**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Lập trình Python (PYPRO)**

MÃ SỐ LỚP HP: **IPPA233277\_03**

Tên chủ đề : **LẬP TRÌNH PYTHON PHÂN TÍCH DỮ LIỆU CÁC ĐỘI BÓNG NGOẠI HẠNG ANH MÙA 06/07 ĐẾN MÙA 17/18.**

**Họ tên sinh viên: NGUYỄN KHANG THỊNH [69]**

**Mã số sinh viên: 21110660**

**Lớp: 211103C [CQ]**

**Ngày nộp: ……./……./2023**

**Ký tên:**

**TP.HCM, ngày .... Tháng .... năm 2023**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Lập trình Python (PYPRO)**

MÃ SỐ LỚP HP: **IPPA233277\_03**

Tên đề tài: **Lập trình Python PHÂN TÍCH DỮ LIỆU CÁC ĐỘI BÓNG NGOẠI HẠNG ANH MÙA 06/07 ĐẾN MÙA 17/18**

**Giảng viên giảng dạy: VÕ XUÂN THỂ**

**Họ tên sinh viên: NGUYỄN KHANG THỊNH [69]**

**Mã số sinh viên: 21110660**

**Lớp: 211103C [CQ]**

**TÊN CÁC FILE SẢN PHẨM ĐỀ TÀI:**

**Tên sản phẩm đề tài: G369Thinh\_DAHP.PyPro\_EPLStats.rar**

**Tập tin mã nguồn: G369Thinh\_EPLStats.py**

**Tập dữ liệu thực nghiệm: G369Thinh\_EPLStats.csv**

**Giới thiệu thư mục bài làm: G369Thinh\_DAHP.Intro.docx**

**Tập các thư viện sử dụng:** # Speech: speech\_recognition, gtts, playsound

# GUI: tkinter

# EDA: pandas, numpy, scipy, sklearn

**NHẬN XÉT CỦA GIẢNG VIÊN**

**………………………………………………………..**

**………………………………………………………..**

**Ngày……./……../202…..**

**Giảng viên: Ký tên**

**TP.HCM, ngày …. tháng …. năm 2023**

**LỜI CẢM ƠN**

Em xin cam kết đồ án học phần này hoàn toàn do em thực hiện, không sao chép của các bạn sinh viên hay các nguồn khác.

Em đã nghiên cứu kỹ đồ án học phần và thực hiện kiểm tra nhiều lần, song không tránh khỏi sai sót. Mong nhận được sự góp ý của thầy để em có thể khắc phục và hoàn thiện hơn đồ án này.

Cuối cùng, em không thể hoàn thành đồ án này nếu không có sự giúp đỡ của thầy Võ Xuân Thể. Em chân thành cảm ơn sự giúp đỡ của thầy trong thời gian qua.

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| CSDL hoặc DB | Cơ sở dữ liệu: DataBase |
| NSD = Users = Account | Người Sử Dụng = là quyền làm việc trên WebApp được đảm bảo bởi tối thiểu là 2 yếu tố: tên đăng nhập (UserName) và mật khẩu (Password) |
|  |  |
| AI | Artificial intelligence: Trí tuệ nhân tạo |
| EDA | Exploratory Data Analysis: Phân tích Khám phá Dữ liệu (Còn gọi là: phân tích dữ liệu thăm dò) |
| GUI | Graphical User Interface: Giao diện đồ hoạ |
| PYPRO | Lập trình Python |

**DANH MỤC CÁC THUẬT NGỮ**

|  |  |
| --- | --- |
| Giao diện người dùng:  User-Interface | Là hệ thống các màn hình giao tiếp cho phép người sử dụng tương tác với các thành phần phần mềm trong HTTT, thường là 1 trong các dong: Win-form, Web-form, Mobile-Form. |
| Tài khoản (Account) | Là một quyền làm việc trên HTTT được cấp phát cho một cá nhân thông qua tên tài khoản (username) và mật khẩu (password). |
| Trợ lý ảo: Voice Assistant | Là các hệ thống có khả năng “nghe” và “nói” với con người, nhờ đó hỗ trợ con người trong một số chức năng như một “trợ lý”. |

**MỤC LỤC**

[**LỜI CẢM ƠN** 3](#_Toc130324658)

[DANH MỤC CÁC TỪ VIẾT TẮT 4](#_Toc130324659)

[**DANH MỤC CÁC THUẬT NGỮ** 4](#_Toc130324660)

[Chương 1: GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 7](#_Toc130324661)

[1.1. Tổng quan về ĐỒ ÁN HỌC PHẦN 7](#_Toc130324662)

[1.2. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN 8](#_Toc130324663)

