|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**  **KHOA CÔNG NGHỆ THÔNG TIN I**    https://lh5.googleusercontent.com/GkgQG79WELM4Z1GCDA1-MqBpaZsNTG6Hz1WFsJUqgvM3NmfOeEBYmduqJ-WssIs8cuZxq8zz0EDJcmryhD1-wVdqH9KpRqoE0-17CI9QxTdLZN5UQOX3_tPIjYyHf-jCQ-fxTXsrfwkqYbX57ArMsQWv-nUgcDvUgWfiJWUkEscSOt5_37mlOwNLnmIFURSC  **BÁO CÁO BÀI TẬP LỚN**  **BÀI TOÁN BẺ ẢNH CAPTCHA**   |  |  | | --- | --- | | Danh sách thành viên DIPF22PRJ37: | | | 1. | Trương Thị Thu Hà – B19DCCN212 | | 2. | Phạm Thị Thanh Hảo – B19DCCN224 | | 3. | Thái Thuỳ Linh – B19DCCN380 | | Học phần: | Xử lý ảnh | | Nhóm: | 02 | | Lớp: | D19-042 | | Giảng viên: | Phạm Văn Sự |   **Hà Nội 2022** |

# Lời cảm ơn

**MỤC LỤC**

[Lời cảm ơn 2](#_Toc121508432)

[I. Giới thiệu chung 4](#_Toc121508433)

[1. Captcha là gì? 4](#_Toc121508434)

[2. Các loại captcha 4](#_Toc121508435)

[3. Bẻ ảnh captcha 5](#_Toc121508436)

[II. Cơ sở lý thuyết 5](#_Toc121508437)

[1. Công cụ lập trình 5](#_Toc121508438)

[1.1. Ngôn ngữ lập trình Python 5](#_Toc121508439)

[1.2. Thư viện sử dụng 6](#_Toc121508440)

[1.3. IDE Visual Studio Code 7](#_Toc121508441)

[2. Thuật toán học máy K – Nearest Neighbor 8](#_Toc121508442)

[2.1. Khái niệm 8](#_Toc121508443)

[2.2. Ý tưởng thuật toán 8](#_Toc121508444)

[3. Lý thuyết xử lý ảnh 9](#_Toc121508445)

[3.1. Không gian màu 9](#_Toc121508446)

[3.2. Phân vùng ảnh theo ngưỡng biên độ 9](#_Toc121508447)

[3.3. Khử nhiễu 9](#_Toc121508448)

[III. Bài toán bẻ ảnh captcha 10](#_Toc121508449)

[1. Sinh ảnh captcha 10](#_Toc121508450)

[2. Xử lý ảnh 11](#_Toc121508451)

[3. Nhận dạng ký tự 13](#_Toc121508452)

[4. Kết quả 13](#_Toc121508453)

[IV. Tài liệu tham khảo 14](#_Toc121508454)

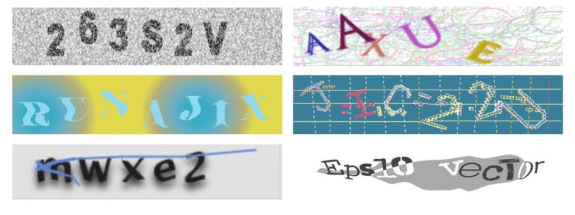
# I. Giới thiệu chung

## 1. Captcha là gì?

* Captcha là viết tắt của “Completely Automated Public Turing test to tell Computers and Humans Apart” (tạm dịch là Bài kiểm tra tính tự động để phân biệt máy tính và con người).
* Captcha ra đời nhằm ngăn chặn Spam trong Website Forum, đảm bảo an ninh bảo mật thông tin dữ liệu, lượng tương tác thật và ngăn chặn sự tấn công với mật khẩu, tạo độ tin cậy cho các thăm dò trực tuyến.

