

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO ĐỀ TÀI**

**PHÁT HIỆN TIN TỨC GIẢ MẠO**

**Giảng viên hướng dẫn**: **VŨ HOÀI NAM**

**Sinh viên thực hiện:**

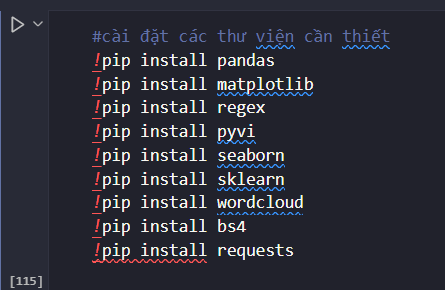
**Nguyễn Văn Khánh – B19DCCN357**

**Nguyễn Văn Dũng – B19DCCN129**

**Đặt vấn đề**

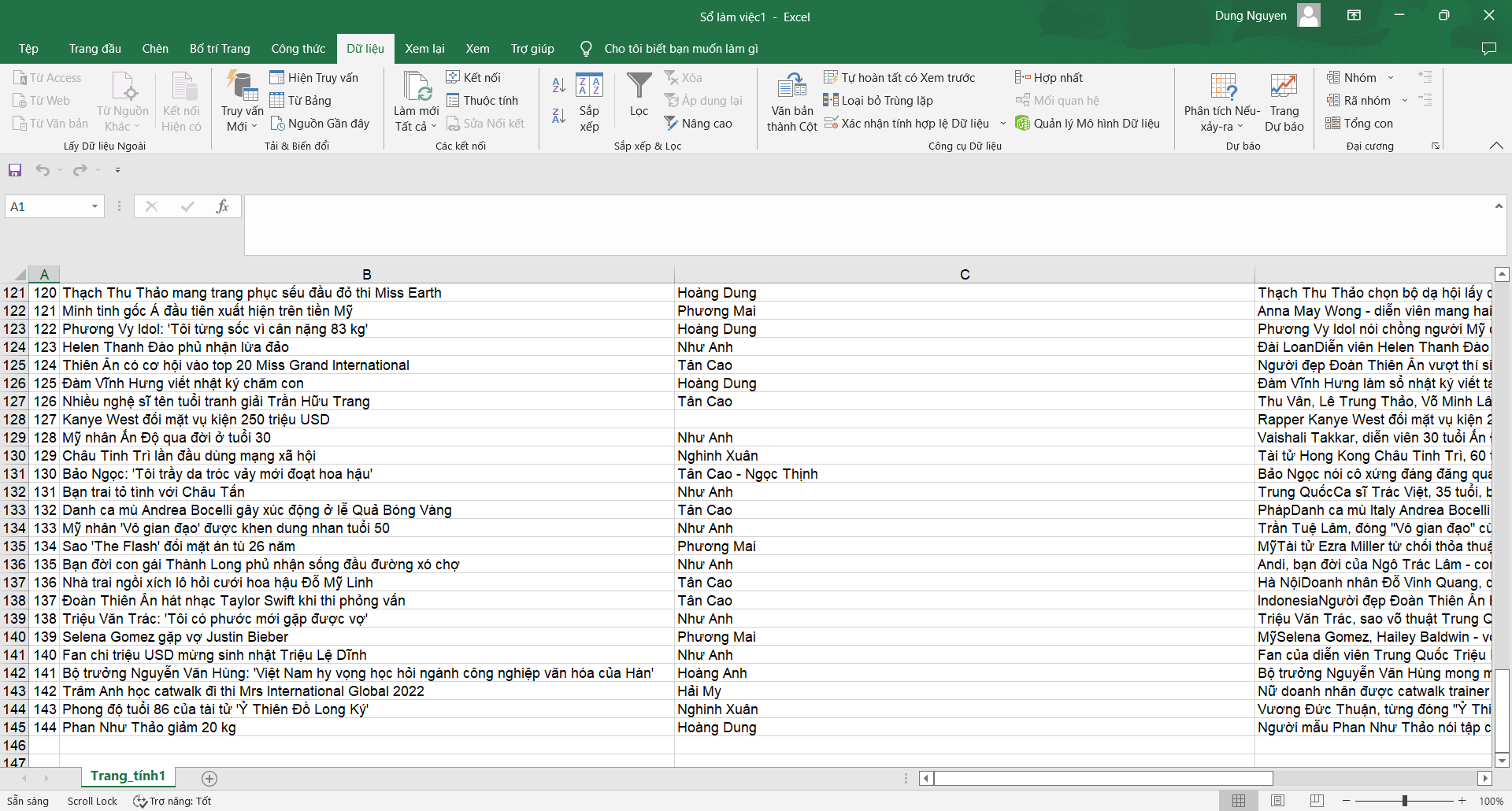
Trong thời đại bùng nổ thông tin, cũng như sự phát triển của mạng xã hội như hiện nay thì việc gặp phải các "Fake News" là điều không thể tránh khỏi. Thậm chí nó đang được nhìn nhận như một trong những mối đe dọa lớn nhất đến sự tiếp cận, nền kinh tế tri thức và tranh luận tự do. Vậy có cách nào để phát hiện fake news hay không? Câu trả lời là có. Bằng cách thực hiện project này chúng ta hoàn toàn có thể phân biệt được sự khác nhau giữa tin thật và tin giả.

