

REPORT

INTELLIGENT SYSTEM

NEW SUMMARY BY MACHINE LEARNING APPROACH

Lecturer: Assoc. Prof. Quản Thành Thơ

Group Members:

Nguyễn Trí Chấn Hưng 1552159

Lê Bá Thành 1552340

Contents

[I. Introduction 3](#_heading=h.30j0zll)

[1.1 Overview 3](#_heading=h.1fob9te)

[1.2 Objective and scope 4](#_heading=h.3znysh7)

[II. Background 5](#_heading=h.2et92p0)

[2.1 Word2vec 5](#_heading=h.tyjcwt)

[2.2 KNN (K Nearest Neighbor) 6](#_heading=h.3dy6vkm)

[2.3 K-means 7](#_heading=h.1t3h5sf)

[III. System Design 8](#_heading=h.4d34og8)

[3.1. Dataset 8](#_heading=h.2s8eyo1)

[3.2 NLP Processing 9](#_heading=h.17dp8vu)

[3.3 Visualization module 10](#_heading=h.3rdcrjn)

[3.3 Extracting summarization from the news 13](#_heading=h.26in1rg)

[IV. Evaluation 15](#_heading=h.lnxbz9)

[V. Conclusion and Future Work 17](#_heading=h.35nkun2)

[5.1 Conclusion 17](#_heading=h.1ksv4uv)

[5.2 Future work 17](#_heading=h.44sinio)

# I. Introduction

## 1.1 Overview

In the modern world, the amount of information that people reached increase day by day, especially with e-news. besides the main content of each news, news articles can be long and often take too much time to get to the point. Due to that, a summary offers reader a chance to understand the main idea of the piece, without needing to read the actual text itself. It helps present the reader with the main points in a shorter amount of time, by excluding details that aren't necessary.

Text Summarization is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. With growing digital media and ever growing publishing – who has the time to go through entire articles / documents / books to decide whether they are useful or not.

Text summarization can broadly be divided into two categories – Extractive Summarization and Abstraction Summarization.

* Extractive summarization: These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method.
* Abstraction Summarization: These methods use advanced NLP (Natural Language Processing) techniques to generate an entirely new summary. Some parts of this summary may not even appear in the original text.

An intelligent system can be used to solve both of method above. However, in this article, we will be focusing in extractive summarization technique

## 1.2 Objective and scope

This system consists of two-part: *NLP processing* and *Extracting summarization from the news*. We want to apply machine learning approaches, as these problems require analysis of billions of contents from the Vietnam news. With automation using artificial intelligence, the users can gain better main content in each article with good performance and efficiency to the users.

For *NLP processing*, provided a list of keywords in the tag from each news, we expect to train a model with a dictionary of the keywords above. Because we believe that the keywords at the end of each article contain the main information of that news.

For *Extracting summarization*, the module is expected to classify the URL new belongs to what type of categories and to extract the main information from the new article URL base on the dictionary which is created above.

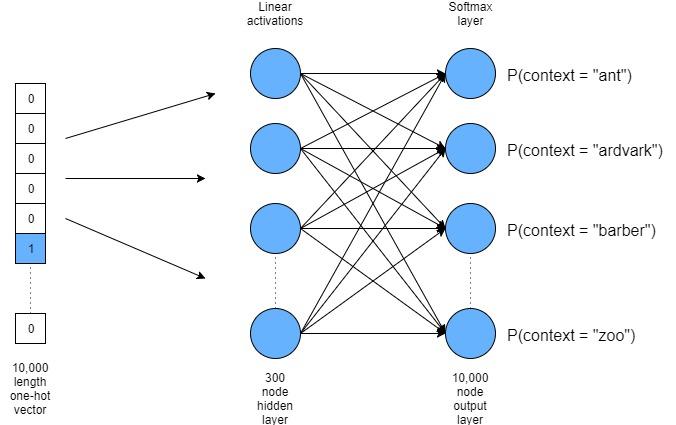
# II. Background

## 2.1 Word2vec

Word2vec is a technique for natural language processing. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosen carefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

Word2vec was created and published in 2013 by a team of researchers led by Tomas Mikolov at Google.