## 2. Các loại captcha

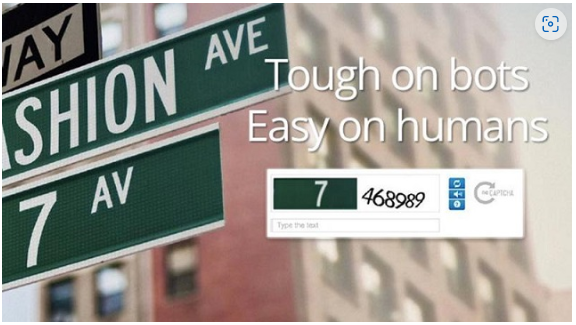
* Captcha văn bản: sử dụng một từ hoặc cụm từ kết hợp giữa chữ cái và chữ số để kiểm tra.



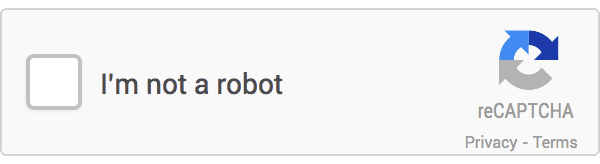
* Captcha hình ảnh: yêu cầu người dùng xác định ảnh phù hợp với chủ đề hoặc không phù hợp.



* Captcha âm thanh:



* no-Captcha:



## 3. Bẻ ảnh captcha

* Sử dụng các kỹ thuật xử lý ảnh số cùng các thuật toán phù hợp để thực hiện nhận dạng các mã captcha.
* Đầu vào: Hình ảnh mã captcha văn bản:



* Đầu ra: Mã captcha dự đoán: A72oez.

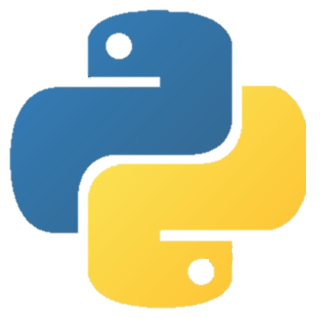
# II. Cơ sở lý thuyết

## 1. Công cụ lập trình

### 1.1. Ngôn ngữ lập trình Python

#### 1.1.1. Python là gì?

* Python là ngôn ngữ lập trình đa dụng phổ biến có thể được sử dụng cho nhiều ứng dụng. Nó bao gồm các cấu trúc dữ liệu cấp cao, gõ động, liên kết động và nhiều tính năng khác giúp nó hữu ích cho việc phát triển ứng dụng phức tạp.
* Python là ngôn ngữ phổ quát được tìm thấy trong nhiều úng dụng khác nhau.



#### 1.1.2. Ưu điểm

* Python là ngôn ngữ lập trình toàn cầu.
* Khả năng thông dịch cao: Quản lý dữ liệu dễ dàng và chuyển đổi code, điều hướng mã lệnh phong phú.
* Lập trình hướng đối tượng.
* Hoạt động tốt trên tất cả các nền tảng hệ điều hành Windows, macOS, Linux.
* Cú pháp đơn giản, dễ học, dễ nhớ, mức độ ứng dụng cao.
* Có thể mở rọng và có thể nhúng dễ dàng mọi thông tin.
* Thư viện lớn, đa dạng hỗ trợ.