[1.3. Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN 8](#_Toc130324664)

[1.4. Bố cục của báo cáo 8](#_Toc130324665)

[Chương 2: CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN 10](#_Toc130324666)

[2.1. GIỚI THIỆU VỀ LẬP TRÌNH PYTHON 10](#_Toc130324667)

[2.1.1. Giới thiệu chung 10](#_Toc130324668)

[2.1.2. Cài đặt “bộ thảo chương” (soạn thảo chương trình IDE) và “bộ dịch” 10](#_Toc130324669)

[2.1.3. Giới thiệu các bộ tương tác lập trình 10](#_Toc130324670)

[2.1.4. Cấu hình mô trường tương tác lập trình 10](#_Toc130324671)

[2.2. TỔNG QUAN VỀ NGÔN NGỮ LẬP TRÌNH PYTHON 10](#_Toc130324672)

[2.2.1. Danh hiệu (Identifier) 10](#_Toc130324673)

[2.2.2. Về viết mã lệnh lập trình (Code) 10](#_Toc130324674)

[2.2.3. Về chú giải (Comment) 10](#_Toc130324675)

[2.2.4. Các phép toán 10](#_Toc130324676)

[2.2.5. Chuỗi 10](#_Toc130324677)

[2.2.6. Lệnh 10](#_Toc130324678)

[2.2.7. Định nghĩa hàm 10](#_Toc130324679)

[2.2.8. Danh sách 10](#_Toc130324680)

[2.2.9. Bộ dữ liệu: Tuple 10](#_Toc130324681)

[2.2.10. Tập hợp: Set 10](#_Toc130324682)

[2.2.11. Từ điển: Dictionary 10](#_Toc130324683)

[2.2.12. Module 10](#_Toc130324684)

[2.2.13. Package 10](#_Toc130324685)

[2.2.14. Lớp 10](#_Toc130324686)

[2.2.15. Xử lý ngoại lệ 10](#_Toc130324687)

[2.3. GIỚI THIỆU CÁC THƯ VIỆN VÀ CÔNG CỤ CĂN BẢN 10](#_Toc130324688)

[2.3.1. Nạp thư viện 10](#_Toc130324689)

[2.3.2. Pandas 10](#_Toc130324690)

[2.3.3. Tensorflow 10](#_Toc130324691)

[2.3.4. NumPy 10](#_Toc130324692)

[2.3.5. SCIPY 10](#_Toc130324693)

[2.3.6. Theano 10](#_Toc130324694)

[2.3.7. Matplotlib 10](#_Toc130324695)

[2.3.8. scikit-learn = sklearn 10](#_Toc130324696)

[2.3.9. Keras 10](#_Toc130324697)

[2.3.10. NLTK 11](#_Toc130324698)

[2.3.11. PyTorch 11](#_Toc130324699)

[2.3.12. LightGBM 11](#_Toc130324700)

[2.3.13. Eli5 11](#_Toc130324701)

[2.4. LẬP TRÌNH PYTHON XỬ LÝ GIỌNG NÓI (Trợ lý ảo: Voice Assistant) 11](#_Toc130324702)

[2.4.1. Giới thiệu chủ đề 11](#_Toc130324703)

[2.4.2. Tài liệu và nguồn thực nghiệm 11](#_Toc130324704)

[2.4.3. Giới thiệu Các thư viện Python liên quan 11](#_Toc130324705)

[2.4.4. Giới thiệu bài toán 11](#_Toc130324706)

[2.4.5. Giới thiệu code tham khảo 11](#_Toc130324707)

[2.5. KỸ THUẬT LẬP TRÌNH ĐỆ QUY: THÁP HÀ NỘI (HANOI TOWER) 11](#_Toc130324708)

[2.6. LẬP TRÌNH PHÂN TÍCH KHÁM PHÁ (EDA) [Thăm dò] 11](#_Toc130324709)

[Chương 3: PHÂN TÍCH VÀ XÁC ĐINH CÁC CƠ SỞ KỸ THUẬT 11](#_Toc130324710)

[3.1. GIỚI THIỆU CHUNG VỀ CHỦ ĐỀ 11](#_Toc130324711)

[3.1.1. Mô tả yêu câu của chủ đề 11](#_Toc130324712)

[3.1.2. Phân tích yều cầu của chủ đề 12](#_Toc130324713)

[3.2. CÁC NỀN TẢNG KỸ THUẬT 12](#_Toc130324714)

[3.2.1. Hệ thống thư viện sử dụng 12](#_Toc130324715)

[3.2.2. Tập dữ liệu thực nghiệm 12](#_Toc130324716)

