**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**VIỆN KỸ THUẬT CÔNG NGHỆ**

****

**ĐỒ ÁN CUỐI KÌ**

**ĐỒ ÁN CHUYÊN NGÀNH**

**Đề tài:**

**XÂY DỰNG ỨNG DỤNG TRÍCH XUẤT THÔNG TIN BÀI BÁO BẰNG THUẬT TOÁN SVM**

*Người hướng dẫn*: Ths.**DƯƠNG THỊ KIM CHI**

*Người thực hiện*: **TRẦN QUỐC DŨNG – 1824801040013**

**HOÀNG KIM TUYẾN – 1824801040043**

Lớp **: D18HT01**

Khoá  **: D18**

**BÌNH DƯƠNG, NĂM 2021**

LỜI CẢM ƠN

Nhóm em xin gửi tới lời cảm ơn chân thành đến cô **Dương Thị Kim Chi** – giảng viên môn “*Đồ Án Chuyên Ngành*” đã trang bị cho chúng em những kiến thức và kỹ năng cần có để thực hiện đề tài báo cáo cuối môn này.

Trong quá trình thực hiện đồ án cuối môn, nhóm em đã cố gắng áp dụng những kiến thức **Cô** đã giảng dạy vào trong đồ án. Tuy nhiên trong quá trình thực hiện đồ án nhóm em còn thiếu sót, hay những đánh giá về đề tài. Rất mong nhận được sự quan tâm, góp ý của **Cô** để đồ án của nhóm em được hoàn chỉnh và tốt hơn.

**Xin chân thành cảm ơn!**

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của **Cô Dương Thị Kim Chi**. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.Trường Đại học Thủ Dầu Một không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Thủ Dầu Một, ngày 6 tháng 12 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Hoàng Kim Tuyến Trần Quốc Dũng*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

TP. Thủ Dầu Một, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[DANH MỤC CÁC HÌNH VẼ 3](#_Toc89667403)

[DANH MỤC CÁC BẢNG 4](#_Toc89667404)

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 5](#_Toc89667405)

[1.1 Lý do chọn đề tài 5](#_Toc89667406)

[1.2 Mục đích nghiên cứu 5](#_Toc89667407)

[1.3 Đối tượng nghiên cứu và phạm vị nghiên cứu 6](#_Toc89667408)

[1.4 Phương pháp nghiên cứu 6](#_Toc89667409)

[1.5 Bố cục của đồ án 6](#_Toc89667410)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 7](#_Toc89667411)

[2.1 Ngôn ngữ lập trình python 7](#_Toc89667412)

[2.2 Supervised learning 8](#_Toc89667413)

[2.3 Support vector machine(SVM) 9](#_Toc89667414)

[2.3.1 Tổng quan về SVM hỗ trợ 10](#_Toc89667415)

[2.4 Text mining 11](#_Toc89667416)

[CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG VÀ THỰC NGHIỆM 12](#_Toc89667417)

[3.1 Các bước xây dựng bài toán 12](#_Toc89667418)

[3.2 Dữ liệu 12](#_Toc89667419)

[3.2.1 Thu thập dữ liệu 12](#_Toc89667420)

[3.2.2 Xử lý và làm sạch dữ liệu dữ liệu 14](#_Toc89667421)

[3.3 Train, test, và predict mode SVM 15](#_Toc89667422)

[3.4 ktinker python lập trình ứng dụng 16](#_Toc89667423)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CHO ĐỀ TÀI 20](#_Toc89667424)

[4.1 Kết luận 20](#_Toc89667425)

[4.2 Hướng phát triển 20](#_Toc89667426)

[TÀI LIỆU THAM KHẢO 21](#_Toc89667427)

DANH MỤC CÁC HÌNH VẼ

[Hình 1: Mô hình học có giám sát 9](#_Toc89665997)

[Hình 2: Code cào dữ liệu từ báo Dân Trí 14](#_Toc89665998)

[Hình 3: File data.csv lưu trữ thông tin bài báo 15](#_Toc89665999)

[Hình 4: Xử lý stopword trong dữ liệu 16](#_Toc89666000)

[Hình 5: train, test dữ liệu 17](#_Toc89666001)

[Hình 6: Kết quả của quá trình train dữ liệu 17](#_Toc89666002)

[Hình 7: Giao diện ứng dụng trích xuất thông tin bài báo 18](#_Toc89666003)

DANH MỤC CÁC BẢNG

[Bảng : Mô tả dữ liệu của đề tài…………………………………………………….18](#_Toc89666003)

1. TỔNG QUAN VỀ ĐỀ TÀI
   1. Lý do chọn đề tài

Theo như chúng ta biết hiện nay tất cả thông tin trên các trang báo đều được đăng tải lên mạng xã hội là dạng văn bản. Nguồn tin được cung cấp tới chúng ta hàng ngày, các trang báo luôn đăng tải thông tin bài viết của họ lên mạng xã hội để cho người đọc biết được sự vật, hiện tượng hay sự kiên nổi bật trong ngày mà mạng xã hội quan tâm và bàn luận.

