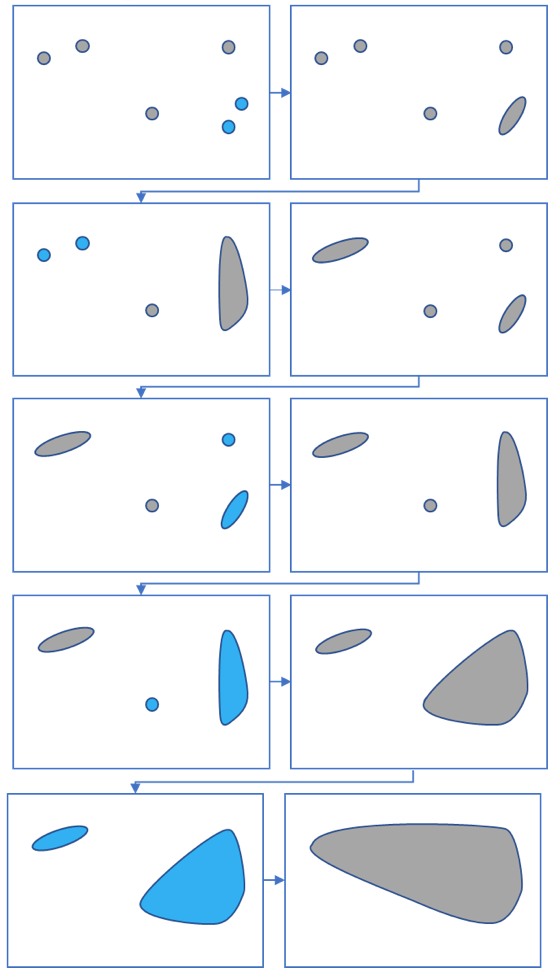


* **Average group:** Khoảng cách giữa 2 cụm là khoảng cách trung bình giữa các điểm trong cụm



* **Centroid distance:** Khoảng cách giữa 2 cụm là khoảng cách giữa 2 tâm cụm





*Sơ đồ hierarchical clustering phân cụm bottom-up theo các bước*

# PHẦN 3: ỨNG DỤNG DEMO

## I. Các công nghệ sử dụng

* Nltk => Để tách câu
* ViTokenizer => Để tách từ trong tiếng việt
* Gensim => Để load model work2vec tiếng việt
* K-means, hierarchical clustering => Để phân cụm

## II. Hình ảnh demo

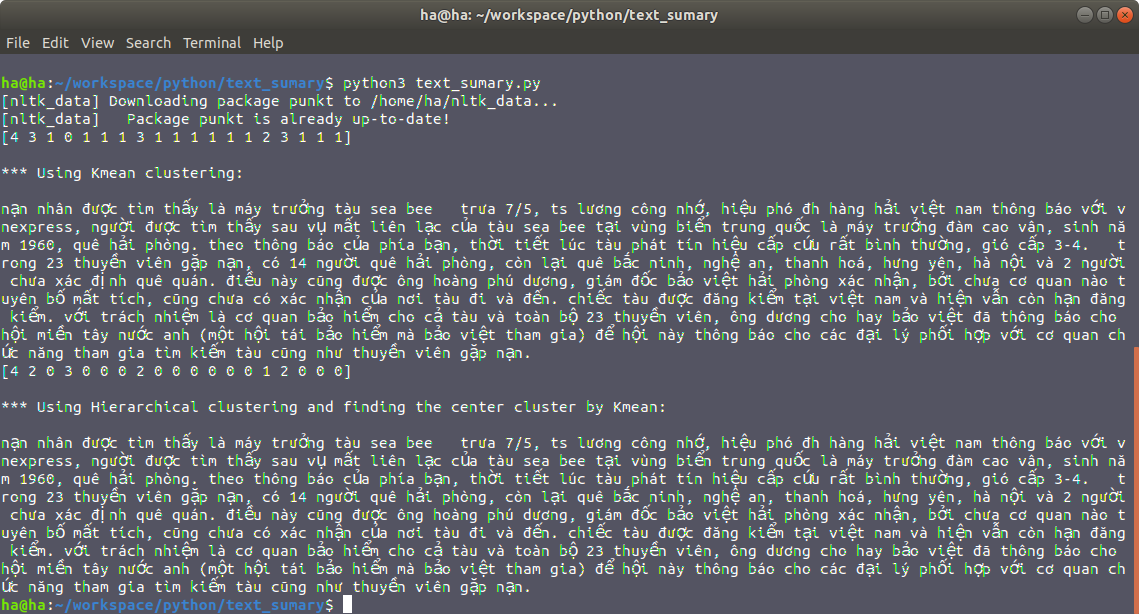
Văn bản đầu vào:

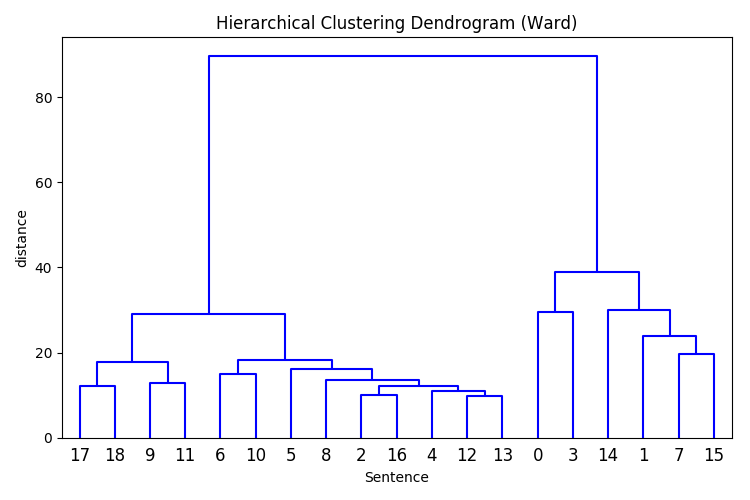
Nạn nhân được tìm thấy là máy trưởng tàu SEA BEE   Trưa 7/5, TS Lương Công Nhớ, Hiệu phó ĐH Hàng hải Việt Nam thông báo với VnExpress, người được tìm thấy sau vụ mất liên lạc của tàu SEA BEE tại vùng biển Trung Quốc là máy trưởng Đàm Cao Vân, sinh năm 1960, quê Hải Phòng.  Ngoài máy trưởng Vân, từ ngày 3/5 đến nay tàu cứu nạn DONGHAIJIU 169 của Trung tâm Cứu nạn Biển Đông Thượng Hải (Trung Quốc) vẫn chưa tìm thêm được nạn nhân nào, cũng chưa xác định tọa độ tàu gặp nạn. Ông Nhớ cho biết, chưa xác định được nguyên nhân tàu mất liên lạc. Theo thông báo của phía bạn, thời tiết lúc tàu phát tín hiệu cấp cứu rất bình thường, gió cấp 3-4.   Trong 23 thuyền viên gặp nạn, có 14 người quê Hải Phòng, còn lại quê Bắc Ninh, Nghệ An, Thanh Hoá, Hưng Yên, Hà Nội và 2 người chưa xác định quê quán. Cao tuổi nhất là thuyền viên Nguyễn Văn Bình, sinh năm 1955, quê Hải Phòng. Nhỏ tuổi nhất là Nguyễn Đình Dương, sinh năm 1985, quê Từ Liêm, Hà Nội. Ông Nhớ cho hay, đến nay chưa thể nói SEA BEE mất tích vì chưa xác định được tọa độ. Điều này cũng được ông Hoàng Phú Dương, Giám đốc Bảo Việt Hải Phòng xác nhận, bởi chưa cơ quan nào tuyên bố mất tích, cũng chưa có xác nhận của nơi tàu đi và đến. SEA BEE xuất cảng Qinghuang Dao (Trung Quốc) ngày 1/5, theo lịch trình sẽ đến Manila (Philippines) ngày 6/5. Nhưng rời cảng được 1 ngày thì tàu mất liên lạc với chủ tàu là Công ty Vận tải biển Đông Long.  Ông Dương cho hay, SEA BEE được Bảo Việt Hải Phòng bảo hiểm với mức 2 triệu USD. Con tàu trọng tải 6.273 DWT này được đóng tại Nhật Bản năm 1980 và năm 2003, Công ty Đông Long đã mua lại. Hiện Đông Long cho Công ty Cryxtal Shipping của Hàn Quốc thuê định kỳ. Chiếc tàu được đăng kiểm tại Việt Nam và hiện vẫn còn hạn đăng kiểm. Với trách nhiệm là cơ quan bảo hiểm cho cả tàu và toàn bộ 23 thuyền viên, ông Dương cho hay Bảo Việt đã thông báo cho Hội miền Tây nước Anh (một hội tái bảo hiểm mà Bảo Việt tham gia) để hội này thông báo cho các đại lý phối hợp với cơ quan chức năng tham gia tìm kiếm tàu cũng như thuyền viên gặp nạn. Về việc chi trả bảo hiểm, ông Dương khẳng định, khi xác định được tàu và các thuyền viên mất tích hoàn toàn thì Bảo Việt Hải Phòng sẽ trả một lần cho gia đình nạn nhân với mức bảo hiểm trách nhiệm cao nhất - 25.000 USD/người.  Trong vòng 2 năm qua, ĐH Hàng hải đã mất đi nhiều thuyền viên. Ngày 20/1, tàu Tiên Phong của Hàn Quốc bị đắm trên vùng biển phía đông CHDCND Triều Tiên, làm thiệt mạng 6 thủy thủ Việt Nam. Trước đó, tháng 2/2004, chiếc tàu Duri cũng bị đắm khiến 18 thủy thủ của Việt Nam bị chết và mất tích.