1. Cài đặt các thư viện cần thiết

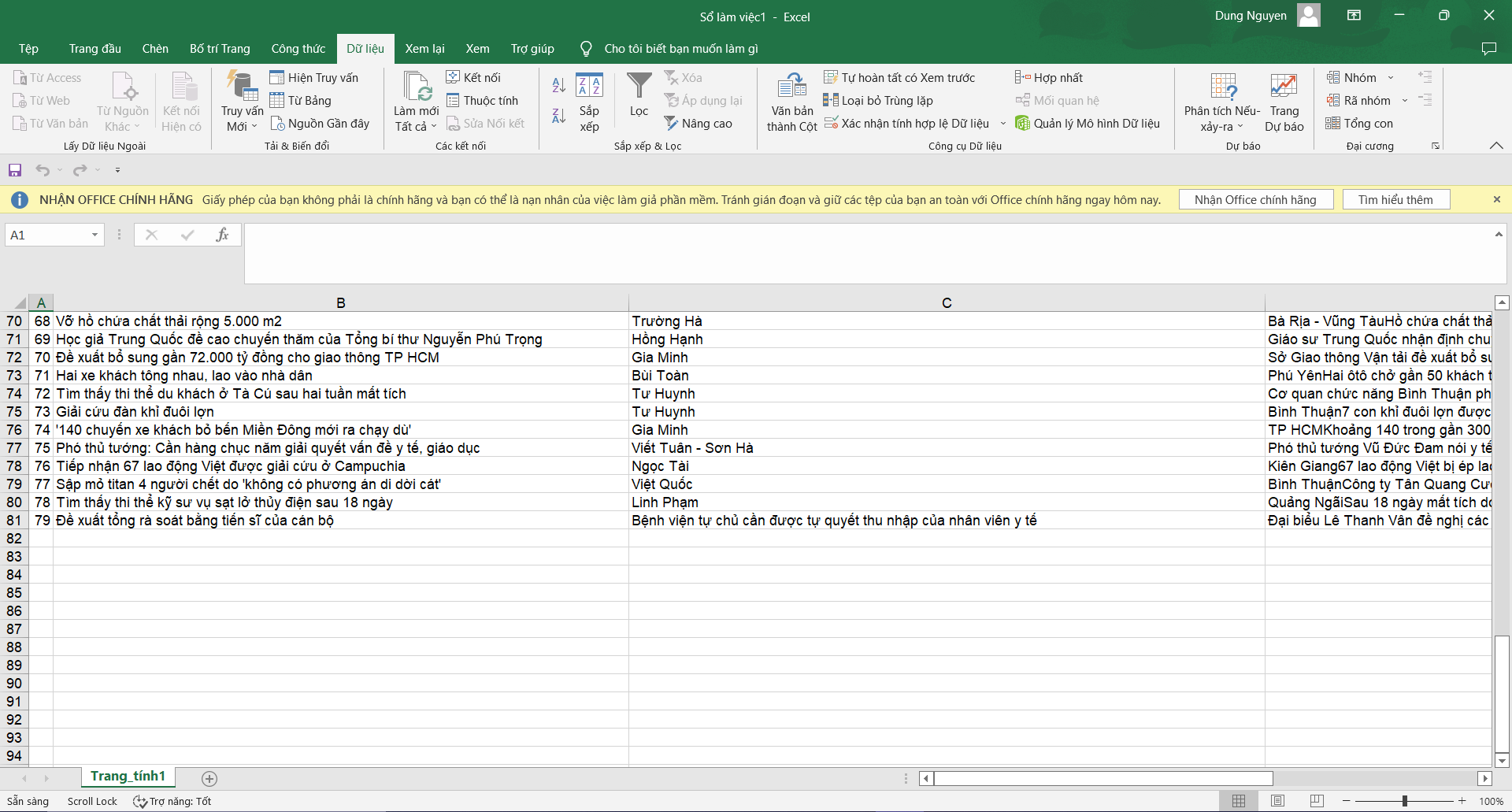


1. Thu thập dữ liệu

* Cào dữ liệu từ vnexpress.net:
* Thể loại giải trí:
* # Thể loại giải trí
* import requests
* from bs4 import BeautifulSoup
* import csv
* data = []
* i = 1
* # Lấy 5 trang
* for index in range(5):
* # Khởi tạo request
* res = requests.get('https://vnexpress.net/giai-tri/gioi-sao-p' + str(index))
* soup = BeautifulSoup(res.content, 'html.parser')
* # Lấy tất cả link trong page hiện tại
* links = soup.select('article.item-news p.description > a')
* # Duyệt mảng links lấy được
* for link in links:
* req = requests.get(link['href'])
* sup = BeautifulSoup(req.content, 'html.parser')
* title = sup.select('h1.title-detail')
* author = sup.select('p.Normal strong')
* description = sup.select('.sidebar-1 p.description')
* # Xử lý missing data
* # Nếu mất trường author thì thêm vào string rỗng
* if(len(author) == 0):
* author.append('')
* # Nếu mất nội dung chính thì bỏ qua
* if(len(description) == 0):
* continue
* # Nếu mất tiêu đề và có nội dung thì lấy câu đầu tiên của nội dung làm tiêu đề
* elif(len(title) == 0):
* title.append(description[0].text.partition('.')[0] + '.')
* try:
* # Thêm vào mảng data
* data.append([i, title[0].text, author[0].text, description[0].text, 1])
* i += 1
* except:
* print(len(author))
* # Xuất ra file csv
* with open('GiaiTri.csv', 'a', encoding='utf8', newline='') as file:
* writer = csv.writer(file)
* writer.writerow(['Stt', 'Tiêu đề', 'Tác giả', 'Nội dung', 'Kết quả'])
* writer.writerows(data)
* file.close()
* Kết quả:



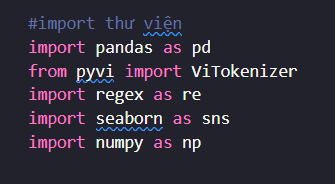
* Thể loại thời sự:
* # Loại thời sự
* import requests
* from bs4 import BeautifulSoup
* import csv
* data = []
* i = 1
* # Lấy 5 trang
* for index in range(5):
* # Khởi tạo request
* res = requests.get('https://vnexpress.net/thoi-su-p' + str(index))
* soup = BeautifulSoup(res.content, 'html.parser')
* # Lấy tất cả link trong page hiện tại
* links = soup.select('article.item-news p.description > a')
* # Duyệt mảng links lấy được
* for link in links:
* req = requests.get(link['href'])
* sup = BeautifulSoup(req.content, 'html.parser')
* title = sup.select('h1.title-detail')
* author = sup.select('p.Normal strong')
* description = sup.select('.sidebar-1 p.description')
* # Xử lý missing data
* # Nếu mất trường author thì thêm vào string rỗng
* if(len(author) == 0):
* author.append('')
* # Nếu mất nội dung chính thì bỏ qua
* if(len(description) == 0):
* continue
* # Nếu mất tiêu đề và có nội dung thì lấy câu đầu tiên của nội dung làm tiêu đề
* elif(len(title) == 0):
* title.append(description[0].text.partition('.')[0] + '.')
* try:
* # Thêm vào mảng data
* data.append([i, title[0].text, author[0].text, description[0].text, 1])
* i += 1
* except:
* print(len(author))
* # Xuất ra file csv
* with open('ThoiSu.csv', 'a', encoding='utf8', newline='') as file:
* writer = csv.writer(file)
* writer.writerow(['Stt', 'Tiêu đề', 'Tác giả', 'Nội dung', 'Kết quả'])
* writer.writerows(data)
* file.close()
* Kết quả:



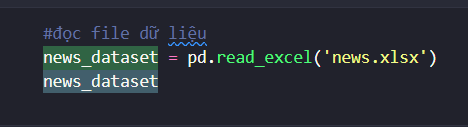
* Do một số tin tức chưa xác nhận được là tin giả hay thật nên được thu thập một cách thủ công trên các page trên Facebook như Theanh28, Việt Tân, ...
* Sau khi thu thập thông tin đồng nhất các dữ liệu về file ‘new.xlsx’.

1. Tiền xử lý dữ liệu

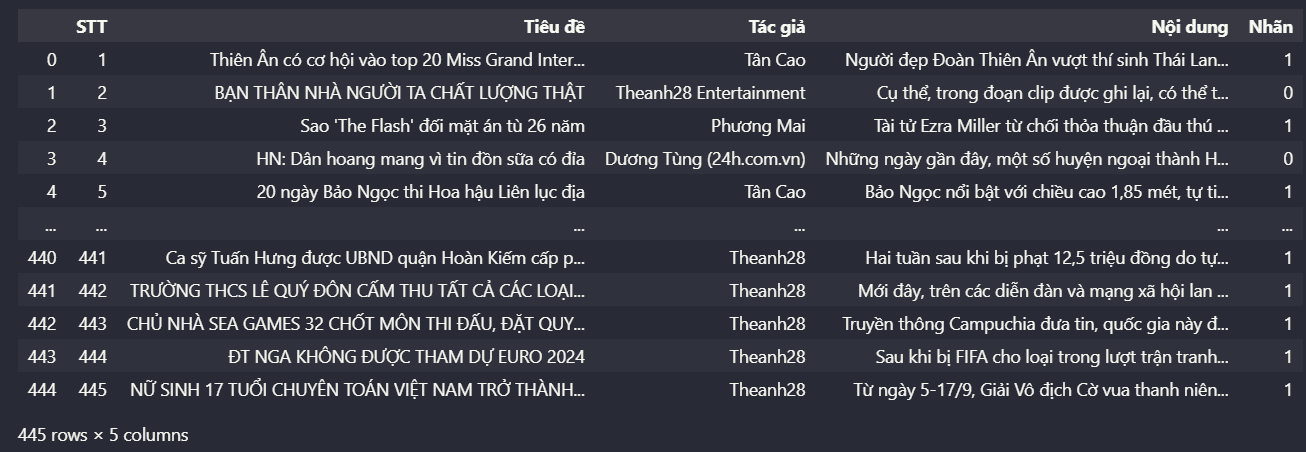
* Import các thư viện cần thiết cho tiền xử lý dữ liệu