Tuy nhiên một tình trạng mà các trang báo nào cũng gặp phải đó là việc đăng tải thông tin nội dung của bài báo lên mạng quá dài, khiến cho người đọc có thể xảy ra trang thái lười đọc hoặc không đọc vì nội dung của bài báo quá dài và không tập trung vào nội dung chính.

Nhưng với thời đại công nghệ 4.0 tiên tiến việc áp dụng công nghệ vào xử lý các các nội dung văn bản là điều cần thiết. Chính vì vậy nhóm em đã chọn đề tài *“Xây dựng ứng dụng trích xuất thông tin bài báo bằng thuật toán SVM*” nhằm tạo ra một hệ thống hỗ trợ người đọc trích xuất nội dung thông của một bài báo một cách ngắn nhất mà vẫn giữ được nội dung chính của bài báo, hệ thống sẽ sử dụng thuật toán SVM để trích xuất đặc trưng và kích thước của bài báo sẽ được giảm xuống.

* 1. Mục đích nghiên cứu

Đồ án *“Xây dựng ứng dụng trích xuất thông tin bài báo bằng thuật toán svm*” hướng tới xây dựng hệ thống ứng dụng trích xuất thông tin của bài báo, hệ thống sẽ dựa vào thuật toán SVM để trích xuất tóm tắt ngắn nhất thông tin của bài báo. Thông qua việc nghiên cứu phương pháp học sâu và ngôn ngữ lập trình python để xây dựng ứng dụng.

* 1. Đối tượng nghiên cứu và phạm vị nghiên cứu

Với mục đích xây dựng được ứng dụng trích xuất thông tin bài báo, đề tài xác định đối đượng và phạm vi nghiên cứu như sau:

* Đối tượng nghiên cứu: nội dung của các bài báo
* Nghiên cứu và sử dụng giải thuật SVM
* Nền tảng ngôn ngữ lập trình python
* Phạm vị nghiên cứu: nội dung các bài báo của báo Dân Trí
  1. Phương pháp nghiên cứu
* Phương pháp nghiên cứu lý thuyết: Nghiên cứu các tài liệu về trích xuất thông tin, nghiên cứu giải thuật SVM.
* Phương pháp nghiên cứu thực tiễn: Sau khi nghiên cứu kĩ tài liệu về khai phá văn bản, xác định được các bước để thực hiện, để đưa ra hướng giải quyết cho đề tài, ứng dụng vào xây dựng hệ thống và đưa ra kết quả của sản phẩm.
  1. Bố cục của đồ án

Nội dung của đồ án được chia thành bố cục như sau:

Chương 1: Tổng quan về đề tài

Chương 2: Cơ sở lý thuyết

Chương 3: Xây dựng ứng dụng và thực nghiệm

Chương 4: Kết luận và hướng phát triển cho đề tài

1. CƠ SỞ LÝ THUYẾT
   1. Ngôn ngữ lập trình python

Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level) ngữ nghĩa động. Python là ngôn ngữ lập trình hướng đối tượng thông dụng dùng để viết các tiện ích hệ thống. Nó cũng được sử dụng như ngôn ngữ kết dính đóng vai trò tích hợp C và C++.

Được tạo ra bởi Guido van Rossum tại Amsterdam năm 1990, Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý.

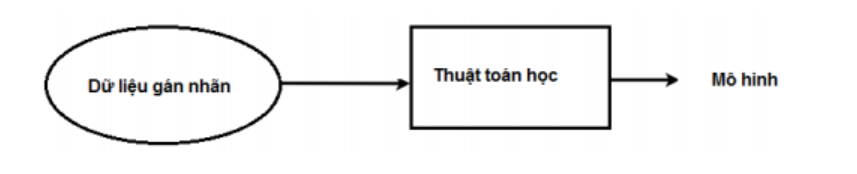
Python là ngôn ngữ có hình thức khá đơn giản và rõ ràng, do đó tạo nên sự dễ dàng tiếp cânh cho những lập trình viên mới bắt đầu.

Ban đầu, Python được phát triển để chạy trên nền Unix, nhưng rồi theo thời gian, nó đã được mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Windows, Linux và các hệ điều hành khác thuộc họ Unix.