[3.2.3. Mô tả giải thuật 13](#_Toc130324717)

[3.3. LẬP TRÌNH XỬ GIỌNG NÓI (TRỢ LÝ ẢO: VOICE ASSISTANT) 14](#_Toc130324718)

[3.3.1. Giới thiệu chung về chủ đề 14](#_Toc130324719)

[3.3.1.1. Mô tả yêu cầu của chủ đề 14](#_Toc130324720)

[3.3.1.2. Phân tích yều cầu của chủ đề 14](#_Toc130324721)

[3.3.2. Các nền tảng kỹ thuật 15](#_Toc130324722)

[3.3.2.1. Nền tảng kỹ thuật cơ bản 15](#_Toc130324723)

[3.3.2.2. Hệ thống thư việc sử dụng 15](#_Toc130324724)

[3.3.3. Thiết kế giao diện 15](#_Toc130324725)

[3.3.4. Mã lệnh lập trình (python) 15](#_Toc130324726)

[Chương 4: LẬP TRÌNH PYTHON PHÂN TÍCH ........................................ 17](#_Toc130324727)

[4.1. Thiết kế giao diện 17](#_Toc130324728)

[4.2. Mã lệnh lập trình (python) 17](#_Toc130324729)

[Chương 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN 21](#_Toc130324730)

[5.1. Kết luận 21](#_Toc130324731)

[5.1.1. Những kết quả đạt được 21](#_Toc130324732)

[5.1.2. Hạn chế 21](#_Toc130324733)

[5.2. Hướng phát triển 21](#_Toc130324734)

[5.2.1. Hướng khắc phục các hạn chế 21](#_Toc130324735)

[5.2.2. Hướng mở rộng ĐỒ ÁN HỌC PHẦN 21](#_Toc130324736)

[DANH MỤC TÀI LIỆU THAM KHẢO 22](#_Toc130324737)

# GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN

## Tổng quan về ĐỒ ÁN HỌC PHẦN

Đồ án Học phần Lập trình Python Phân tích dữ liệu các đội bóng Ngoại hạng Anh mùa 06/07 đến mùa 17/18: **EPLStats** vận dụng kiến thức và kỹ năng về Lập trình Python để xây dựng giao diện đồ hoạ (GUI) cho phép: tương tác với người dùng bằng giọng nói (Trợ lý ảo: Voice Assistant) và phân tích thăm dò (EDA: Epolore Data Analysing) đối với tập dữ liệu thực nghiệm về Dữ liệu các đội bóng Ngoại hạng Anh mùa 06/07 đến mùa 17/18.

## Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN

Bao gồm 3 nội dung lập trình chính dùng ngôn ngữ lập trình Python:

+ Trợ lý ảo

+ Phân tích dữ liệu khám phá || thăm dò

+ Giao diện đồ họa dạng WinForm

## Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN

+ Ngôn ngữ lập trình **Python**

+ Công cụ lập trình và biên dịch: **Spyder (Anaconda3)**

+ Các thư viện chính:

# B1: NẠP THƯ VIỆN

# Speech

import speech\_recognition as sr

from gtts import gTTS

import playsound

# tkinter

import tkinter as tk

from tkinter import messagebox

# thư viện OS (lập thư mục, files)

import os # THƯ VIỆN OS MS. WINDOWS = Lập thư mục & lưu file

# Nạp các thư viện cần thiết EDA

import numpy as np #Numeric Python: Thư viện về Đại số tuyến tính tính

import pandas as pd #Python Analytic on Data System: For data processing(Thư viện xử lý dữ liệu)

from scipy import stats # thư viện cung cấp các công cụ thống kê [statistics]   
 sub-lib của science python [các công cụ khoa học]

from sklearn import preprocessing # Thư viện tiền xử lý DL (XL ngoại lệ: Isolated)

from sklearn.feature\_selection import SelectKBest, chi2   
 # Nạp hàm Thư viện phân tích dữ liệu thăm dò

## Bố cục của báo cáo

Báo cáo gồm những nội dung như sau:

Chương 1: Giới thiệu đồ án của Học phần

Chương 2: Các cơ sở lý thuyết của Học phần phục việc thực hiện đề tài

Chương 3: Phân tích và xác đinh các cơ sở kỹ thuật thực hiện đề tài

Chương 4: Lập trình xây dựng sản phầm đề lài: PHÂN TÍCH ........................................: Keyword

Chương 5 Tổng kết các kết quản đạt được và còn hạn chế của đồ án, đồng thời đề xuất hướng khắc phục hạn chế và phát triển Đồ án.

# CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN

## GIỚI THIỆU VỀ LẬP TRÌNH PYTHON

### Giới thiệu chung

### Cài đặt “bộ thảo chương” (soạn thảo chương trình IDE[[1]](#footnote-1)) và “bộ dịch”

### Giới thiệu các bộ tương tác lập trình

### Cấu hình mô trường tương tác lập trình

## TỔNG QUAN VỀ NGÔN NGỮ LẬP TRÌNH PYTHON

### Danh hiệu (Identifier)

### Về viết mã lệnh lập trình (Code)

### Về chú giải (Comment)

### Các phép toán

### Chuỗi

### Lệnh

### Định nghĩa hàm

### Danh sách

### Bộ dữ liệu: Tuple

### Tập hợp: Set

### Từ điển: Dictionary

### Module

### Package

### Lớp

### Xử lý ngoại lệ

## GIỚI THIỆU CÁC THƯ VIỆN VÀ CÔNG CỤ CĂN BẢN

### Nạp thư viện

### Pandas

### Tensorflow

### NumPy

### SCIPY

### Theano

### Matplotlib

### scikit-learn = sklearn

### Keras

### NLTK

### PyTorch

### LightGBM

### Eli5

## LẬP TRÌNH PYTHON XỬ LÝ GIỌNG NÓI (Trợ lý ảo: Voice Assistant)

### Giới thiệu chủ đề

### Tài liệu và nguồn thực nghiệm

### Giới thiệu Các thư viện Python liên quan

### Giới thiệu bài toán

### Giới thiệu code tham khảo

**[1] Nghe tiếng Việt => Text**

**[2] Text => đọc tiếng Việt**

## KỸ THUẬT LẬP TRÌNH ĐỆ QUY: THÁP HÀ NỘI (HANOI TOWER)

## LẬP TRÌNH PHÂN TÍCH KHÁM PHÁ (EDA) [Thăm dò]

1. **PHÂN TÍCH VÀ XÁC ĐINH CÁC CƠ SỞ KỸ THUẬT**

## GIỚI THIỆU CHUNG VỀ CHỦ ĐỀ

### Mô tả yêu câu của chủ đề

Vận dụng kỹ thuật lập trình EDA kết hợp với GUI và Voice Assistant cho phép dùng tập dữ liệu thực nghiệm về doanh số bán hàng của các chi nhánh của hệ siêu thị để dự đoán và tư vấn các nhà quản lý hoạch định kế hoạch kinh doanh của hệ thống siêu thi.

Dữ liệu thực nghiệm gồm doanh số bán hàng trong quá khứ của hệ thống siêu thị:

sttHoTen\_weatherAUS.csv : dạng file Excel .csv

. Số records: ……….

. Số thuộc tính input: …………. để đoán thuộc tính thứ ???: target variable:

= “RainTomorrow”[Y|N]

### Phân tích yều cầu của chủ đề

+ Chủ đề lập trình python này là cơ sở của các ứng dụng về DataWarehouse, Bigdata, Data Analysis, Data Sciences, . . .

+ Đây là xử lý nền tảng cho các nội dung chuyên môn (nêu trên) chủ yếu là giai đoạn tiền xử lý dữ liệu đầu vào (Input Preporcessing), như: xử lý NULL, lọc dữ liệu, tích hợp dữ liệu (Integration),…

**Một số trường hợp tiền xử lý dữ liệu đầu vào (input) của các hệ thống EDA**

[1] Xử lý NULL là ……………

[2] Xử lý dữ liệu quá khác biệt = “ngoại lệ” (isolated: ….) là……………

## CÁC NỀN TẢNG KỸ THUẬT

### Hệ thống thư viện sử dụng

**Bước 1: Nhập các thư viện cần thiết**

# AI/weatherAUS.py

# For linear algebra (Thư viện về Đại số tuyến tính tính)

import numpy as np

# For data processing (Thư viện xử lý dữ liệu)

import pandas as pd

### Tập dữ liệu thực nghiệm

**Bước 2: Tải tập dữ liệu**

#Load the data set (Nạp tập dữ liệu)

df = pd.read\_csv('./weatherAUS.csv')

#Display the shape of the data set (Vẽ biểu đồ tập dữ liệu)

print('Độ lớn của bảng [frame] dữ liệu thời tiết:',df.shape)

#Display data (Hiển thị dữ liệu dạng mảng 5 dòng đầu)

print(df[0:5])

Chạy python G3B5EX3DA\_sttHoTen.py sẽ hiển thị kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

### Mô tả giải thuật

**Bước 3: Xử lý dữ liệu**

# Checking for null values (Kiểm tra giá trị null = đếm số dòng có dữ liệu ứng từng thuộc tính)

print(df.count().sort\_values())

