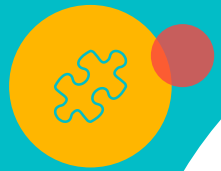
ĐỒ ÁN CUỐI KỲ
NHẬP MÔN
KHOA HỌC DỮ LIỆU

Nội dung chính

1. Ý tưởng
2. Thu thập, khám phá dữ liệu
3. Tiền xử lý dữ liệu và xây dựng mô hình
4. Tổng kết

Ý tưởng



Liệu có thể phân loại được các bài báo theo từng chủ đề hay không?

Nếu xây dựng thành công mô hình này ta sẽ giúp cho những trang báo điện tử có thể tự động hóa trong việc phân chia các bài báo một cách nhanh nhất

Thu thập
Khám phá dữ liệu

Thu thập dữ liệu

Trên mạng xã hội hiện nay có rất nhiều các trang báo điện tử.

→ Việc tìm trang báo điện tử uy tín để lấy dữ liệu là rất quan trọng.

VnExpress

Thời sự

Thế giới

Thể thao

Pháp luật

Xe bọc thép cùng 6.000 người xuất quân bảo vệ Đại hội Đản

Kinh doanh

HÀ NỘ

người cùng xe bọc thép, xe chuyên dụng đã tham gia lễ xuất quân bảo vệ Đại hội Đảng toàn quốc lần thứ 13 tại

Thời s

Giáo dục

Trục vớt mảnh máy bay Indonesia từ độ sâu 23 mét

Thợ lặn Indonesia vớt các bộ phận máy bay của hãng Sriwijaya Air từ vùng biển sâu khoảng 23 mét sau khi phát hiện tín hiệu nghi là hộp đen.

Tìm thấy mảnh thi thể vụ rơi máy bay Indonesia

Các phần thi thể và mảnh vỡ được trục vớt từ một vùng biển gần thủ đô Jakarta, nơi một máy bay chở 62 người lao xuống hôm 9/1. 27

Nỗi cô đơn của ông Park

Ông Park làm việc với đội tuyển Quốc gia và U22 đêm Giáng sinh, rồi bay về Hàn Quốc khi quê hương trong làn sóng dịch thứ ba.

Việt Tâm

37

190 nghị sĩ soạn điều khoản xem xét bãi nhiệm Trump

jeu soạn

Kinh doanh

Quốc tế | Doanh nghiệp | Chứng khoán | Bất động sản | Startup | Vhome

Giá xăng, dầu ngày mai có

Chiến lược kinh tế với

Thu thập dữ liệu

- Bước 1: Ta sẽ truy cập vào trang web theo từng chủ đề theo mẫu:

  https://vnexpress.net/{subject}-p{page}
- Bước 2: Từ địa chỉ đó ta sẽ thu thập địa chỉ url của các bài báo theo chủ đề đó.
- Bước 3: Truy cập vào danh sách url của từng chủ đề và lấy nội dung của từng bài báo.

Thời sự

Chủ nhật, 10/1/2021, 14:52 (GMT+7)

'Quái xế' náo loạn đường phố Sài Gòn

Hàng trăm thanh niên chạy xe "cọp" chặn nhiều tuyến đường, quốc lộ vùng ven TP HCM so kè tốc độ gây náo loạn cả ngày lẫn đêm.