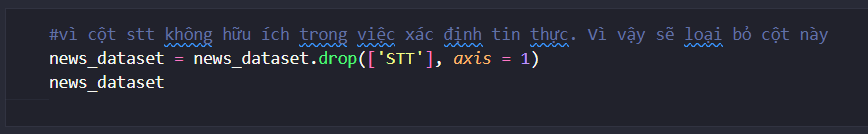
* Đọc dữ liệu từ file ‘new.xlsx’



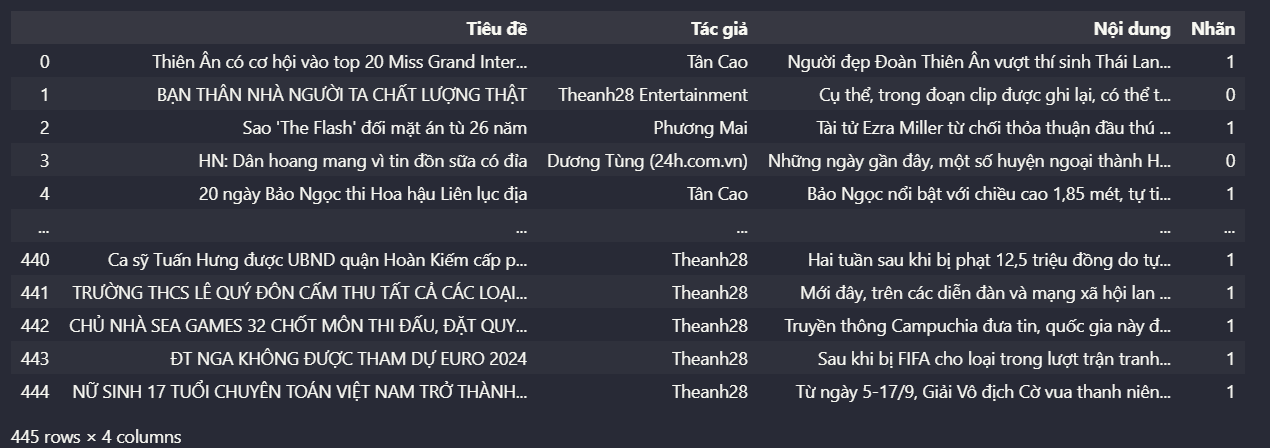
File dữ liệu gồm có 445 dòng và 5 cột:



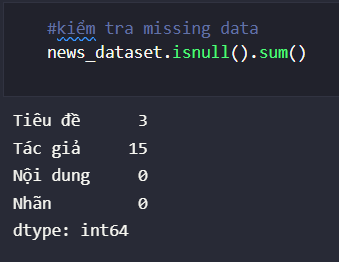
* Loại bỏ cột dữ liệu không hữu ích cho việc phân tích



Dữ liệu lúc này còn 445 dòng và 4 cột:



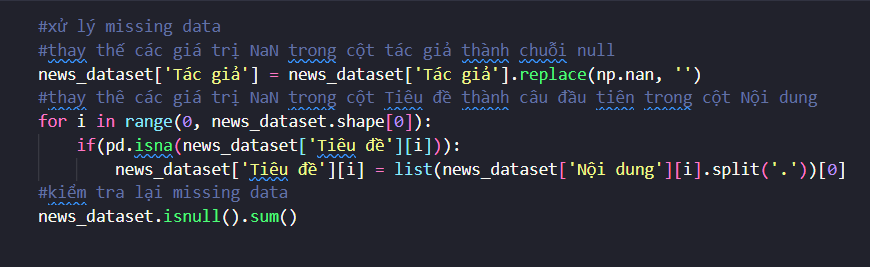
* Kiểm tra missing data



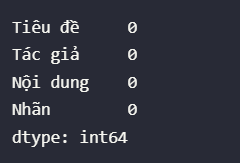
Có 3 ô trong cột ‘Tiêu đề’ và 15 ô trong cột ‘Tác giả’ bị thiếu dữ liệu.

* Xử lý missing data

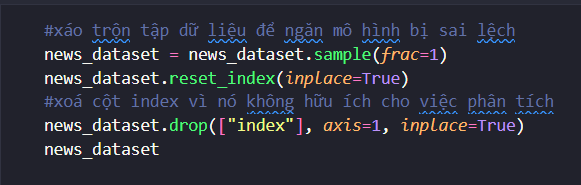
Ở đây các ô bị thiếu dữ liệu trong cột ‘Tác giả’ sẽ được thay thế bằng một chuỗi null và những ô bị thiếu dữ liệu trong cột ‘Tiêu đề’ sẽ được thay bằng câu đầu tiên trong dữ liệu cột ‘Nội dung’. Sau khi xử lý ta kiểm tra lại