Một số tính năng nổi bật của python như:

* **Miễn phí, mã nguồn mở**
* **Ngôn ngữ lập trình đơn giản, dễ đọc**
* **Khả năng di chuyển**
* **Khả năng mở rộng và có thể nhúng**
* Ngôn ngữ thông dịch cấp cao
* **Hướng đối tượng**
* **Thư viện tiêu chuẩn lớn**
* **Khoa học, phân tích số liệu**
  1. Supervised learning

Học có giám sát (supervised learning) là một kỹ thuật của ngành học máy nhằm mục đích xây dựng một hàm f từ dữ tập dữ liệu huấn luyện (Training data). Dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào và đầu ra mong muốn. Đầu ra của hàm f có thể là một giá trị liên tục hoặc có thể là dự đoán một nhãn phân lớp cho một đối tượng đầu vào. Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm f cho một đối tượng đầu vào hợp lệ bất kì, sau khi đã xét một số mẫu dữ liệu huấn luyện (nghĩa là các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa từ các dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách hợp lý.



Hình 1: Mô hình học có giám sát

* Cách giải một bài toán học có giám sát

Để giải một bài toán học có giám sát ta thực hiện theo các bước sau:

**Bước 1:** Xác định loại của các dữ liệu huấn luyện: Trước tiên ta cần phải quyết định xem loại dữ liệu nào sẽ được sử dụng làm dữ liệu huấn luyện. Ta có thể chọn dữ liệu một kí tự viết tay đơn lẻ, toàn bộ một từ viết tay, hay toàn bộ một dòng chữ viết tay, …

**Bước 2:** Thu thập tập dữ liệu huấn luyện. Khi thu thập tập dữ liệu huấn luyện cần phải đảm bảo được sự đặc trưng cho thực tế sử dụng của hàm chức năng. Do đó tập các dữ liệu đầu vào và đầu ra tương ứng phải được thu thập từ các chuyên gia hoặc từ việc đo đạc tính toán.

**Bước 3:** Xác định việc biễu diễn các đặc trưng đầu vào cho hàm mục tiêu cần tìm. Độ chính xác của mục tiêu phụ thuộc rất lớn vào các đối tượng đầu vào được biểu diễn như thế nào. Đa số các đối tượng đầu vào được chuyển đổi thành một véc tơ đặc trưng chứa các đặc trưng cơ bản của đối tượng đó. Chú ý số lượng các đặc trưng không được lớn quá, để tránh sự bùng nổ tổ hợp tuy nhiên nó phải đủ lớn để đảm bảo dự đoán chính xác đầu ra.

**Bước 4:** Xác định cấu trúc của hàm mục tiêu cần tìm và giải thuật học tương  
ứng. Ví dụ, ta có thể sử dụng mạng nơ-ron nhân tạo, cây quyết định, …

**Bước 5:** Hoàn thiện và thiết kế chương trình.

* Dưới đây là một số bài toán thuộc loại học có giám sát:
* Phân loại thư rác
* Phân loại chuyên mục bài báo
* Dự đoán thẻ tag cho bài báo
* Xác định ảnh một con vật có phải mèo không?
* Xác định giới tính qua khuôn mặt trong ảnh
* Nhận dạng chữ số viết tay
* Nhận dạng tiếng nói
* Nhận dạng ký tự quang học
  1. Support vector machine(SVM)

Máy vectơ hỗ trợ (SVM - viết tắt tên tiếng Anh support vector machine) là một khái niệm trong [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) và [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) cho một tập hợp các phương pháp [học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t) liên quan đến nhau để [phân loại](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_(h%E1%BB%8Dc_m%C3%A1y)&action=edit&redlink=1) và [phân tích hồi quy](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy). SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó SVM là một [thuật toán phân loại](https://vi.wikipedia.org/w/index.php?title=Thu%E1%BA%ADt_to%C3%A1n_ph%C3%A2n_lo%E1%BA%A1i&action=edit&redlink=1) [nhị phân](https://vi.wikipedia.org/wiki/Ph%C3%A2n_lo%E1%BA%A1i_nh%E1%BB%8B_ph%C3%A2n). Với một bộ các ví dụ luyện tập thuộc hai thể loại cho trước, thuật toán luyện tập SVM xây dựng một mô hình SVM để phân loại các ví dụ khác vào hai thể loại đó. Một mô hình SVM là một cách biểu diễn các điểm trong không gian và lựa chọn ranh giới giữa hai thể loại sao cho khoảng cách từ các ví dụ luyện tập tới ranh giới là xa nhất có thể. Các ví dụ mới cũng được biểu diễn trong cùng một không gian và được thuật toán dự đoán thuộc một trong hai thể loại tùy vào ví dụ đó nằm ở phía nào của ranh giới.