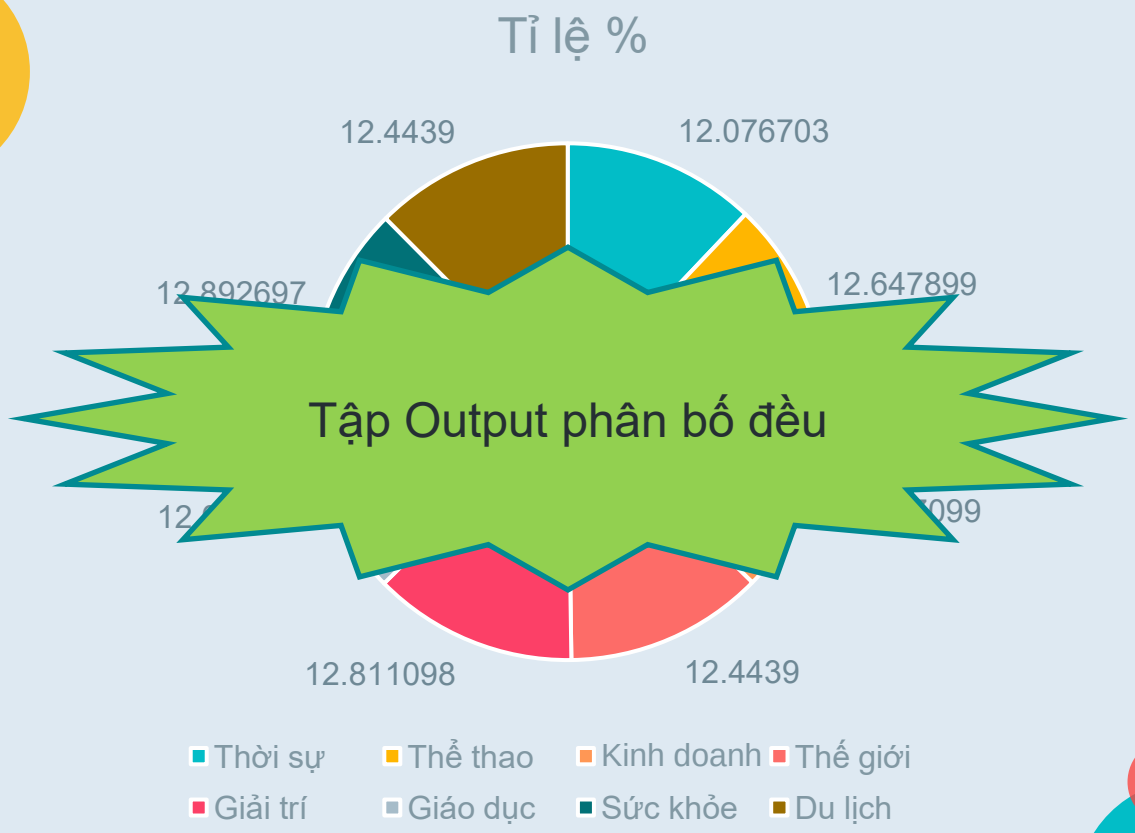
Liên tiếp những ngày đầu năm, nhất là cuối tuần, nhiều thanh niên chạy xe máy đã được được độ, chế liên tục hẹn các "đài" - cuộc tập hợp của hàng trăm, có khi đến hàng nghìn người từ nhiều nơi về một địa điểm để đua xe trái phép.

Khám phá dữ liệu

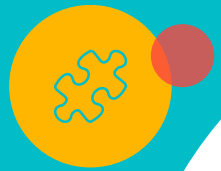
Ta sẽ trả lời một số câu hỏi:

- Dữ liệu thu thập được có bao nhiêu dòng, cột?
- Dữ liệu có bị lặp hay không?
- Kiểu dữ liệu của các cột?
- Có dòng nào không thu thập được nội dung văn bản không?
- Cột output có bị thiếu giá trị không?
- Tỉ lệ các giá trị trong cột Output như thế nào?

Khám phá dữ liệu



Tiền xử lý
Xây dựng mô hình



Tiền xử lý dữ liệu

- Xóa các ký tự đặc biệt trong văn bản.
- Tách từ trong văn bản

Đây ta sử dụng thư viện ***pyvi*** để tách một từ trong văn bản

- Xóa những dòng có số ký tự ***ít hơn 1000***

Loại bỏ đi những dòng không lấy được dữ liệu hoặc lấy được quá ít

- Chuyển cột Output thành dạng số.

Xây dựng mô hình

- Chúng ta sẽ đưa dữ liệu dạng văn bản đã được xử lý về dạng vector thuộc tính có dạng số học.
- Có nhiều phương pháp:
  - *Count Vectors as features*
  - *TF-IDF Vectors as features*
  - *Word Embeddings as features*

Phương pháp TF-IDF

- Đây là một phương pháp cực kì phổ biến trong xử lý văn bản. Nó được tính theo công thức:
- $$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$
- $$IDF(t) = \log e\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right)$$
- Sử dụng N-gram: Kết hợp n thành phần (từ) liên tiếp nhau.

VD: “Thủ_tướng Đức nhận_lời tham_dự lễ kỷ_niệm"

Các thuật toán áp dụng cho phân loại văn bản

Naïve Bayes

Áp dụng định lý Bayes với giả định độc lập để phân loại văn bản

Support Vector Machine

Mô hình học có giám sát, thể hiện dữ liệu là các điểm trong không gian.

Logistic Regression

Mô hình hồi quy Logistic dùng để dự đoán đầu ra liên tục

K – nearest neighbor

Tính toán khoảng cash giữa các điểm dữ liệu để gom nhóm và xác định thể loại văn bản

Neural Network

Sử dụng mạng nơron nhân tạo để phân lớp văn bản

Decision Tree

Từ dữ liệu ban đầu, cây quyết định sinh ra các luật để dự đoán lớp của các dữ liệu

Phương pháp Naïve Bayes

- Sử dụng mô hình Multinomial Naïve Bayes.
- Công thức tính Naïve Bayes tổng quát:

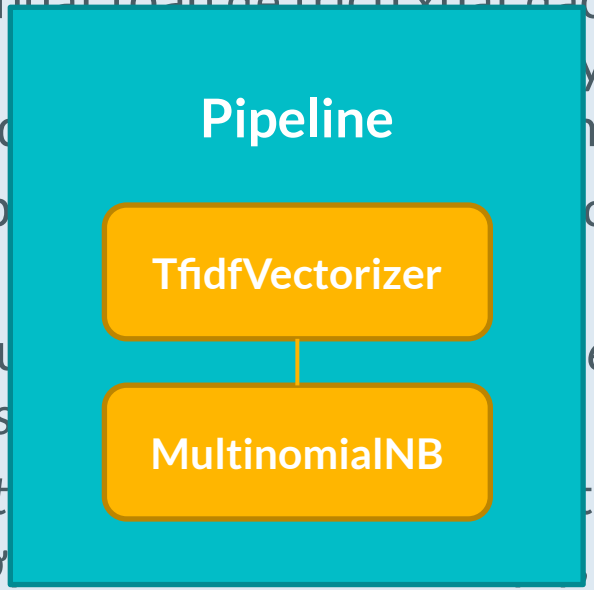
$$\Pr(c|ti) = \frac{\Pr(c)\Pr(ti|c)}{\Pr(ti)}$$

- Để tìm hiểu về công thức phân lớp tham khảo thêm ở:

https://www.cs.waikato.ac.nz/ml/publications/2004/kibriya_et_al_cr.pdf

Pipeline

- Một *pipeline* trong *sklearn* là một tập các chuỗi thuật toán để trích xuất đặc trưng, tiền xử … yện dữ liệu sử … máy cụ thể.
- Mỗi pip… ức nhất định.
- Bước cu… e được gọi là *es*…
- Một *est*… t toán phân lớ… một mạng nơ-ron hay có thể là một thuật toán học máy không giám sát.



Test các siêu tham số

Có 3 siêu tham số mà ta sẽ thử:

- Tham số ***α*** trong **MultinomialNB**.

Khi tính xác suất có trường hợp p = 0. Do đó tử số sẽ cộng cho α để tránh trường hợp đó.

- Tham số ***min_df*** trong **TfidfVectorizer**

Được sử dụng để xóa các từ xuất hiện không quá thường xuyên trong các văn bản.

- Tham số ***max_df*** trong **TfidfVectorizer**

Được sử dụng để xóa các từ xuất hiện quá thường xuyên trong các văn bản.

Test mô hình vừa huấn luyện

- Để cho khách quan, dữ liệu test sẽ được thu thập ở một trang báo điện tử khác.
- Ta chọn trang https://vietnamnet.vn/
- Ta sẽ dựa vào các chủ đề để lấy các bài báo sau đó ghi ra một file *test.csv*.
- Ta sẽ đọc lấy dữ liệu test đó để predict thử và kiểm tra kết quả.

Độ lỗi mô hình

- Độ chính xác trên tập train là: 99,58%
- Độ chính xác trên tập validation là: 92.91%
- Độ chính xác trên tập test là: 81.45%

=> Nhận xét:

- Mô hình dự đoán trên tập test và validation cho kết quả dự đoán khá cao.
- **Nhưng** khi mang mô hình ra để dự đoán các bài báo thu thập ở một trang khác (https://vietnamnet.vn/) thì cho kết quả thấp hơn nhưng vẫn ở mức cao.

Thư viện sử dụng trong đồ án

- Matplotlib
- Seaborn
- Requests
- Numpy
- BeautifulSoup
- Sklearn
- Pyvi

Tài liệu tham khảo

- https://viblo.asia/p/phan-loai-van-ban-tu-dong-bang-machine-learning-nhu-the-nao-4P856Pa1ZY3
- https://viblo.asia/p/phan-loai-van-ban-tu-dong-bang-machine-learning-nhu-the-nao-phan-2-4P856PqBZY3
- https://scikit-learn.org/stable/


Tổng kết

Quá trình làm đồ án

- Điều học được qua đồ án này:
- ✓ Tự đặt và tìm ra các phương pháp để có thể giải quyết vấn đề một cách tốt, hiệu quả nhất.
- ✓ Cách thu thập dữ liệu từ trang web.
- ✓ Lựa chọn và huấn luyện một mô hình.
- ✓ Cách làm việc nhóm hiệu quả.

Quá trình làm đồ án

- Khó khăn
  - Trong việc tìm đề tài và chọn câu hỏi.
  - Trong việc tìm kiếm nguồn dữ liệu .
  - Trong việc chọn mô hình, siêu tham số phù hợp, phương pháp xử lý nào là tốt nhất, thuận tiện nhất.

Nếu có thêm thời gian em sẽ làm gì?

- Kết quả dự đoán của mô hình vẫn chưa thực sự được cao như em kỳ vọng.

Vì vậy nếu có thêm thời gian em sẽ thử xây dựng các mô hình khác để xem kết quả có thể tăng lên hay không?

Cảm ơn thầy đã lắng nghe!

Thành viên nhóm

- Nguyễn Huy Hải – MSSV: 18120023.
- Nguyễn Thanh Tùng – MSSV: 18120104