Họ tên: Nguyễn Tiến Hào

20214037

Check spam email:

import pandas as pd

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.model\_selection import train\_test\_split

from sklearn.naive\_bayes import MultinomialNB

# Bước 1: Đọc dữ liệu từ file CSV

data = pd.read\_csv('spam.csv', encoding='ISO-8859-1')

# Bước 2: Giữ lại các cột cần thiết và đổi tên các cột

data = data[['v1', 'v2']]

data.columns = ['label', 'text']

# Bước 3: Làm sạch dữ liệu

data['text'] = data['text'].str.replace(r'\W', ' ').str.lower()

# Bước 4: Chia dữ liệu thành tập huấn luyện và tập kiểm tra

X\_train, X\_test, y\_train, y\_test = train\_test\_split(data['text'], data['label'], test\_size=0.2, random\_state=42)

# Bước 5: Chuyển đổi văn bản thành đặc trưng số sử dụng TF-IDF

vectorizer = TfidfVectorizer(max\_features=3000)

X\_train\_tfidf = vectorizer.fit\_transform(X\_train)

X\_test\_tfidf = vectorizer.transform(X\_test)

# Bước 6: Xây dựng mô hình Naive Bayes

model = MultinomialNB()

model.fit(X\_train\_tfidf, y\_train)

# Bước 7: Hàm kiểm tra email nhập vào là spam hay ham

def predict\_email(email\_text):

    # Làm sạch email

    clean\_email = email\_text.lower().replace(r'\W', ' ')

    # Chuyển đổi email thành đặc trưng số

    email\_tfidf = vectorizer.transform([clean\_email])

    # Dự đoán và trả về kết quả

    prediction = model.predict(email\_tfidf)

    return prediction[0]

# Bước 8: Kiểm tra nhiều email liên tục

while True:

    email = input("Nhập nội dung email cần kiểm tra (hoặc gõ 'exit' để thoát): ")

    if email.lower() == 'exit':

        print("Đã thoát chương trình.")

        break

    result = predict\_email(email)

    print(f"Nội dung email được xác định là: {result}\n")

kết quả: 