**Kết quả:**

Sunshine 86442

Evaporation 94845

Cloud3pm 105192

Cloud9am 111072

Pressure9am 170333

Pressure3pm 170342

WindDir9am 176693

WindGustDir 176954

WindGustSpeed 177055

WindDir3pm 183702

Humidity3pm 183959

Temp3pm 184959

WindSpeed3pm 185025

RISK\_MM 186393

RainTomorrow 186393

RainToday 186394

Rainfall 186394

Humidity9am 187603

WindSpeed9am 188080

Temp9am 188599

MinTemp 188671

MaxTemp 188876

Location 191431

Date 191431

dtype: int64

Text

Description automatically generated with medium confidence

## LẬP TRÌNH XỬ GIỌNG NÓI (TRỢ LÝ ẢO: VOICE ASSISTANT)

## Giới thiệu chung về chủ đề

### Mô tả yêu cầu của chủ đề

+ Lập trình App xử lý "lời nới" (Speech): Ứng dụng “Trợ lý ảo” (Voice Assistant)

+ Gồm 2 phần (Part) cơ bản sau

1. Nghe tiếng Việt => Text: ……………….

2. Trả lời bằng tiếng Việt : Text => Nói tiếng Việt: ……………….

### Phân tích yều cầu của chủ đề

+ Chủ đề lập trình python này là cơ sở của các ứng dụng về AI, ML, Deep ML.

+ Các hệ thống này thường có 3 phần :

Part1: Speech….: chính là part1 của chủ đề này

Part2: Các kỹ thuật AI

Part2: Voice….: chính là part2 của chủ đề này

## Các nền tảng kỹ thuật

### Nền tảng kỹ thuật cơ bản

### Hệ thống thư việc sử dụng

B1: Cài đặt thư viện

conda install …. Các thư viện sau:

------------------gốc hướng dẫn / dùng python trong visual (pip3)

pip3 install SpeechRecognition

*pip3 install SpeechRecognition pydub # Long audio source*

pip3 install pyaudio

B2 : Nạp thư viện vào App

import speech\_recognition as sr

**import** speech\_recognition **as** sr

**from** gtts **import** gTTS

**import** os

**import** time

**import** playsound

## Thiết kế giao diện

+ Xử lý trong cửa sổ Console Output của Spyder

## Mã lệnh lập trình (python)

**FULL CODE**

# -\*- coding: utf-8 -\*-

"""

Created on Sat Jul 16 19:59:48 2023

@author: VOXUAN

"""

# B1: NẠP THƯ VIỆN

import speech\_recognition as sr

from gtts import gTTS

import playsound

# B2: CHỌN PHƯƠNG ÁN NHẬP ÂM THANH TỪ MICROPHONE

r = sr.Recognizer()

# B3: XỬ LÝ NHẬN DIỆN MIC

#PHẦN 1: IN RA TEXT THEO LỜI NÓI = Nghe tiếng Việt => Text

with sr.Microphone() as Source:

#hiệu chỉnh mic để chuẩn bị nói

print("Hieu chinh nhieu trươc khi noi!")

r.adjust\_for\_ambient\_noise(Source, duration=1)

#nhận lời nói của người dùng từ MIc mặc định lưu dữ liệu âm thanh vào audio\_data

print("Nói tiếng Việt đi, sau 5s sẽ in ra văn bản!")

audio\_data = r.record(Source, duration = 5)

#In ra văn bản text

print("KẾT QUẢ NHẬN DIỆN ..................")

#chuyển lời nói thành văn bản

try:

text = r.recognize\_google(audio\_data,language="vi")

except:

text = "Quý vị nói gì nghe không rõ...!"

#in kết quả ra

print("Quý vị đã nói là : ",format(text))

#PHẦN 2: XUẤT RA LỜI NÓI THEO VĂN BẢN ĐÃ NHẬP = Trả lời bằng tiếng Việt : Text => Nói tiếng Việt

# ĐỌC CHỮ VIỆT => NÓI RA TIẾNG VIỆT

def VXT(t):

vx = gTTS(text=t, lang = 'vi')

xuan = '06VXT.mp3'

vx.save(xuan)