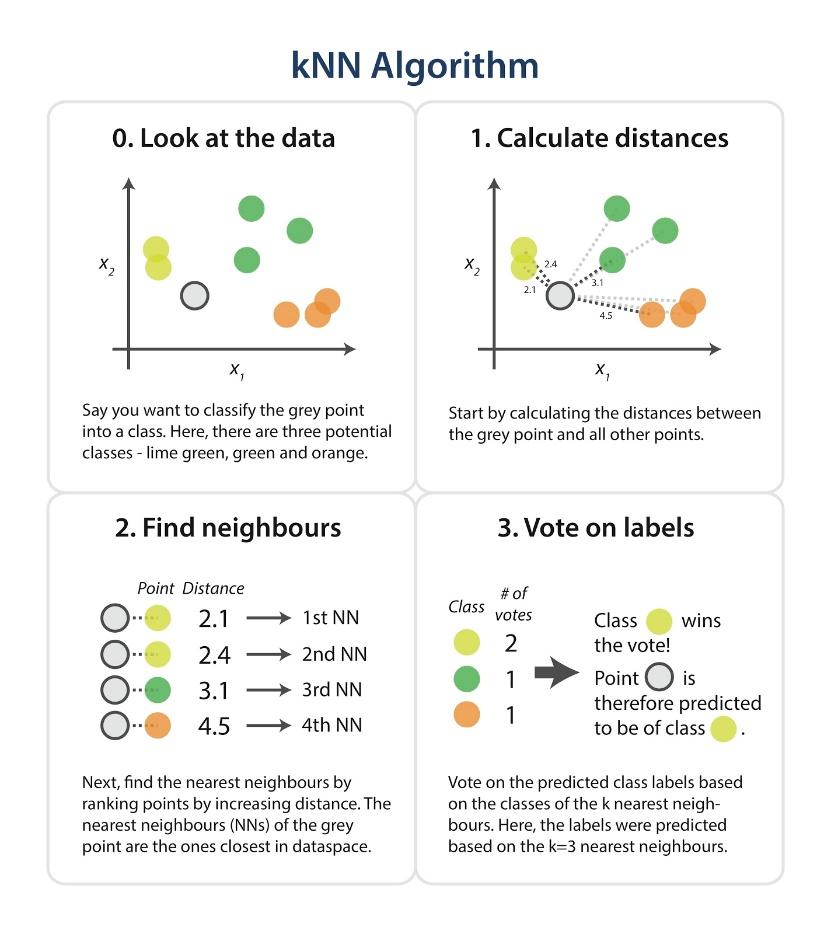


## 2.2 KNN (K Nearest Neighbor)

K Nearest Neighbor is one of the fundamental algorithms in machine learning. Machine learning models use a set of input values to predict output values. KNN is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbor is classified.

KNN classifies the new data points based on the similarity measure of the earlier stored data points. For example, if we have a dataset of tomatoes and bananas. KNN will store similar measures like shape and color. When a new object comes it will check its similarity with the color (red or yellow) and shape. The main point of KNN is selecting the K to represent the number of the nearest neighbors we used to classify new data points. There are two ways to select the K:

* The value of K equal to square root the total number data points available in the datasets.
* Odd value of is always selected to avoid confusion between 2 classes.



## 2.3 K-means

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances)

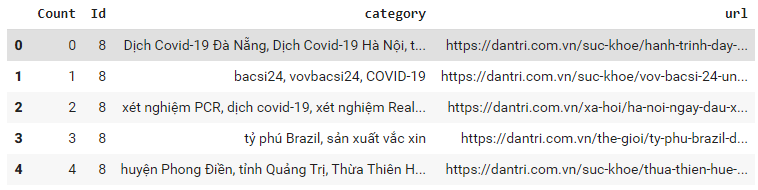
To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

# III. System Design

## 3.1. Dataset

In this section, we want to describe the process to implement the dictionary which is a list of keywords is present under vector from tags in the news articles from many trusted sources (bamoi, dantri, vnexpress,..)

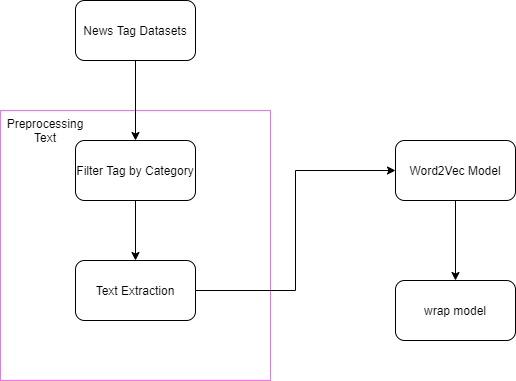




The dataset contains the list of keywords is divided by a list of categories (Politic, Sport, Education, Entertainment, Technology, Economy, Law, Film, Covid19) which is label by Id column. From this dataset, we can get the main keyword features of each news.