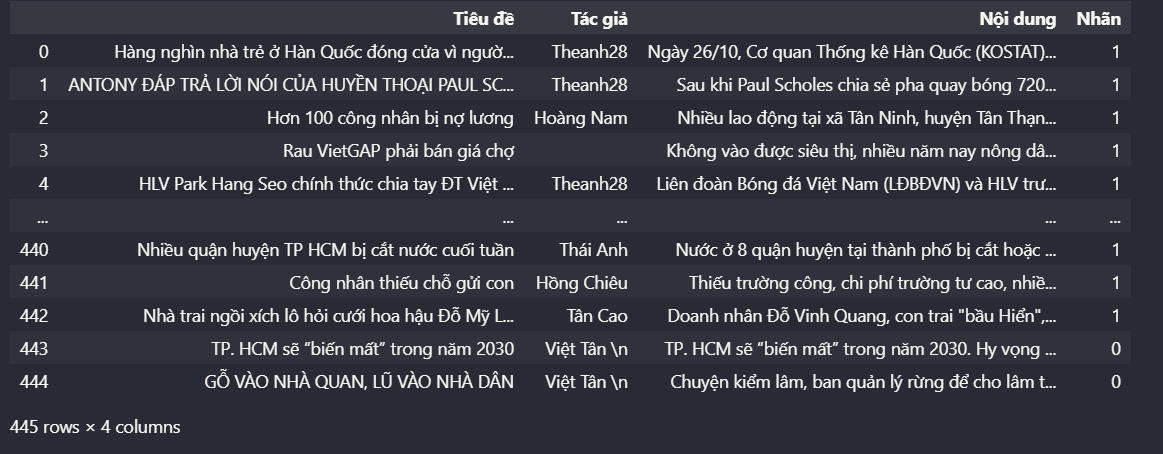
Tất cả dữ liệu bị thiếu đã được xử lý



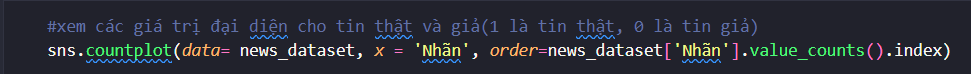
* Xáo trộn tập dữ liệu

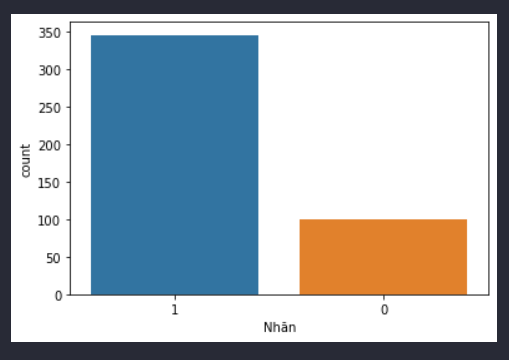


Tập dữ liệu đã được xáo trộn



* Xem các giá trị

Biểu đồ thể hiện các giá trị biểu thị cho tin giả và tin thật



* Xử lý dữ liệu tiếng việt

Ta sẽ chuẩn hoá tập dữ liệu về một kiểu Unicode dựng sẵn để chính xác trong việc so sánh các chuỗi cùng giá trị

Sau đó ta sẽ thực hiện tách các từ để phân loại được các từ đơn và từ ghép

Tiếp đến ta sẽ chuyển tất cả dữ liệu về dạng chữ viết thường

Cuối cùng ta sẽ loại bỏ hết các ký tự đặc biệt và khoảng trắng thừa

#xử lý dữ liệu Tiếng Việt

#chuẩn hoá Unicode Tiếng Việt (đưa về 1 chuẩn Unicode dựng sẵn)

uniChars = "àáảãạâầấẩẫậăằắẳẵặèéẻẽẹêềếểễệđìíỉĩịòóỏõọôồốổỗộơờớởỡợùúủũụưừứửữựỳýỷỹỵÀÁẢÃẠÂẦẤẨẪẬĂẰẮẲẴẶÈÉẺẼẸÊỀẾỂỄỆĐÌÍỈĨỊÒÓỎÕỌÔỒỐỔỖỘƠỜỚỞỠỢÙÚỦŨỤƯỪỨỬỮỰỲÝỶỸỴÂĂĐÔƠƯ"

unsignChars = "aaaaaaaaaaaaaaaaaeeeeeeeeeeediiiiiooooooooooooooooouuuuuuuuuuuyyyyyAAAAAAAAAAAAAAAAAEEEEEEEEEEEDIIIOOOOOOOOOOOOOOOOOOOUUUUUUUUUUUYYYYYAADOOU"

def loaddicchar():

    dic = {}

    char1252 = 'à|á|ả|ã|ạ|ầ|ấ|ẩ|ẫ|ậ|ằ|ắ|ẳ|ẵ|ặ|è|é|ẻ|ẽ|ẹ|ề|ế|ể|ễ|ệ|ì|í|ỉ|ĩ|ị|ò|ó|ỏ|õ|ọ|ồ|ố|ổ|ỗ|ộ|ờ|ớ|ở|ỡ|ợ|ù|ú|ủ|ũ|ụ|ừ|ứ|ử|ữ|ự|ỳ|ý|ỷ|ỹ|ỵ|À|Á|Ả|Ã|Ạ|Ầ|Ấ|Ẩ|Ẫ|Ậ|Ằ|Ắ|Ẳ|Ẵ|Ặ|È|É|Ẻ|Ẽ|Ẹ|Ề|Ế|Ể|Ễ|Ệ|Ì|Í|Ỉ|Ĩ|Ị|Ò|Ó|Ỏ|Õ|Ọ|Ồ|Ố|Ổ|Ỗ|Ộ|Ờ|Ớ|Ở|Ỡ|Ợ|Ù|Ú|Ủ|Ũ|Ụ|Ừ|Ứ|Ử|Ữ|Ự|Ỳ|Ý|Ỷ|Ỹ|Ỵ'.split(

        '|')

    charutf8 = "à|á|ả|ã|ạ|ầ|ấ|ẩ|ẫ|ậ|ằ|ắ|ẳ|ẵ|ặ|è|é|ẻ|ẽ|ẹ|ề|ế|ể|ễ|ệ|ì|í|ỉ|ĩ|ị|ò|ó|ỏ|õ|ọ|ồ|ố|ổ|ỗ|ộ|ờ|ớ|ở|ỡ|ợ|ù|ú|ủ|ũ|ụ|ừ|ứ|ử|ữ|ự|ỳ|ý|ỷ|ỹ|ỵ|À|Á|Ả|Ã|Ạ|Ầ|Ấ|Ẩ|Ẫ|Ậ|Ằ|Ắ|Ẳ|Ẵ|Ặ|È|É|Ẻ|Ẽ|Ẹ|Ề|Ế|Ể|Ễ|Ệ|Ì|Í|Ỉ|Ĩ|Ị|Ò|Ó|Ỏ|Õ|Ọ|Ồ|Ố|Ổ|Ỗ|Ộ|Ờ|Ớ|Ở|Ỡ|Ợ|Ù|Ú|Ủ|Ũ|Ụ|Ừ|Ứ|Ử|Ữ|Ự|Ỳ|Ý|Ỷ|Ỹ|Ỵ".split(

        '|')

    for i in range(len(char1252)):

        dic[char1252[i]] = charutf8[i]

    return dic

dicchar = loaddicchar()

# Đưa toàn bộ dữ liệu qua hàm này để chuẩn hóa lại

def convert\_unicode(*txt*):

    return re.sub(

        r'à|á|ả|ã|ạ|ầ|ấ|ẩ|ẫ|ậ|ằ|ắ|ẳ|ẵ|ặ|è|é|ẻ|ẽ|ẹ|ề|ế|ể|ễ|ệ|ì|í|ỉ|ĩ|ị|ò|ó|ỏ|õ|ọ|ồ|ố|ổ|ỗ|ộ|ờ|ớ|ở|ỡ|ợ|ù|ú|ủ|ũ|ụ|ừ|ứ|ử|ữ|ự|ỳ|ý|ỷ|ỹ|ỵ|À|Á|Ả|Ã|Ạ|Ầ|Ấ|Ẩ|Ẫ|Ậ|Ằ|Ắ|Ẳ|Ẵ|Ặ|È|É|Ẻ|Ẽ|Ẹ|Ề|Ế|Ể|Ễ|Ệ|Ì|Í|Ỉ|Ĩ|Ị|Ò|Ó|Ỏ|Õ|Ọ|Ồ|Ố|Ổ|Ỗ|Ộ|Ờ|Ớ|Ở|Ỡ|Ợ|Ù|Ú|Ủ|Ũ|Ụ|Ừ|Ứ|Ử|Ữ|Ự|Ỳ|Ý|Ỷ|Ỹ|Ỵ',

        lambda *x*: dicchar[*x*.group()], *txt*)

def text\_preprocess(*document*):