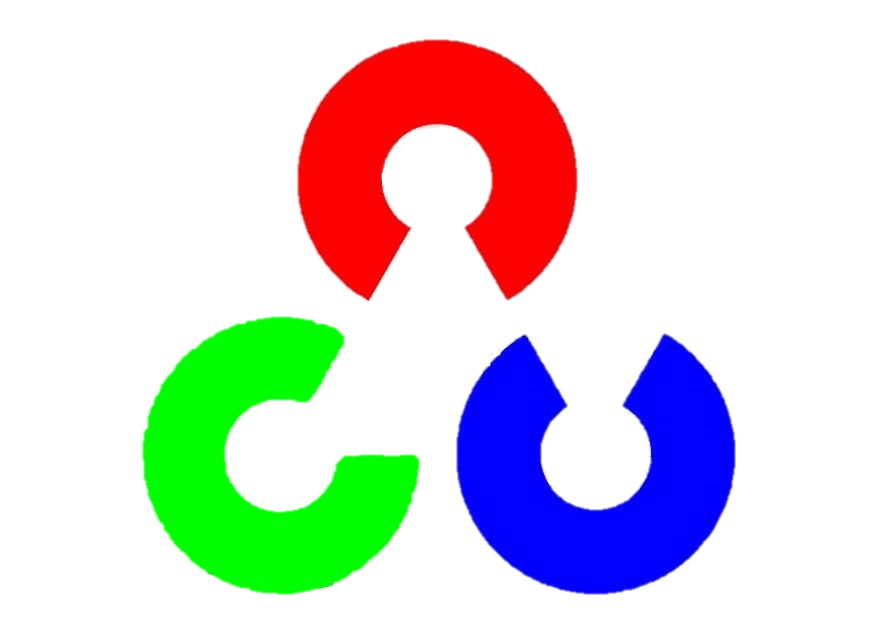
#### 1.1.3. Ứng dụng

* Phát triển web (phía máy chủ).
* Các tập lệnh Python giúp tự động hoá tác vụ.
* Khoa học dữ liệu.
* Kỹ thuật Robot.
* Python giúp kiểm thử phần mềm được tự động hoá.

### 1.2. Thư viện sử dụng

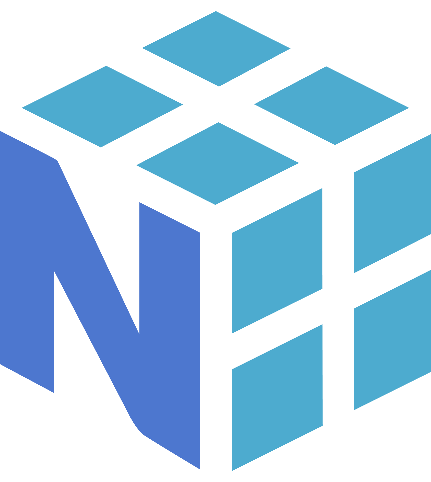
#### 1.2.1. OpenCV

* OpenCV viết tắt cho Open Source Computer Vision Library. OpenCV là thư viện nguồn mở hàng đầu cho Computer Vision và Machine Learning.
* Ứng dụng:
* Hình ảnh street view
* Kiểm tra và giám sát tự động
* Robot và xe hơi tự lái
* Phân tích hình ảnh y học
* Tìm kiếm và phục hồi hình ảnh/video
* Phim – cấu trúc 3D từ chuyển động
* Nghệ thuật sắp đặt tương tác



#### 1.2.2. Numpy

* Numpy (Numeric Python): là một thư viện toán học rât phổ biến và mạnh mẽ của Python. NumPy được trang bị các hàm số đã được tối ưu, cho phép làm việc hiệu quả với ma trận và mảng, đặc biệt là dữ liệu ma trận và mảng lớn với tốc độ xử lý nhanh hơn nhiều lần khi chỉ sử dụng Python đơn thuần.
* Sử dụng numpy:
* Các phép toán toán học và logic trên mảng.
* Các biến đổi Fourier và các quy trình để thao tác shape.
* Các phép toán liên quan đến đại số tuyến tính. Numpy tích hợp sẵn các hàm cho đại số tuyến tính và tạo số ngẫu nhiên.



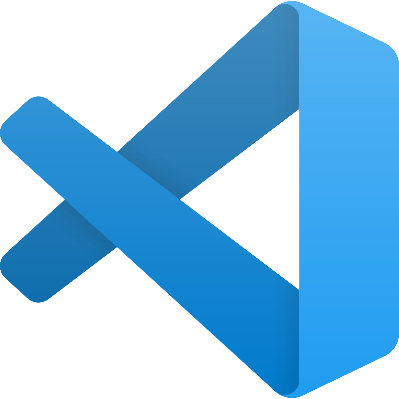
#### 1.2.3. Scikit-learn

* Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling.



