**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

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

****

**CAPSTONE PROJECT REPORT**

**SUBJECT:**

**Websites Categorization**

| **Giáo viên hướng dẫn:** | Thân Quang Khoát  Trần Việt Trung  Nguyễn Kiêm Hiếu  Bùi Thị Mai Anh  Nguyễn Thị Oanh |
| --- | --- |
| **Sinh viên thực hiện:** | Nguyễn Thụ Hiếu  Nguyễn Minh Tuấn  Nguyễn Khánh Linh  Bùi Trần Hải Quân  Vũ Đình Minh |

**TABLE OF CONTENTS**

[**I.**](#_heading=h.30j0zll) **Problem Statement** 2

[**II.**](#_heading=h.1fob9te) **Data Science methodology** 3

[**1.**](#_heading=h.3znysh7) **Data Crawling** 3

[**2.**](#_heading=h.2et92p0) **Data Preprocessing** 5

[**3.**](#_heading=h.tyjcwt) **Data Visualization** 6

[**4.**](#_heading=h.3dy6vkm) **Model Training** 15

[**III.**](#_heading=h.1t3h5sf) **Result** 18

[**1.**](#_heading=h.4d34og8) **Final Result** 18

[**2.**](#_heading=h.2s8eyo1) **Difficulties:** 20

[**3.**](#_heading=h.17dp8vu) **Future Improvement** 20

[**IV.Conclusions** 21](#_heading=h.lnxbz9)

[IV.](#_heading=h.35nkun2) Reference 22

**Foreword**

Today, the development of information technology has penetrated into most aspects of social life, supporting people in many fields. Internet and Websites are indispensable parts of human life, helping us to communicate, connect, exchange, buy, sell, study and update news etc… With great advances in Data Science and Machine Learning, our team proposes an approach for Websites Categorization.

We would like to express our sincere thanks to Mr. Than Quang Khoat, Mr. Tran Viet Trung, Mrs. Bui Thi Mai Anh , Mr. Nguyen Kiem Hieu, Mrs. Nguyen Thi Oanh for trusting our team to choose this topic, and for their enthusiastic guidance during the last semester, helping us learn a lot of new knowledge and especially giving different perspectives on various fields in the information technology industry.

Due to both subjective and objective reasons, also due to limited knowledge and the shortage of time, the research we carried out is still imperfect. Our team is looking forward to receiving guidance to let our group improve this subject in the future.

| Our Team |
| --- |
| Nguyễn Thụ Hiếu  Nguyễn Minh Tuấn  Nguyễn Khánh Linh  Bùi Trần Hải Quân  Vũ Đình Minh |

# **Problem Statement**

Categorization can be defined as “the process by which individuals group objects or events into categories” (Norman, 1966). In other words, categorization is the act of grouping similar things together so that they can be stored in memory for future reference.

There are several use cases for website categorization across a variety of industries. One significant use of website categorization is in cybersecurity, where we group websites into categories such as spam, phishing, or other "problematic" websites that we do not want users of our networks, such as employees or clients, to access.

Marketing is also an important use of website classification. If we want to position our adverts on publishers' websites, we need to do so on web pages that fall under the same heading or industry as the goods or services we are promoting. We need to have possible partner websites that are correctly classified in order to execute this effectively.

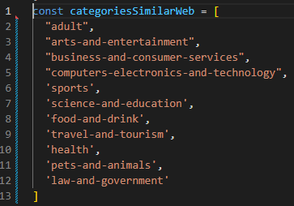
In this subject, we provide a solution to classify websites. The input is the content of website while output is a website label for each content.

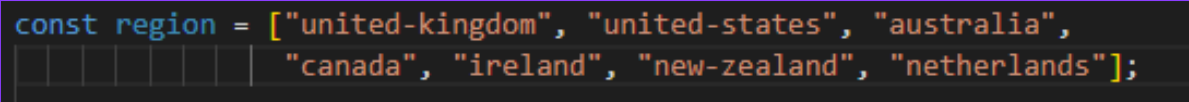
In addition, the machine learning approach we apply is Naïve Bayes classification .