    # chuẩn hóa unicode

*document* = convert\_unicode(*document*)

# tách từ(từ đơn và từ ghép)

*document* = ViTokenizer.tokenize(*document*)

# đưa về chữ viết thường

*document* = *document*.lower()

    #loại bỏ ký tự đặc biệt

*document* = re.sub(r'[^*\s\w*áàảãạăắằẳẵặâấầẩẫậéèẻẽẹêếềểễệóòỏõọôốồổỗộơớờởỡợíìỉĩịúùủũụưứừửữựýỳỷỹỵđ\_]',' ',*document*)

    # xóa khoảng trắng thừa

*document* = re.sub(r'*\s*+', ' ', *document*).strip()

    return *document*

#Thực hiện xử lý dữ liệu với từng cột trong tập dữ liệu

for i in range(0, len(news\_dataset['Tiêu đề'])):

    news\_dataset['Tiêu đề'][i] = text\_preprocess(news\_dataset['Tiêu đề'][i])

for i in range(0, len(news\_dataset['Tác giả'])):

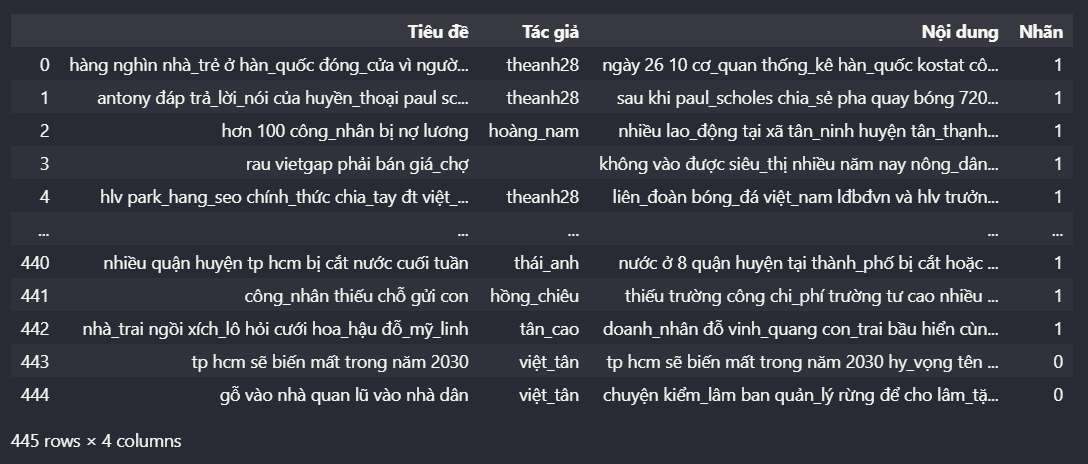
    news\_dataset['Tác giả'][i] = text\_preprocess(news\_dataset['Tác giả'][i])

for i in range(0, len(news\_dataset['Nội dung'])):

    news\_dataset['Nội dung'][i] = text\_preprocess(news\_dataset['Nội dung'][i])

news\_dataset

Sau xử lý ta thu được tập dữ liệu sau



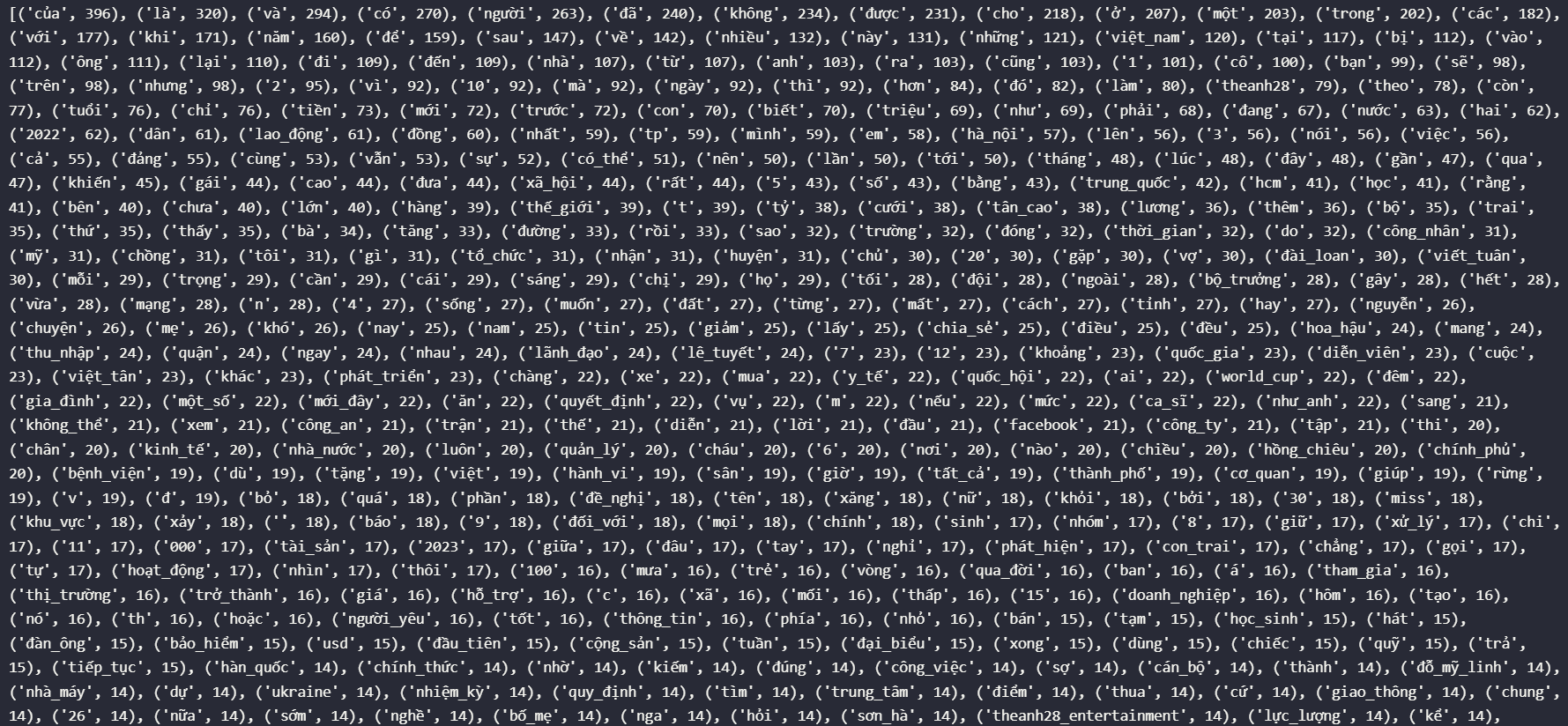
* Tìm các từ dừng

Duyệt qua tập dữ liệu sau đó đếm số lần xuất hiện của các từ trong đó.