## 3.2 NLP Processing

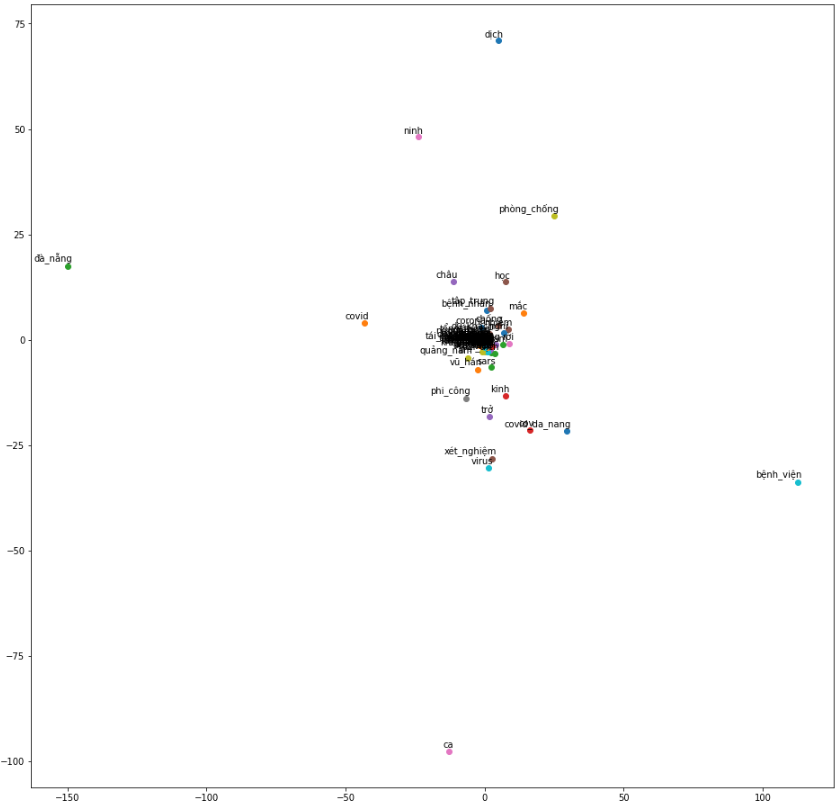
This module is built on python using word2vec model. The system is presented in Figure 4.

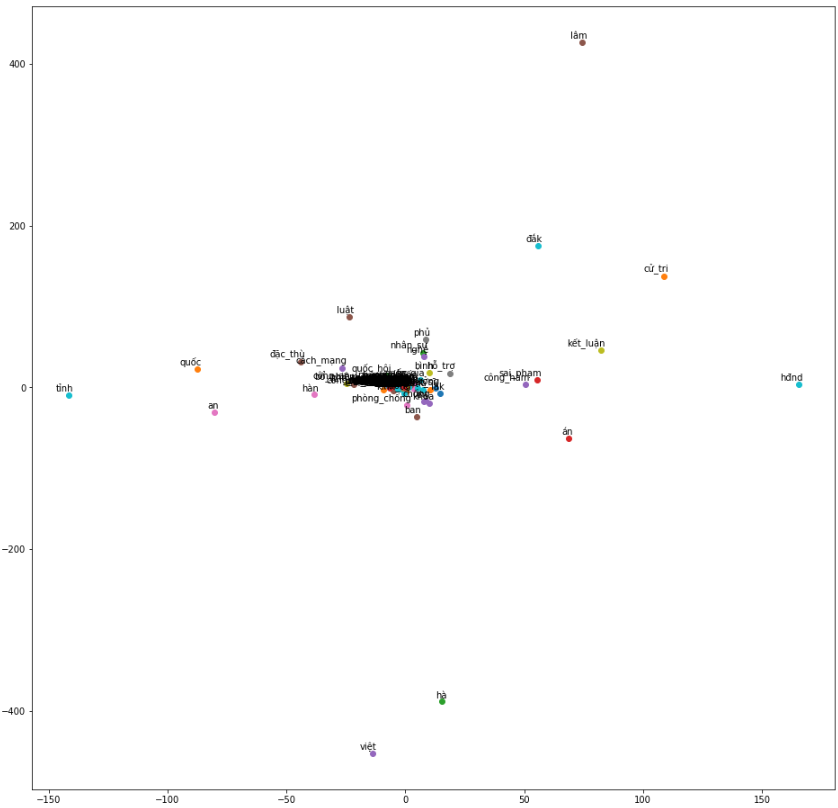


The list of keywords is filtered by each category. Then the stop word list will be provided to remove all noise. After that, we will split the sentences to words. However, Vietnam language cannot full meaning with single word. So we need one more step with Vitonkenizer library to link all single word which have full meaning.