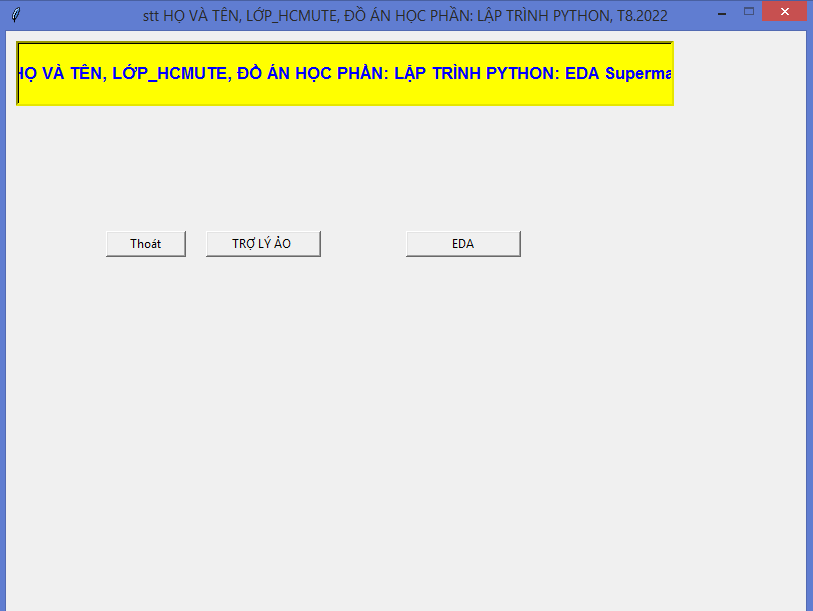
playsound.playsound(xuan)

VXT(text)

#VXT("số thứ tự 06, họ tên: VÕ XUÂN THỂ là giảng viên viên của TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH")

1. **LẬP TRÌNH PYTHON PHÂN TÍCH ........................................**

## Thiết kế giao diện



## Mã lệnh lập trình (python)

# -\*- coding: utf-8 -\*-

"""

Created on Sat Jul 16 19:59:48 2023

@author: VOXUAN

"""

# B1: NẠP THƯ VIỆN

# Speech

import speech\_recognition as sr02vtth

from gtts import gTTS

import playsound

# tkinter

import tkinter as tk02vtth

from tkinter import messagebox

# thư viện OS (lập thư mục, files)

import os # THƯ VIỆN OS MS. WINDOWS = Lập thư mục & lưu file

# Nạp các thư viện cần thiết EDA

import numpy as np #Numeric Python: Thư viện về Đại số tuyến tính tính

import pandas as pd #Python Analytic on Data System: For data processing (Thư viện xử lý dữ liệu)

from scipy import stats # thư viện cung cấp các công cụ thống kê [statistics] sub-lib của science python [các công cụ khoa học]

from sklearn import preprocessing # Thư viện tiền xử lý DL (XL ngoại lệ: Isolated)

from sklearn.feature\_selection import SelectKBest, chi2 # Nạp hàm Thư viện phân tích dữ liệu thăm dò

# B2: KHAI BÁO TÊN THƯ MỤC & FILE LƯU CÁC THÔNG TIN BÀI LÀM

vtth02\_FILE = "vtth020.mp3" # lưu tên file Input

vtth02\_DIR = 'vtth02' # Thư mục lưu các file [trên]

os.makedirs(vtth02\_DIR, exist\_ok=True) # TẠO THƯ MỤC LƯU (từ thư viện os - của OS MS. Windows))

"""

CÁC HÀM THỰC HIỆN CÁC CHỨC NĂNG: HỆ THỐNG

"""

def Thoat():

traloi = messagebox.askquestion("Xác nhận","Thiệt thoát không (Y/N)?")

if traloi == "yes": wn.destroy()

"""

CÁC HÀM THỰC HIỆN CÁC CHỨC NĂNG = SPEECH

"""

def Lenh(): # NHẬP ÂM THANH TỪ MICROPHONE

r = sr02vtth.Recognizer()

with sr02vtth.Microphone() as Source:

#hiệu chỉnh mic để chuẩn bị nói

messagebox.showinfo("Nhắc nhở", "Hieu chinh nhieu trươc khi noi!")

r.adjust\_for\_ambient\_noise(Source, duration=1)

#nhận lời nói ra lệnh từ người dùng thông qua MIc [mặc định] lưu dữ liệu âm thanh vào audio\_data

messagebox.showinfo("Cảnh báo", "Bấm OK để bắt đầu Chọn lệnh bằng tiếng Việt, trong 3s" )

audio\_data = r.record(Source, duration = 3)

try:

vlenh = r.recognize\_google(audio\_data,language="vi")

except:

vlenh = "Quý vị nói gì nghe không rõ...!"