# **Data Science methodology**

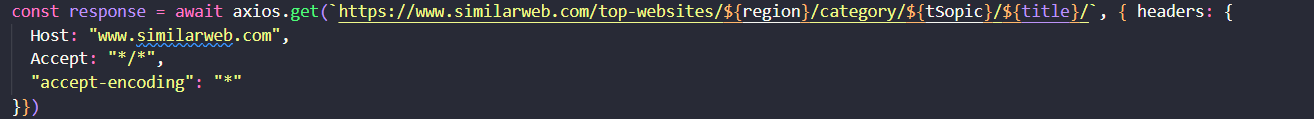
## **Data Crawling**

Our team will first use https://www.similarweb.com to compile a list of comparable websites under 11 broad categories: arts and entertainment, business and consumer services, computers electronics and technology, food and drink, health, science and education, sports, adult, travel and tourism, pets and animals, law, and government. For each issue, an average of 50–60 websites are collected, and if the information is unsatisfactory, up to 100 websites. There are just 11 websites with an adult-specific topic.



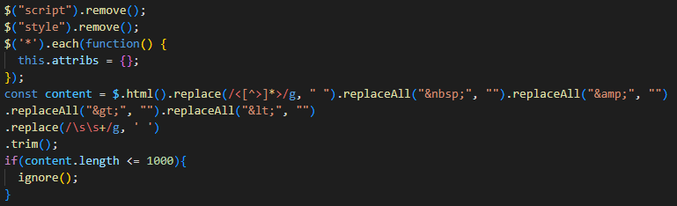
We only select websites from nations where English is widely spoken:

We crawl from websites in 3 nations using the **axios** (a library dedicated to fetching the Javascript API) tool and run a for loop across each website in each topic:



Our team removes irrelevant html, css, and js elements from data using the cheerio tool:





However, because there is a lot of crawl data from faulty websites, we have to take an extra step to filter websites that contain invalid data. Manually each member filter 2 topics, remove websites that contain non-English characters or cannot be read.

Finally, my team exported the processed data to a JSON file and then convert it to an Excel file for further processing in Python.

Ảnh có chứa văn bản

Mô tả được tạo tự động

## **Data Preprocessing**

After the completion of the data crawling stage, the team proceeds to preprocess the data using the following steps [2]:

\*Lowering all the text:



\*Remove all hyperlink , url in all documents:



\*Remove all punctuation such as “;”, “?” ,“!” and other unnecessary character or symbol:



\*Remove all html tag :



\*Remove all stop word such as a , an ,the ,…. :



\*Using SnowballStemmer to stem the all the word:





## **Data Visualization**

Our team visualizes the data after it has been cleaned and normalized in order to provide a first assessment of the obtained data set.

* 1. **The graph shows the number and percentage of valid websites crawled by topic**

1. The bar chart illustrates the amount of crawled websites

Chart, bar chart

Description automatically generated

1. The pie chart representing the percentage of the number of websites in each topic

A picture containing electronics, compact disk

Description automatically generated

* 1. **The WordCloud displays the most frequently used words by subject on websites**

*The text occurs in that topic more often when its icon is bigger.*

*Example:*

1. Sports



1. Science and Education



1. Pets and Animals

Text

Description automatically generated with low confidence

1. Travel and Tourism

Text

Description automatically generated

**3.3 The bar chart that displays top 10 most frequently used words by subject on websites:**

*Example:*

1. Sport

Chart, bar chart

Description automatically generated

1. Science and education

Chart, bar chart

Description automatically generated

c)Pet and animals

Chart, bar chart

Description automatically generated

d)Travel and tourism

Chart, bar chart, histogram

Description automatically generated

**3.4 .Vectorization for feature extraction:**

Vectorization is one of the most useful approach for converting raw data input into vectors of real number that make machine learning code become more efficient.