### 1.3. IDE Visual Studio Code

* Visual Studio Code (VS Code) là một trình soạn thảo mã nguồn mở gọn nhẹ nhưng có khả năng vận hành mạnh mẽ trên 3 nền tảng là Windows, Linux và macOS được phát triển bởi Microsoft.
* VS Code hỗ trợ cho JavaScript, Node.js và TypeScript, cũng như cung cấp một hệ sinh thái mở rộng vô cùng phong phú cho nhiều ngôn ngữ lập trình khác.
* VS Code là sự kết hợp giữa một trình soạn thảo mã nguồn với những công cụ phát triển mạnh mẽ như Git, Debug hay Syntax Highlighter.



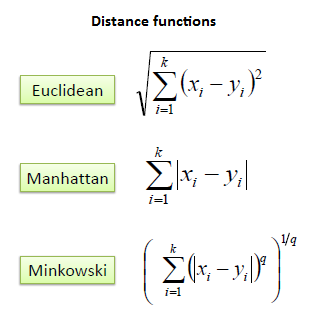
## 2. Thuật toán học máy K – Nearest Neighbor

### 2.1. Khái niệm

* K-Nearest Neighbor (KNN), thuật toán K láng giềng gần nhất.
* Thuật toán KNN là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có.

### 2.2. Ý tưởng thuật toán

* Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất.
* Việc tìm khoảng cách giữa 2 điểm có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp:



## 3. Lý thuyết xử lý ảnh

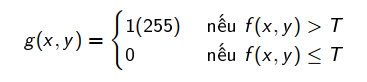
### 3.1. Không gian màu

* Không gian màu là một hệ thống xác định các màu sắc.
* Một số không gian màu phổ biến: RGB, HSV, CMYK, GRAY, YIQ, HSL, YUV,…



### 3.2. Phân vùng ảnh theo ngưỡng biên độ

* Phân vùng ảnh theo ngưỡng biên độ là thực nhiện phân vùng các ảnh một cách trực tiếp thành các vùng dựa trên các giá trị mức xám và/hoặc các tính chất của các giá trị mức xám:



* Những yếu tố chính ảnh hưởng đến việc xác định ngưỡng:
* Sự phân tách giữa các đỉnh
* Mức độ nhiễu của ảnh
* Kích thước tương dối của đối tượng so với nền
* Tính đồng nhất của nguồn chiếu sáng
* Tính đồng nhất của các tính chất phản xạ của ảnh

### 3.3. Khử nhiễu

* Nhiễu trong ảnh số được coi là sự thay đổi đột ngột của tín hiệu ảnh trên một khoảng rất nhỏ.
* Các mô hình nhiễu: Nhiễu Gausse, nhiễu Rayleigh, nhiễu xung, nhiễu Erlang, nhiễu hàm mũ, nhiễu phân bố đều,…
* Có thể ước lượng tham số của nhiễu thông qua phân tích phổ, chụp ảnh mẫu, khảo sát lược đồ xám vùng ảnh đồng nhất.
* Khử nhiễu hình ảnh đề cập đến quá trình khôi phục tín hiệu từ hình ảnh nhiễu.
* Khử nhiễu được thực hiện để loại bỏ nhiễu không mong muốn khỏi hình ảnh để phân tích hình ảnh ở dạng tốt hơn.

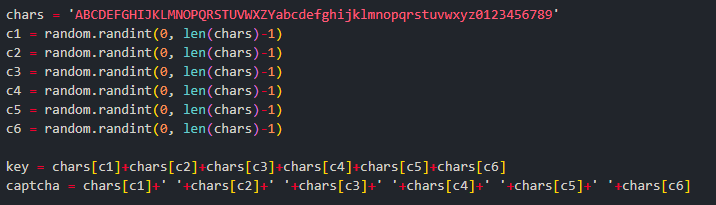
# III. Bài toán bẻ ảnh captcha

## 1. Sinh ảnh captcha

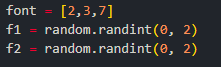
* Khởi tạo ảnh trắng kích thước 40x185:



* Sinh ngẫu nhiên mã captcha:



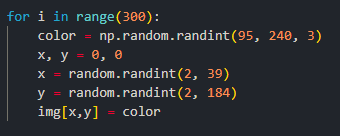
* Chọn ngẫu nhiên font có sẵn trong OpenCV:



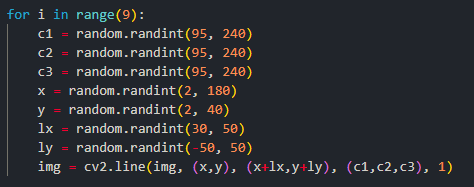
* Thêm mã captcha đã sinh vào ảnh trắng:



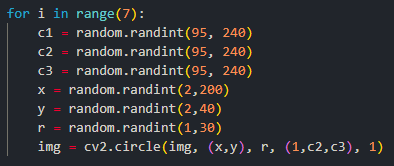
* Thêm nhiễu:
* Điểm ảnh màu:



* Đường thẳng màu:



* Đường tròn màu:



* Tăng gấp 2 lần các kích thước ảnh:



* Lưu ảnh captcha:



* Ảnh được sinh:



## 2. Xử lý ảnh

* Loại bỏ nhiễu ảnh màu:
* Ước lượng tham số của nhiễu sử dụng ảnh mẫu kích thước 20x20, là các ảnh được cắt ra từ ảnh captcha gốc.
* Kích thước bộ lọc: 80x80





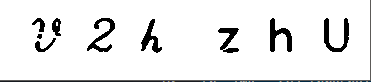
* Chuyển đổi không gian màu từ RGB sang GRAY:



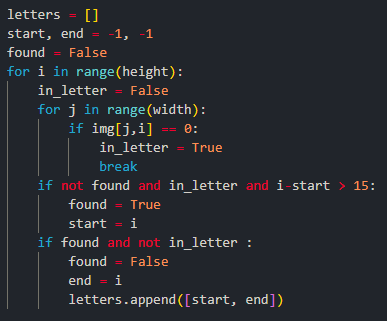


* Phân ngưỡng ảnh với ngưỡng T=150:

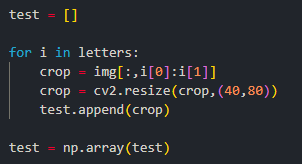




* Tìm vị trí các ký tự mã captcha:



* Cắt ảnh:



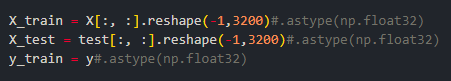


## 3. Nhận dạng ký tự

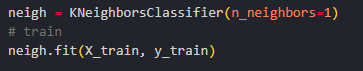
* Sử dụng thuật toán KNN.
* Tải dữ liệu huấn luyện:



* Điều chỉnh kích thước dữ liệu huấn luyện:



* Huấn luyện với K=1:



* Dự đoán mã captcha:



## 4. Kết quả

* Ảnh captcha:



* Dữ liệu đầu vào:



* Dữ liệu đầu ra:



# IV. Tài liệu tham khảo

1. <https://quantrimang.com/cong-nghe/nguon-goc-cua-captcha-va-recaptcha-159522>

2. <https://bkhost.vn/blog/captcha/>

3. <http://www.e-ptit.edu.vn/ngon-ngu-lap-trinh-python/>

4. <https://mona.media/ngon-ngu-lap-trinh-python/>

5.<https://opencv.org/>

6. <https://numpy.org/>

7. <https://scikit-learn.org/stable/>

8. <https://code.visualstudio.com/>

9.<https://vietnambiz.vn/thuat-toan-k-lang-gieng-gan-nhat-k-nearest-neighbor-knn-la-gi-2020022911113334.htm>

10.<https://codelearn.io/sharing/thuat-toan-k-nearest-neighbors-knn>

11. Giáo trình Xử lý ảnh, PGS. TS. Nguyễn Quang Hoan, Học viện Công nghệ Bưu chính Viễn thông.