# xuất kết quả ra

messagebox.showinfo("Quý vị đã nói là", format(vlenh))

vText = gTTS(text=vlenh, lang = 'vi')

#vFile = '06VXT.mp3'

vText.save(vtth02\_FILE)

playsound.playsound(vtth02\_FILE)

"""

CÁC HÀM THỰC HIỆN CÁC CHỨC NĂNG = EDA

"""

def EDA():

df = pd.read\_csv('./GxttHoTen\_Topic.csv')

# Display the shape of the data set (xem lượng dòng & cột dữ liệu của tập DL gốc)

messagebox.showinfo("Độ lớn của bảng [frame] dữ liệu thời tiết", df.shape)

""""""""""""""""""""""""""""""""""""""""""""

""" CÁC NỘI DUNG SAU CHƯA THỰC HIỆN

""""""""""""""""""""""""""""""""""""""""""

# Bước 3: Xử lý CỘT dữ liệu NULL quá nhiều OR không có giá trị phân tích

# Checking for null values (Kiểm tra giá trị null = đếm số dòng có dữ liệu ứng từng thuộc# tính)

messagebox.showinfo("CÁC CỘT DỬ LIỆU SẮP XẾP THEO THỨ TỰ CÓ DỮ LIỆU TỪ ÍT => NHIỀU",df.count().sort\_values()) #df.count(): đếm số lượng dòng có dữ liệu của df, .sort\_values() sx tăng dân

df = df.drop(columns=['Sunshine','Evaporation','Cloud3pm','Cloud9am','Location','Date','RISK\_MM'],axis=1)

#df = df.drop(columns=['Sunshine','Evaporation','Cloud3pm','Cloud9am','Pressure9am',# 'Pressure3pm','WindDir3pm', 'WindDir9am', 'WindGustDir',# 'WindGustSpeed','Location','Date','RISK\_MM'],axis=1)

messagebox.showinfo("Độ lớn của bảng [frame] dữ liệu SAU KHI XỬ LÝ CỘT NULL", df.shape) # kiểm tra lại số lượng cột & dòng của df sau khi XL NULL cột

# Bước 4: Xử lý DÒNG dữ liệu NULL

# Removing null values (Xóa tất cả các dòng có giá trị null trong tập FRAME dữ liệu.)

df = df.dropna(how='any')

messagebox.showinfo("Độ lớn của bảng [frame] dữ liệu SAU KHI XỬ LÝ DÒNG NULL", df.shape) # kiểm tra lại số lượng cột & dòng của df sau khi XL NULL các dòng DL

# Bước 5: Xử lý loại bỏ các giá trị ngoại lệ (cá biệt): isolated

#kiểm tra tập dữ liệu có bất kỳ ngoại lệ nào không

z = np.abs(stats.zscore(df.\_get\_numeric\_data())) # Dò tìm và lấy các giá trị cá biệt trong tập dữ liệu gốc thông qua điểm z (z\_score)

messagebox.showinfo("MA TRAN Z-SCORE", z) # in ra tập (ma trận) các giá trị z-score từ tập dữ liệu gốc

df= df[(z < 3).all(axis=1)] # kiểm tra và chỉ giữ lại trong df các giá trị số liệu tưng ứng với z-score < 3 # {loại các giá trị >= 3} vì các giá trị z-score >=3 tướng ứng với số liệu quá khác biệt so với các số liệu còn lại (“cá biệt” = “ngoại lệ” = isolated}

messagebox.showinfo("Độ lớn của bảng [frame] dữ liệu SAU KHI XỬ LÝ NGOẠI LỆ", df.shape)# xác định số dòng & cột dữ liệu sau khu xử lý các giá trị cá biệt

# Bước 6: Thay thế các vị trí giá trị 0 và 1 bởi CÓ (Yes) và KHÔNG (No).

#Thay thế yes (CÓ) and no (KO) vào vị trí giá trị 1 (Y) và 0 (N) tương ứng cột|biến RainToday và# RainTomorrow

df['RainToday'].replace({'KHONG': 'No', 'CO': 'Yes'},inplace = True)

df['RainTomorrow'].replace({'KHONG': 'No', 'CO': 'Yes'},inplace = True)

#Bước 7: Chuẩn hóa (Rời rạc hóa) tập dữ liệu Input dùng ..MaxMin

# # CHUẨN HÓA DL

# scaler = preprocessing.MinMaxScaler() #preprocessing là Sub-Library của thư viện sklearn=> hàm .MinMaxScaler() Rời rạc hóa tập dữ liệu Input

# scaler.fit(df)

# df = pd.DataFrame(scaler.transform(df), index=df.index, columns=df.columns) # Rời rạc hóa số liệu theo thang đo scaler

# df.iloc[4:10]

# messagebox.showinfo("Độ lớn của bảng [frame] dữ liệu SAU KHI CHUẨN HÓA DL", df.shape)# xác định số dòng & cột dữ liệu sau khu xử lý các giá trị cá biệt