CountVectorizer() from sklearn library is one of the helpful feature that help us keep dictionary of every word and also the frequency of each word in each document .However, there is a problem : as can be seen from the TOP 10 WORD’s bar chart ,the word “new” appears in most of categories and it is not really a special word that support us classify categories.TF-IDF helps us to resolve that issue by decreasing the important of the word “new”

**3.4.1 .TF-IDF Overview[3]:**

**a.TF**

TF stands for Term Frequency. It estimates the frequency of a word in an document. The formula of TF is:

**b.IDF**

IDF stands for Inverse Document Frequency .It estimates the importance of a word amongst the documents. The formula of IDF is :

Obviously, by taking IDF , we can downscale the frequent words while making the infrequent words have higher impact

The final formula to caculate TF-IDF score:

**3.4.2 .Apply TfIdfVectorizer:**



We use TfidVectorizer to vectorize our clean text.There are some parameters we use:

- min\_df=0.01 means we ignore all the word that only appear less than 1% of all documents

-max\_df=0.85 means we ignore all the word that appear more than 85% of all documents

-max\_features=1500 means we only extract maximun 1500 features

-ngram\_range=(1,3) means we also consider a sequence of 2 words ,3 words

After apply and extract top importance word, we can see the word “new” not appear in top 10 importance words of pets-and-animals or travel-and-tourism

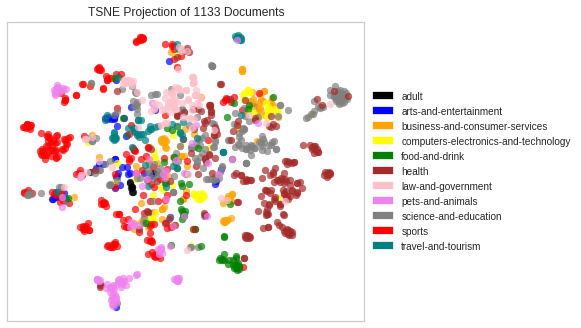
Ảnh có chứa bàn

Mô tả được tạo tự động Ảnh có chứa bàn

Mô tả được tạo tự động Ảnh có chứa bàn

Mô tả được tạo tự động

**3.4.3: t-SNE Visualization**



## **Model Training**

**4.1:Overview about Naïve Bayes Classifier[4]**

**4.1.1 MAP hypothesis:**

Given the set H of possible hypotheses, the learner find the most probable hypothesis h H given the observed data D

Such a maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis:

By Bayes theorem the above expression can convert to:

Since P(D) is the same for all classes, we can have the final formula:

**4.1.2 Naïve Bayes classifier**

Given a training set D, where each training instance x represented as an n-dimensional attribute vector : (x1,x2,…xn).A pre-defined set of classes {c1,c2,…,cm}.Given a new instance z,where should we classify z into:

From MAP hypothesis, we have:

Apply Bayes theorem :

Since is the same for all classes, the most probable class for z is identified by:

Apply conditional probability rule:

Apply Chain rule, can decompose to:

However, the above formula is not easy to compute In this case, we use assumption in Naïve Bayes classifier: the attributes are conditionally independent given classification

Finally, we have Naïve Bayes classifier to finds the most probable class for z:

**4.2.Split dataset:**

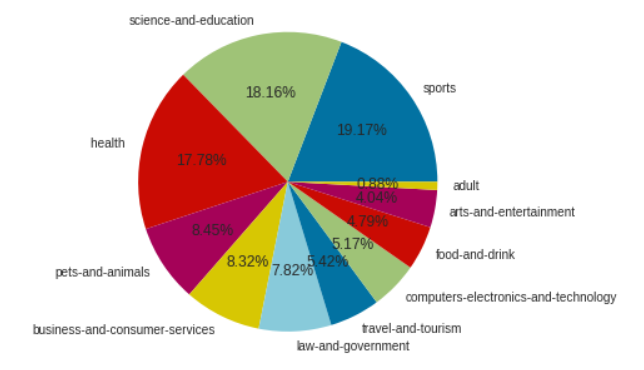
We split the dataset into train set and test set (with ratio 7:3) and also keep the ratio between categories in training set and test set similar to the original ratio between them