In next step, we setup the hyper parameter for the word2vec model (learning rate, word size, alpha, min\_count,…). Then we wrap the model by each category.

## 3.3 Visualization module

In this section, we want to describe, the visualization of vector from category module above.

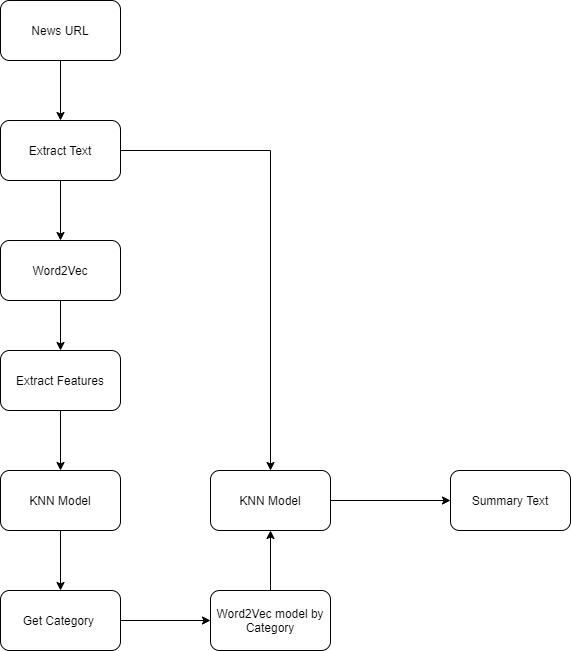


*Figure 6. Politic Word2vec model*

As we can see, every vector is close to each other which is show that almost vector is relevant together. So we can sure that we can extractive summarization from this model.

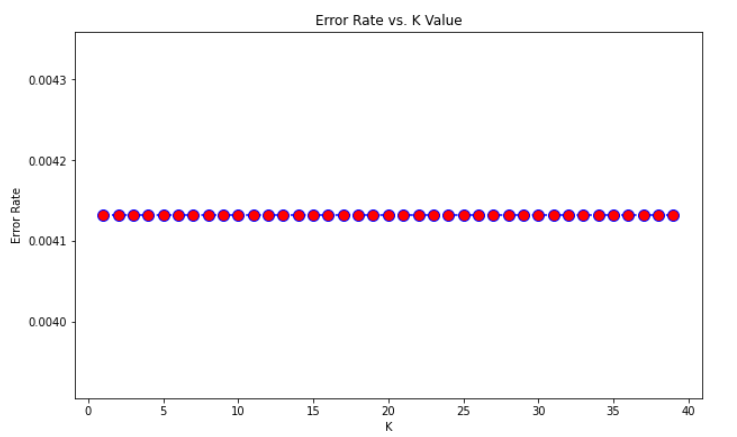
## 3.3 Extracting summarization from the news

This module is built as a backend system to extractive summarization. The module be presented in Figure 7.



This module will get the new URL from the user, then the system will extract the content of the news from the URL and then apply to word2vec model to get all the vector of news. In the next step we will train the categories model above with KNN algorithm to classification the new type of news belongs to what type of category.

To select the number K, we run the KNN algorithm mutil time and get the mean error as below.





In next step, we will classification every vector of the news into the KNN model and get the category label which have highest number of vector belonged to.

In next step we will run the vector of news through the dictionary of keywords we store to vector. If the vector of the news belongs to the model, we will store it. Then we append new of vectors to K-means algorithm. In this step, our concept to extractive summarization that the list of vectors will be split to 3 clusters. So if the new vectors nearly to cluster that words will also contain importance information of the news.

# IV. Evaluation

* In the conclusion, we want to show the result of our module

**Origin text**:

Chiều 1/12, theo báo cáo nhanh của Bộ Y tế về bệnh nhân 1347, cơ quan chức năng bước đầu điều tra có 513 người tiếp xúc với người này. Trong đó, 99 người thuộc diện F1 đã được cách ly, lấy mẫu xét nghiệm toàn bộ.

