ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

----------o0o----------

BÁO CÁO ĐỒ ÁN MÔN HỌC

MACHINE LEARNING

Đề tài: ***FAKE NEWS DETECTION***

GV: **PHẠM NGUYỄN TRƯỜNG AN**

LÊ ĐÌNH DUY

LỚP: **CS114.L21.KHCL**

SV: **ĐẶNG XUÂN MAI – 19521820**

NGUYỄN THỊ THẢO HIỀN – 19521488

LỚP: **CS114.L22.KHCL**

SV: **NGUYỄN HOÀI NAM – 18521126**

NGUYỄN THỊ CẨM HƯỚNG - 19521594

Tóm tắt lại bài toán

Tin giả còn được gọi là tin rác hoặc tin tức giả mạo, là các thông tin giả được lan truyền qua phương tiện truyền thông truyền thống (như in và phát sóng, báo chí) hoặc phương tiện truyền thông xã hội trực tuyến...[1]

Nguyên nhân mà người ta hay tạo ra các tin giả vì [1]:

Trình độ nhận thức hạn chế, tung tin cho vui, không nghĩ đến hậu quả.

Thích "câu like", tạo sự chú ý để trở thành người quan trọng trong mắt người khác.

Tung tin giả có chủ đích, nhằm gây mất trật tự, an ninh xã hội, kích động người dân.

Tin thật, chính thống thường dùng trong hoạt động truyền thông, báo chí - để chỉ một sự kiện, sự việc nào đó có tính mới và là nguồn tin được đưa trên sự thật đã xảy ra, chứ không bịa đặt,... [2]

Dùng những kiến thức đã tìm hiểu được xây dựng 1 chương trình có thể dự đoán đoạn văn bản đưa vào để kiểm tra là tin thật hay tin giả.

Input: 1 đoạn văn bản có độ dài bất kỳ

Output: 1 – nếu chương trình dự đoán thấy đó là tin giả

0 – nếu chương trình dự đoán thấy đó là tin thật

Thống kê lại dữ liệu

Bộ dữ liệu tổng cộng có 485556 dữ liệu, mỗi dữ liệu sẽ có 4 thông tin bao gồm ngày bài báo được đăng tải, tiêu đề của bài báo, 3 dòng đầu tiên trong đoạn văn của bài báo và nhãn bài báo (nhãn = '1' bài báo là False, nhãn = '0' bài báo là True).

Trong đó có tổng cộng 231618 dữ liệu là bài báo giả, được thu thập từ 22 trang tin giả, và có 253938 là bài báo dược thu thập từ những trang báo chính thống, tổng cộng là 10 trang báo.

Sau đó dữ liệu được chia thành 80% train (tương đương với 388444 bài báo) – 20% test (tương đương với 97112 bài báo).

Dữ liệu được thu thập có tất cả 4 thông tin là date, title, text và is_fake. Khi chuyển về file train.json và test.json thì đã được loại bỏ bớt 2 cột, nên thông tin của mỗi bài báo trong file train và test chỉ còn có phần text và is_fake.

Trong đó, file train.json và test_clear.json được đính kèm đã được tiền xử lý dữ liệu như loại bỏ stopwords, loại bỏ những ký tự không phải là chữ,… Còn file test.json thì có nội dung giống như trong bản gốc lấy từ các trang báo trên mạng, có nghĩ là sẽ có các dấu $, %, !, các ký tự in hoa,…

Như vậy thì khi chạy thử model thầy sẽ quan sát được rõ hơn sự khác biệt giữa có tiền xử lý và không tiền xử lý dữ liệu để test.

Những cập nhật thay đổi so với lần báo cáo vấn đáp

Các thay đổi lần lượt là:

1. Stemming data bị lược bớt

Lúc báo cáo vấn đáp với thầy, nhóm có nói là trong các bước tiền xử lý dữ liệu thì bao gồm cả stemming dữ liệu.

Ví dụ minh hoạ về trước và sau khi stemming [3]:

Trước khi stemming	Sau khi steamming
program	program
programs	
programmer	
programmers	
programming	

Nhưng sau khi xem xét kỹ càng lại đoạn code thì hàm steamming của nhóm đang không làm gì cả, không hề đưa dữ liệu về đúng dạng như định nghĩa, nên nhóm xin được trình bày lại là không có bước stemming dữ liệu ở trong đồ án, chỉ có các bước: loại bỏ stopwords (in, at, the, a…), loại bỏ punctuation (các dấu chấm câu) và chuyển tất cả các ký tự về chữ thường, không còn ký tự in hoa nào nữa.