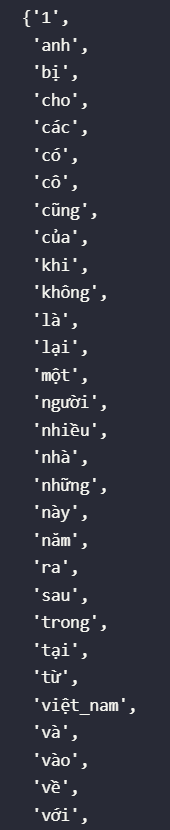
Sau đó sẽ lọc ra tất cả các từ có số lần xuất hiện lớn hơn hoặc bằng 100 là các từ dừng



Từ điển thể hiện số lần xuất hiện của các từ



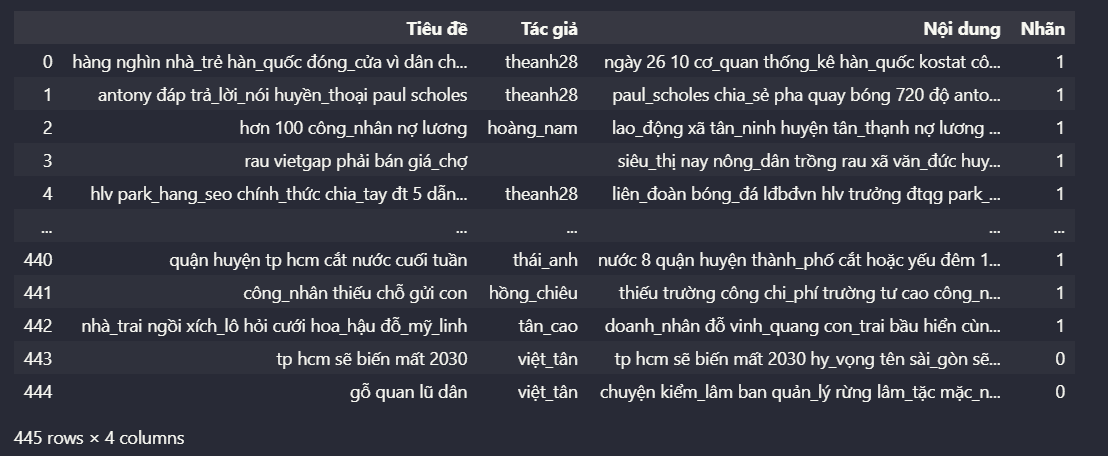
Tập hợp các từ dừng



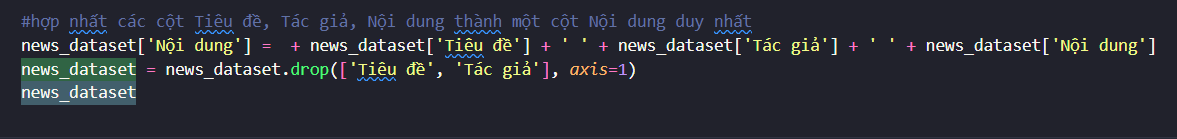
* Loại bỏ các từ dừng



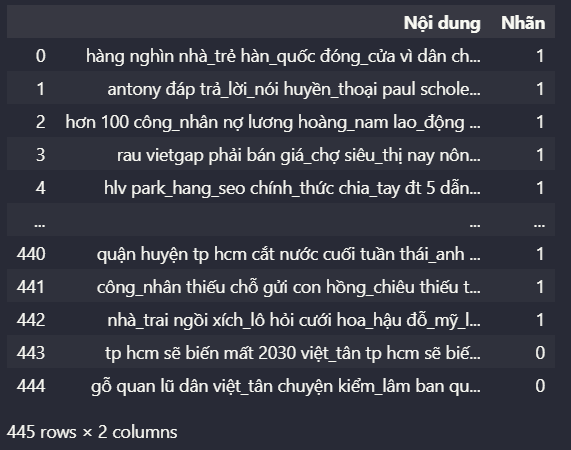
Dữ liệu thu được sau khi loại bỏ các từ dừng