Kết quả cho thấy có 81 trường hợp âm tính, 18 trường hợp đang chờ kết quả.

Trường hợp F2 là 414 người, được cách ly tại nhà, đã lấy mẫu xét nghiệm 337 trường hợp, 123 trường hợp âm tính, số còn lại đang chờ kết quả xét nghiệm.

Tại cuộc họp khẩn trực tuyến với tất cả cơ sở y tế để bàn về các giải pháp phòng chống Covid-19 sau khi xuất hiện ca nhiễm mới do Sở Y tế TP.HCM chủ trì chiều nay,

Phó Giám đốc Sở Y tế TP.HCM Tăng Chí Thượng đề nghị tất cả bệnh viện trên địa bàn cập nhật ngay các địa điểm này trên phiếu khai báo y tế.

Ông cho biết thêm hiện Bệnh viện dã chiến Củ Chi vẫn hoạt động, Bệnh viện Điều trị Covid-19 Cần Giờ mở cửa, Bệnh viện Bệnh Nhiệt đới TP.HCM sẵn sàng tiếp nhận các ca nặng.

Liên quan ca lây nhiễm do không tuân thủ cách ly, Sở Y tế TP.HCM có văn bản đề xuất Ban Chỉ đạo Quốc gia phòng, chống dịch Covid-19 không cho phép tiếp viên, tổ bay kết thúc cách ly tập trung sớm như quy định hiện hành,

ngay cả khi chuyến bay không có người dương tính.

Liên quan đến đề xuất của Sở Y tế TP.HCM, Bộ Y tế kiến nghị Thủ tướng Chính phủ chỉ đạo ngừng việc cách ly thành viên tổ bay tại các cơ sở cách ly tập trung của các hãng hàng không,

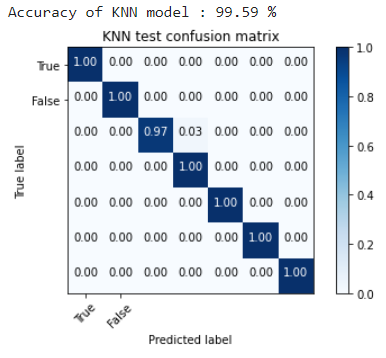
giao việc tổ chức cách ly tập trung tổ bay cho các cơ sở cách ly tập trung do UBND các tỉnh, thành phố quản lý.

Bộ Y tế tiếp tục chỉ đạo và phối hợp chặt chẽ với ngành y tế TP.HCM điều tra mở rộng, khẩn trương truy vết, cách ly và lấy mẫu xét nghiệm những trường hợp liên quan ca bệnh, theo dõi chặt chẽ diễn biến của ổ dịch.

"""

**Summary text**:

* Chiều 1/12, theo báo cáo nhanh của Bộ Y tế về bệnh nhân 1347, cơ quan chức năng bước đầu điều tra có 513 người tiếp xúc với người này. Kết quả cho thấy có 81 trường hợp âm tính, 18 trường hợp đang chờ kết quả. Trường hợp F2 là 414 người, được cách ly tại nhà, đã lấy mẫu xét nghiệm 337 trường hợp, 123 trường hợp âm tính, số còn lại đang chờ kết quả xét nghiệm.Tại cuộc họp khẩn trực tuyến với tất cả cơ sở y tế để bàn về các giải pháp phòng chống Covid-19 sau khi xuất hiện ca nhiễm mới do Sở Y tế TP.HCM chủ trì chiều nay, Phó Giám đốc Sở Y tế TP.HCM Tăng Chí Thượng đề nghị tất cả bệnh viện trên địa bàn cập nhật ngay các địa điểm này trên phiếu khai báo y tế.
* The result of the KNN module, is evaluate with confusion matrix graph



# V. Conclusion and Future Work

## 5.1 Conclusion

Finish the project, we have more opportunity to study and earn more knowledge in the field of NLP and other machine learning technique. We have successful build the new summary system which provides helps the user can summary the new from origin new.

## 5.2 Future work

In the future, our group want to change the KNN algorithm to another apply of machine learning such as SVM or Decision Tree with help us avoid overfit problem and get better accuracy and performance. Besides that, we want to improve our metric with new summary instead of using machine learning. We want to work on deep learning algorithm