Văn bản đầu ra:

nạn nhân được tìm thấy là máy trưởng tàu sea bee   trưa 7/5, ts lương công nhớ, hiệu phó đh hàng hải việt nam thông báo với vnexpress, người được tìm thấy sau vụ mất liên lạc của tàu sea bee tại vùng biển trung quốc là máy trưởng đàm cao vân, sinh năm 1960, quê hải phòng. theo thông báo của phía bạn, thời tiết lúc tàu phát tín hiệu cấp cứu rất bình thường, gió cấp 3-4.   trong 23 thuyền viên gặp nạn, có 14 người quê hải phòng, còn lại quê bắc ninh, nghệ an, thanh hoá, hưng yên, hà nội và 2 người chưa xác định quê quán. điều này cũng được ông hoàng phú dương, giám đốc bảo việt hải phòng xác nhận, bởi chưa cơ quan nào tuyên bố mất tích, cũng chưa có xác nhận của nơi tàu đi và đến. chiếc tàu được đăng kiểm tại việt nam và hiện vẫn còn hạn đăng kiểm. với trách nhiệm là cơ quan bảo hiểm cho cả tàu và toàn bộ 23 thuyền viên, ông dương cho hay bảo việt đã thông báo cho hội miền tây nước anh (một hội tái bảo hiểm mà bảo việt tham gia) để hội này thông báo cho các đại lý phối hợp với cơ quan chức năng tham gia tìm kiếm tàu cũng như thuyền viên gặp nạn.

Hình ảnh chạy



HÌnh ảnh của hierarchical clustering sau khi phân cụm một văn bản

*Tài liệu tham khảo*

* <http://travisaltman.com/honeypot-honeyd-tutorial-part-1-getting-started/>
* <http://bis.net.vn/forums/p/571/1485.aspx>
* <https://gist.github.com/aparrish/2f562e3737544cf29aaf1af30362f469>
* <https://thorpham.github.io/blog/2018/04/24/word2vec/>