Ảnh có chứa văn bản

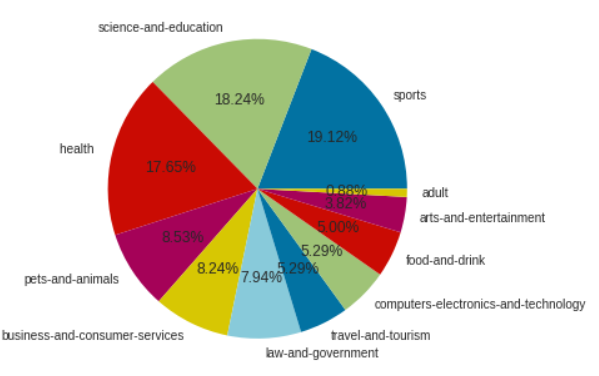
Mô tả được tạo tự động

Train set include 793 records while this number in test set is 340:

**-**Train set :

.

-Test set:



**4.3. Apply Multinomial Naïve Bayes:**

Ảnh có chứa văn bản

Mô tả được tạo tự động

# **Result**

## **Final Result**

**a)Training:**

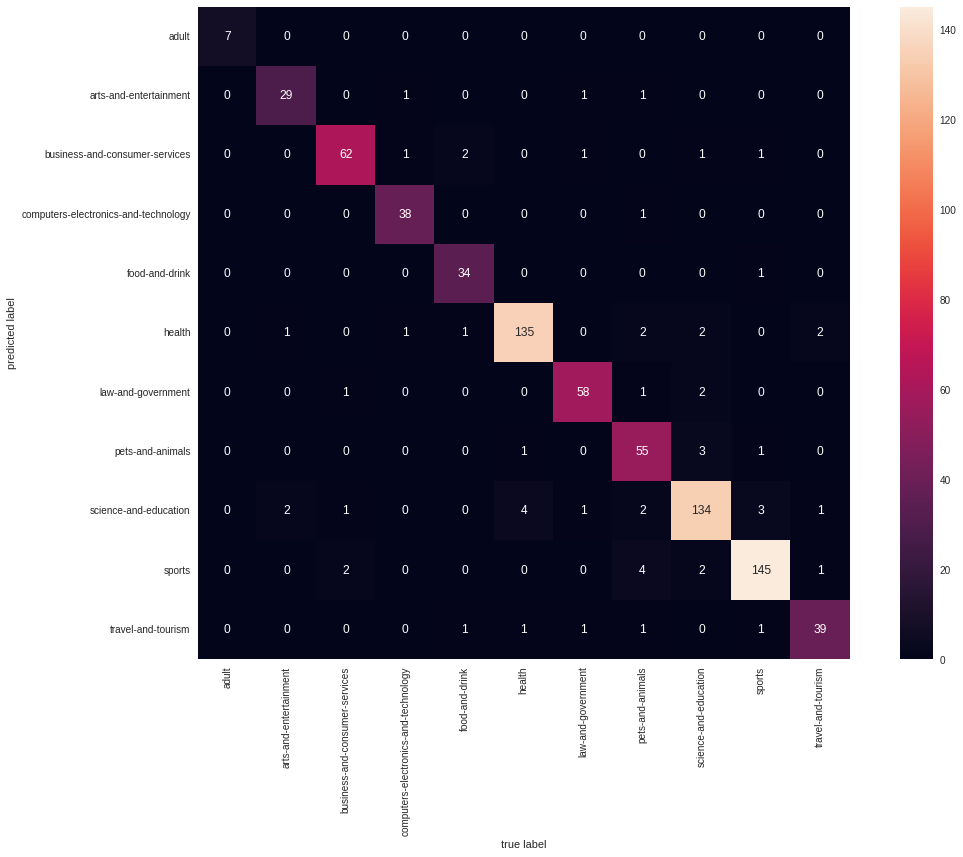
\*Accuracy: 0.93

\*Classification report:

Ảnh có chứa bàn

Mô tả được tạo tự động

Confusion Matrix:



**b)Test:**

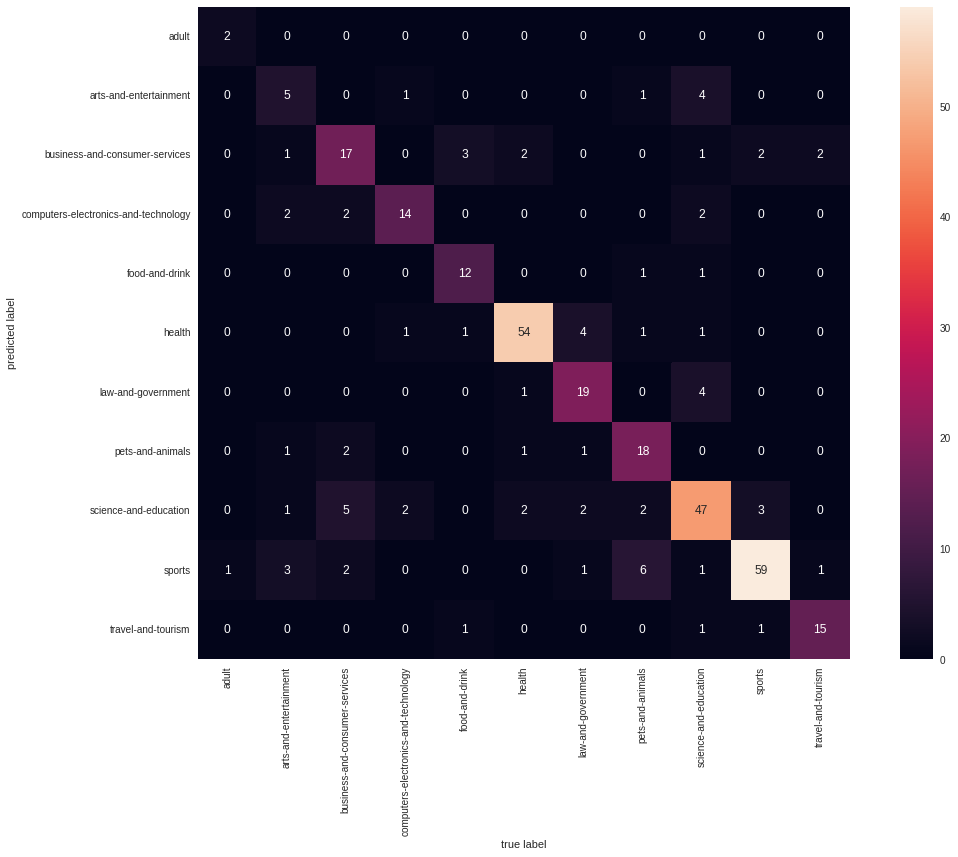
\*Accuracy:0.77

\*Classification Report:

Ảnh có chứa bàn

Mô tả được tạo tự động

\*Confusion Matrix:



## **Difficulties:**

- The collected data is not large but it is time-consuming to collect. Difficult to continue if there is an error while crawling.

- The model accuracy is not bad but also not good. Especially , art-and-entertainment category always have a low precision in test.

## **Future Improvement**

## **-** In the future, for improvement, we are going to collect more data to increase the accuracy and detect if there are any noises in our dataset

## - In addition, we will try our best to handle the multi-class problem . Because some website we have had crawled may belong to more than 1 category and this can be the reason that make our model not work so well

-We will also try another model (SVM, Random Forest ,…) to make more improvement

## 

## **IV.Conclusions**

The report summarized the process our team did, with data science methodology. From the steps of problem statement, data collection, cleaning, model training as well as model evaluation with test set.

In this project, we tried our best to collect data, although we had to deal with a lot of faulty websites and websites that preventing us from crawling. We also learned Naïve Bayes classifier -an easy-to-implement algorithm for text classification .Despite its conditional independence assumption that rarely happening in real life, Naïve Bayes shows not bad performance in website categoring

Finally, completing this project in accordance with the data science methodology gave the team a true understanding of the process. It is not a process of downloading an existing dataset and then training to refine the model, but a practical process.

# Reference

1. <https://medium.com/website-categorization/website-categorization-api-ca6c3e0f6c4d>
2. Previous project : Email spam filtering. Nguyen Minh Tuan, Nguyen Thu Hieu
3. Understanding tf idf: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>
4. Slide L5-Probabilistic\_learning ,PhD Nguyen Nhat Quang,Ha noi University of Science and Technology