* Hợp nhất các cột ‘Tiêu đề’, ‘Tác giả’ và ‘Nội dung’ với nhau

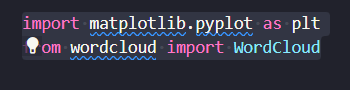


Dữ liệu thu được sau khi hợp nhất



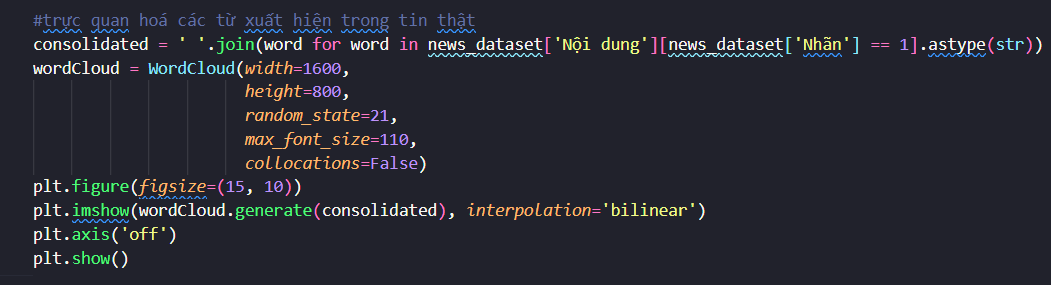
1. Phân tích dữ liệu

Import các thư viện cần thiết



Trực quan hoá các từ xuất hiện trong tin giả và tin thật

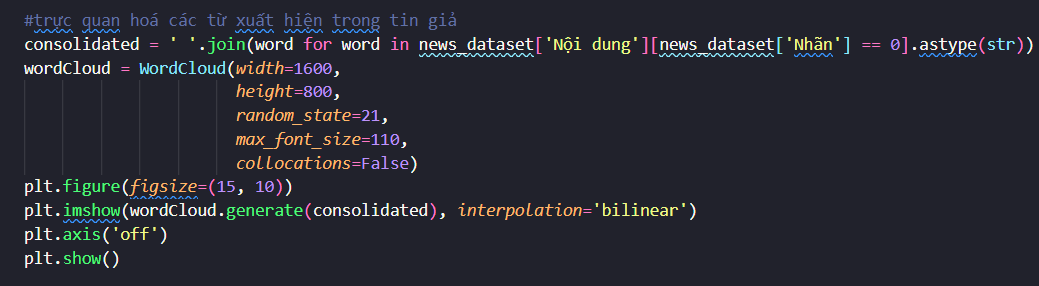
+ Tin thật



Tần suất xuất hiện của các từ trong tin thật

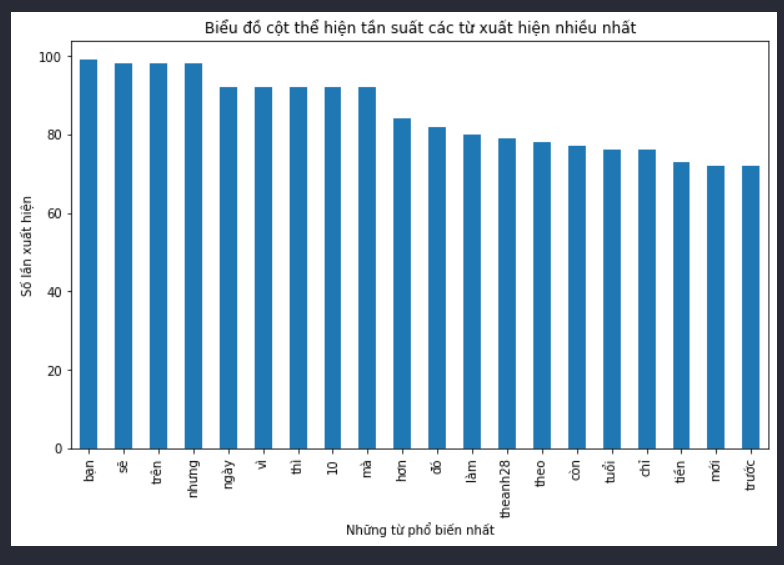


+ Tin giả



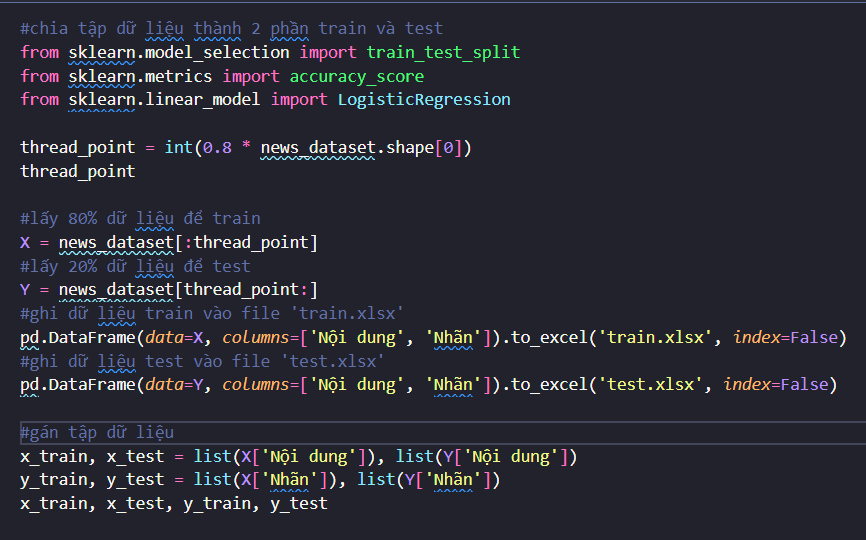
Tần suất xuất hiện của các từ trong tin giả

- Vẽ biểu đồ cột thể hiện 20 từ có tần suất xuất hiện lớn nhất

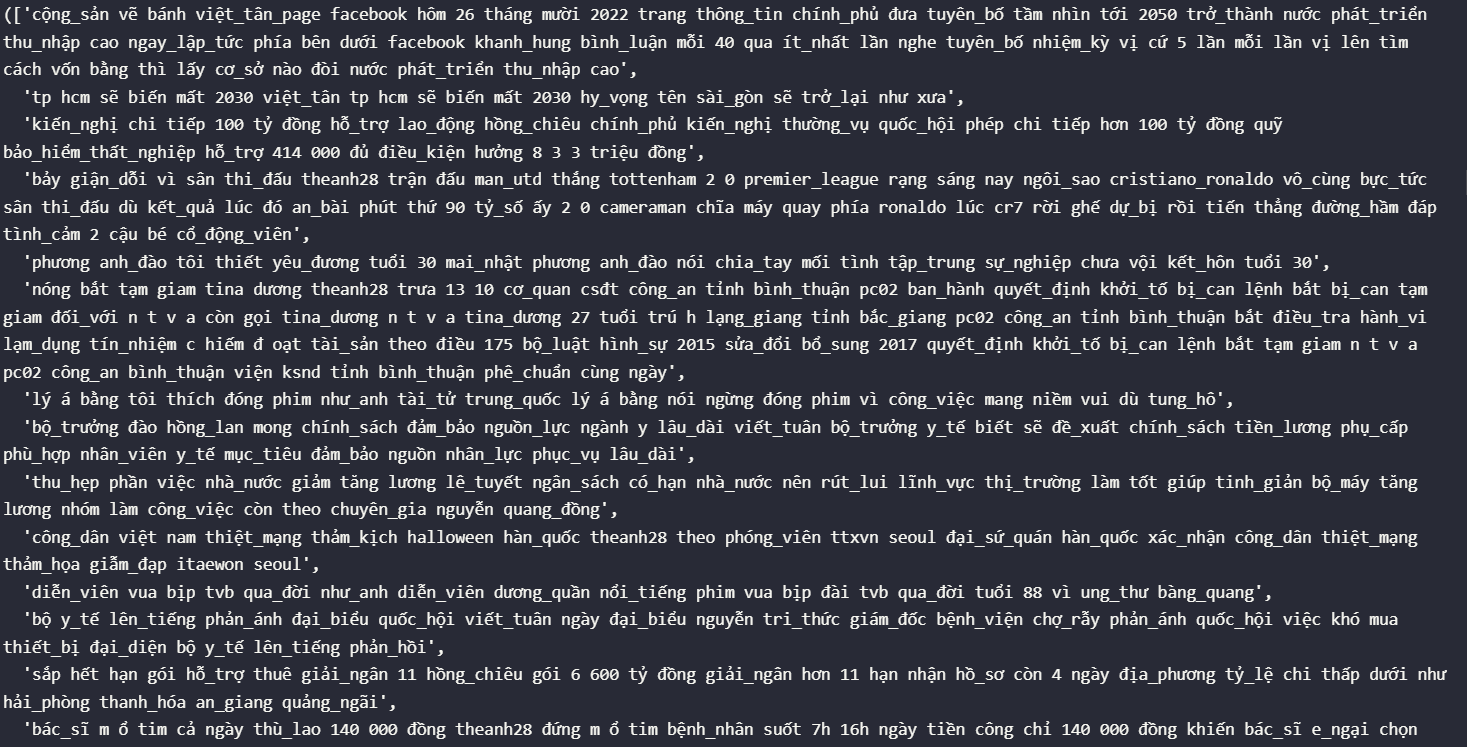


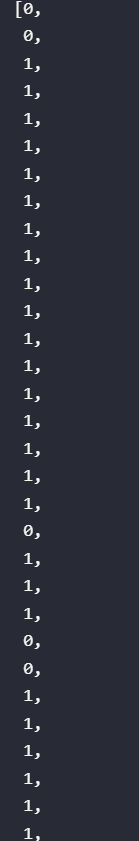
1. Chuyển đổi văn bản thành vector

* Chia đôi tập dữ liệu và ghi vào file tương ứng



Kết quả thu được sau khi chia





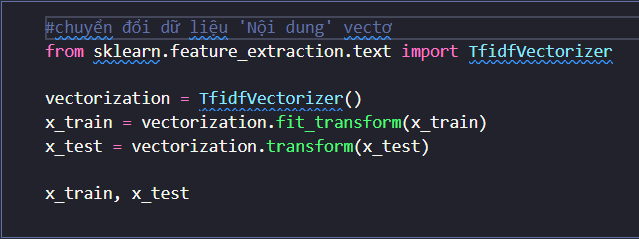
File ‘train.xlsx’



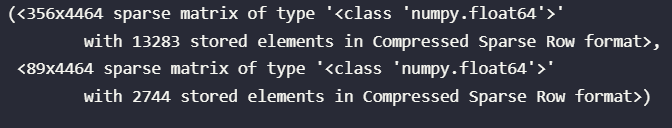
File ‘test.xlsx’