2. Bổ sung thêm Tf – idf và n_grams để nâng cao accuracy

Sau khi báo cáo vấn đáp, nhóm có thêm phần Tf – idf và n-grams (do anh Nguyễn Hoài Nam làm thêm) để nâng cao hơn accuracy. Vì đã thay đổi phương pháp vector hoá các đoạn văn bản, điều đó dẫn đến accuracy của các model cũng khác so với lúc báo báo với thầy. Khi dùng Bag of Word thì model Logistic Regression được xem là tốt nhất, tuy nhiên, khi dùng Tf -idf và ngrams thì model dùng LinearSVC của Support Vector Machine đạt được accuracy cao hơn, nên nhóm cũng đổi luôn model tốt nhất để phát triển ứng dụng của nhóm từ Logistic Reggression thành LinearSVC.

3. Thêm phần phát triển ứng dụng

Khi đã có được model có kết quả tốt nhất, nhóm lưu lại thành file .sav và sử dụng thêm thư viện pickle để xây dựng 1 chương trình có giao diện người dùng kết hợp với model đã có để dự đoán được đoạn văn bản người dùng gõ vào (hoặc copy paste vào) là tin thật hay tin giả.

Thứ tự các file

Các file data_fake_news.json, data_real_news.json là tổng hợp của tất cả các trang báo thuộc từng loại (tin giả hoặc tin thật).

File train.json là dữ liệu dùng để train model (đã được cleaning và loại bỏ bớt các cột không dùng đến).

File test.json là dữ liệu test không được cleaning.

File test_clear.json là dữ liệu test đã được cleaning.

File FakeNewsDetection_FinalModel.ipynb là file chứa mã nguồn dùng để train model theo phương pháp Bag of words.

File Fake_news_detection_4.2.ipynb là file chứa mã nguồn dùng để train model theo phương pháp Tf – idf ngrams.

Các file app.py, bgm.mp3, model.sav và tf-idf.sav là các file để phát triển ứng dụng. Cách để chạy thử ứng dụng là gom 4 file trên vào cùng 1 thư mục sau đó chạy trên PyCharm hoặc terminal.

Những đường link liên quan đến đồ án của nhóm

Link colab tổng hợp tất cả các dữ liệu của các trang web:

https://colab.research.google.com/drive/1do_gDLJmpzV279AkrzopmDeS-ukPOYqx

Link colab xây dựng model:

https://colab.research.google.com/drive/1IHuHmpujWmdCPTjuOWSa5pBYXuMRtnfp#scrollTo=Cdr-ewe9pDw5

Link colab phần tf-idf phát triển thêm:

https://colab.research.google.com/drive/1XtwqndMjvba9Vh06fYmrgCpc6d7Mm1SP?usp=sharing

Link github của nhóm, tổng hợp tất cả các file liên quan đến đồ án: https://github.com/dxmai/CS114.L21.KHCL/tree/main/FinalProject

Link demo chương trình ứng dụng:

https://www.youtube.com/watch?v=giCGo96i_vU&t=2s

Tài liệu tham khảo

[1]“Tin giả là gì? Đăng tin giả bị xử phạt như thế nào?,” *Luật Hoàng Phi*, Mar. 03, 2021. https://luathoangphi.vn/tin-gia-la-gi-dang-tin-gia-bi-xu-phat-nhu-the-nao/ (accessed Aug. 18, 2021).

[2]“01. Tin, tin tức là gì? - Viết và phát hành một Thông cáo báo chí.” https://sites.google.com/a/ecolaw.vn/viet-va-phat-hanh-mot-thong-cao-bao-chi/mot-so-khai-niem-co-ban-trong-truyen-thong-bao-chi/tin-tin-tuc-bao-chi-la-gi (accessed Aug. 18, 2021).

[3]“Python | Stemming words with NLTK,” *GeeksforGeeks*, Oct. 30, 2018. https://www.geeksforgeeks.org/python-stemming-words-with-nltk/ (accessed Aug. 18, 2021).