# # GIAI ĐOẠN 3: PHÂN TÍCH DỮ LIỆU THĂM DÒ : EDA

# # Bước 8: Nạp các thuộc tính quan trọng vào Dataset

# #The important features are put in a data frame

# df = df[['Humidity3pm','Rainfall','RainToday','RainTomorrow']]

# # Bước 9: thực hiện các tính toán trên mô hình phân tích

# #To simplify computations we will use only one feature (Humidity3pm) to build the model

# X = df

# X = df[['Humidity3pm']]

# y = df[['RainTomorrow']]

# X = df.loc[:,df.columns!='RainTomorrow']

# y = df[['RainTomorrow']]

# selector = SelectKBest(chi2, k=3)

# selector.fit(X, y)

# X\_new = selector.transform(X)

# df(['Rainfall', 'Humidity3pm', 'RainToday'], dtype='object')

# messagebox.showinfo(" KÊT QUẢ", X.columns[selector.get\_support(indices=True)])# xác định số dòng & cột dữ liệu sau khu xử lý các giá trị cá biệt

"""

"""

THỦ TỤC CHÍNH = GUI

"""

# B3: LẬP GUI (EX4)

# Tạo một cửa sổ mới

wn = tk02vtth.Tk()

#Thêm tiêu đề cho cửa sổ

wn.title("stt HỌ VÀ TÊN, LỚP\_HCMUTE, ĐỒ ÁN HỌC PHẦN: LẬP TRÌNH PYTHON, T8.2023")

#Đặt kích thước của cửa sổ

wn.geometry('800x600')

#Không cho thay đổi size

wn.resizable(tk02vtth.FALSE, tk02vtth.FALSE)

#Tiêu đề Form = tên đề tài

t = "stt HỌ VÀ TÊN, LỚP\_HCMUTE, ĐỒ ÁN HỌC PHẦN: LẬP TRÌNH PYTHON: EDA Supermarket"

lblDT = tk02vtth.Label(wn, text=t, background = "yellow", fg = "blue", relief = tk02vtth.SUNKEN, font=("Arial Bold", 13), borderwidth = 3, width = 65, height = 3)

lblDT.place(x = 10, y = 10)

# B4: CÁC NÚT LỆNH: EX1 = Speech

#Thoát

btnThoat = tk02vtth.Button(wn, text = "Thoát", width = 10, command = Thoat)

btnThoat.place(x =100, y = 200)# căn cứ vào kích thước form [wn.geometry("800x600")] => canh vị trí Button "thoát"

#Xử lý lời nói = speech = VOICE ASSISTANT

btnNoi = tk02vtth.Button(wn, text = "TRỢ LÝ ẢO ", width = 15, command = Lenh)

btnNoi.place(x = 200, y = 200)

# B5: CÁC NÚT LỆNH: EX3 = EDA = DỰ PHÂN TÍCH DỮ LIỆU KHÁM PHÁ

btnEDA = tk02vtth.Button(wn, text = "EDA", width = 15, command = EDA)

btnEDA.place(x = 400, y = 200)

#Lặp vô tận để hiển thị cửa sổ

wn.mainloop()

1. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN

## Kết luận

## Những kết quả đạt được

ĐỒ ÁN HỌC PHẦN đã thực hiện việc Lập trình Python …………..; bao gồm:

+ Giao diện đồ họa: GUI

+ Xử lý giọng nói trợ lý ảo: Voice Assistant

+ Phân tích khám phá trên tập dữ liệu thực nghiệm: …………………….

## Hạn chế

Vì thời gian có hạn nên chưa liên kết voice assistant để ra lệnh phân tích khám phá bằng lời nói.

## Hướng phát triển

## Hướng khắc phục các hạn chế

Đồ án sẽ tiếp tục thực hiện iên kết voice assistant để ra lệnh phân tích khám phá bằng lời nói

## Hướng mở rộng ĐỒ ÁN HỌC PHẦN

Sẽ tiếp tục mở rộng sản phẩm đề tài với các ứng dụng thực tiễn.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Võ Xuân Thể (2023), Tài liệu học tập Lập trình Python, Trường Đại học Sư phạm kỹ thuật Tp.HCM, Tài liệu lưu hành nội bộ được cung cấp theo từng buổi học
2. Võ Xuân Thể (2023), Bài tập thực hành Lập trình Python, Trường Đại học Sư phạm kỹ thuật Tp.HCM, Tài liệu lưu hành nội bộ được cung cấp theo từng buổi học
3. …

1. IDE (Integrated Development Environment) là môi trường tích hợp dùng để viết code để phát triển ứng dụng [↑](#footnote-ref-1)