* Chuyển đổi dữ liệu ‘Nội dung’ thành vector

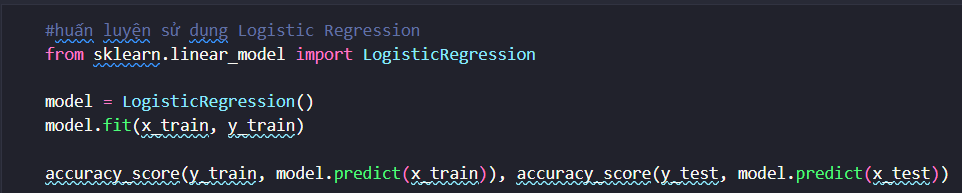


Kết quả thu được



1. Huấn luyện, đánh giá và dự đoán

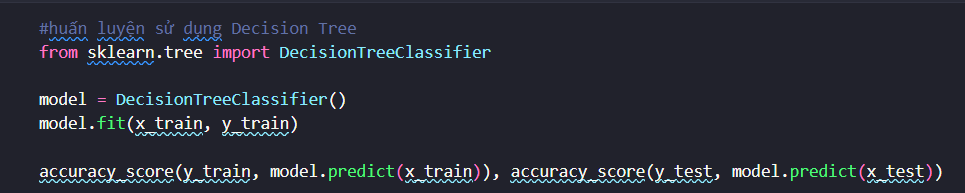
* Huấn luyện sử dụng Logistic Regression



Kết quả sự chính xác thu được



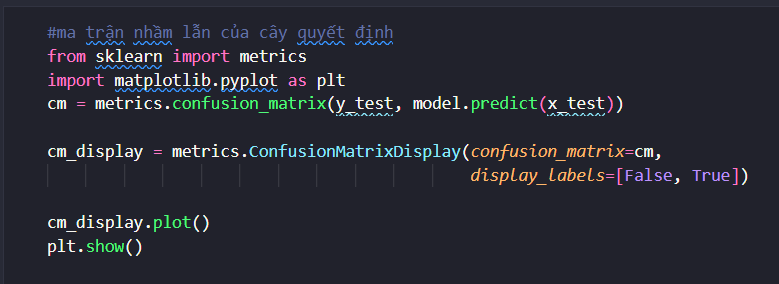
* Huấn luyện sử dụng Decision Tree



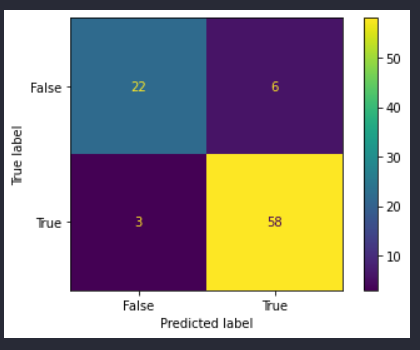
Kết quả sự chính xác thu được



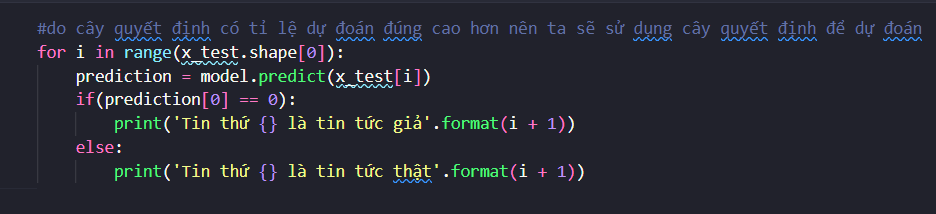
* Vẽ ra ma trận nhầm lẫn của DescisionTree



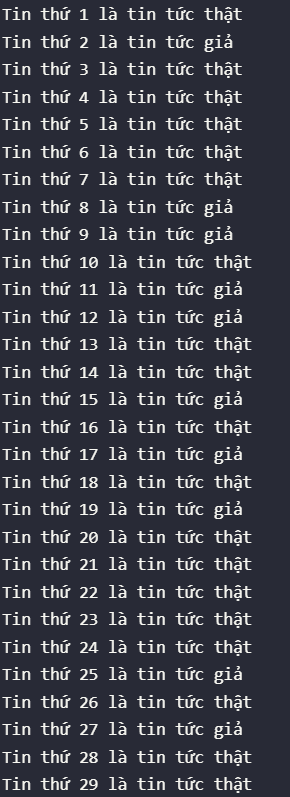
Ma trận nhầm lẫn của Decision Tree



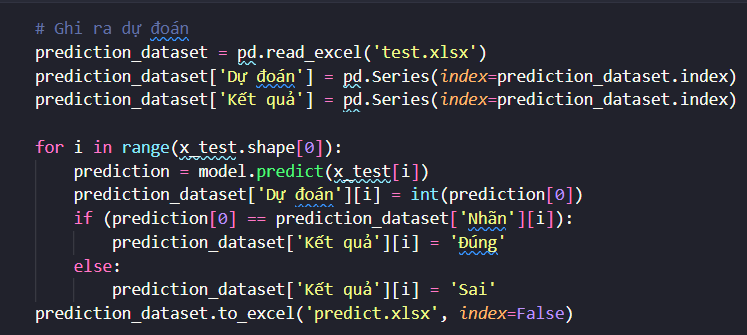
* Do Decision Tree cho tỉ lệ dự đoán đúng cao hơn nên ta sử dụng Decision Tree để thực hiện dự đoán



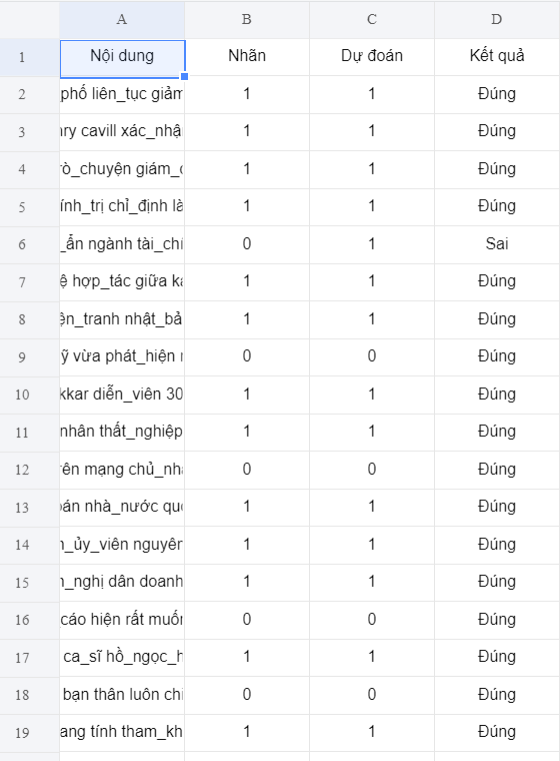
Kết quả dự đoán



* Xuất file dự đoán



File thu được khi dự đoán dữ liệu trong file ‘test.xlsx’



**Lời kết**

Cuối cùng em muốn gửi lời cảm ơn đến thầy đã hướng dẫn chúng em hoàn thành bài báo cáo của mình. Do chưa có nhiều kinh nghiệm về viết báo cáo và còn có nhiều hạn chế về mặt kiến thức nên trong bài báo cáo không tránh khỏi có những sai sót. Em mong muốn nhận được sự đóng góp , phê bình từ phía thầy để bài báo cáo của em được hoàn thiện hơn

Em xin chân thành cảm ơn !