* + 1. Tổng quan về SVM hỗ trợ

Một máy vectơ hỗ trợ xây dựng một [siêu phẳng](https://vi.wikipedia.org/wiki/Si%C3%AAu_ph%E1%BA%B3ng) hoặc một tập hợp các siêu phẳng trong một không gian nhiều chiều hoặc vô hạn chiều, có thể được sử dụng cho phân loại, hồi quy, hoặc các nhiệm vụ khác. Một cách trực giác, để phân loại tốt nhất thì các siêu phẳng nằm ở càng xa các điểm dữ liệu của tất cả các lớp (gọi là hàm lề) càng tốt, vì nói chung lề càng lớn thì [sai số tổng quát hóa](https://vi.wikipedia.org/w/index.php?title=Sai_s%E1%BB%91_t%E1%BB%95ng_qu%C3%A1t_h%C3%B3a&action=edit&redlink=1) của thuật toán phân loại càng bé.

Trong nhiều trường hợp, không thể phân chia các lớp dữ liệu một cách tuyến tính trong một không gian ban đầu được dùng để mô tả một vấn đề. Vì vậy, nhiều khi cần phải [ánh xạ](https://vi.wikipedia.org/wiki/%C3%81nh_x%E1%BA%A1) các điểm dữ liệu trong không gian ban đầu vào một không gian mới nhiều chiều hơn, để việc phân tách chúng trở nên dễ dàng hơn trong không gian mới. Để việc tính toán được hiệu quả, ánh xạ sử dụng trong thuật toán SVM chỉ đòi hỏi [tích vô hướng](https://vi.wikipedia.org/wiki/T%C3%ADch_v%C3%B4_h%C6%B0%E1%BB%9Bng) của các vectơ dữ liệu trong không gian mới có thể được tính dễ dàng từ các tọa độ trong không gian cũ. Tích vô hướng này được xác định bằng một hàm hạt nhân K(x,y) phù hợp. Một siêu phẳng trong không gian mới được định nghĩa là tập hợp các điểm có tích vô hướng với một vectơ cố định trong không gian đó là một hằng số. Vectơ xác định một siêu phẳng sử dụng trong SVM là một tổ hợp tuyến tính của các vectơ dữ liệu luyện tập trong không gian mới với các hệ số αi. Với siêu phẳng lựa chọn như trên, các điểm x trong không gian đặc trưng được ánh xạ vào một siêu mặt phẳng là các điểm thỏa mãn:

Σ*i* *αi* *K*(*xi*,*x*) = hằng số.

Ghi chú rằng nếu K(x,y) nhận giá trị ngày càng nhỏ khi y xa dần khỏi x thì mỗi số hạng của tổng trên được dùng để đo độ tương tự giữa x với điểm xi tương ứng trong dữ liệu luyện tập. Như vậy, tác dụng của tổng trên chính là so sánh khoảng cách giữa điểm cần dự đoán với các điểm dữ liệu đã biết. Lưu ý là tập hợp các điểm x được ánh xạ vào một siêu phẳng có thể có độ phức tạp tùy ý trong không gian ban đầu, nên có thể phân tách các tập hợp thậm chí không lồi trong không gian ban đầu.

* 1. Text mining

Khai thác văn bản (text mining hoặc text data mining) là một quá trình xử lý và trích xuất [thông tin](https://vi.wikipedia.org/wiki/Th%C3%B4ng_tin) nằm trong [văn bản](https://vi.wikipedia.org/wiki/V%C4%83n_b%E1%BA%A3n), quá trình này là một phần của việc [phân tích văn bản](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%ADch_v%C4%83n_b%E1%BA%A3n&action=edit&redlink=1) trong [khai thác dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u). Thông tin được thể hiện dưới dạng các mẫu, xu hướng, thứ tự sắp xếp được trích xuất thông qua các luật hoặc thông qua quá trình học dựa trên các mẫu [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA).

Khai thác văn bản bao gồm các bước cơ bản như: tiền xử lý, học mô hình, phán đoán, tổng hợp phân tích và trình bày kết quả. Tiền xử lý có thể gồm việc phân tách đoạn văn bản thành các đoạn nhỏ hơn, làm giàu văn bản bằng các tri thức bên ngoài, hoặc loại bỏ những thông tin nhiễu trong văn bản. Quá trình học là quá trình tìm ra các mẫu trong một tập các văn bản đã được tiền xử lý hoặc chưa qua tiền xử lý, kết quả quá trình học là một mô hình biểu diễn các mẫu được tìm thấy. Quá trình [phán đoán](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A1n_%C4%91o%C3%A1n&action=edit&redlink=1) là quá trình áp dùng mô hình vừa học được trên các văn bản mới, văn bản mới sẽ được gán nhãn thêm thông tin. Cuối cùng là quá trình tổng hợp và trình bày kết quả. Khai phá văn chia thành các vấn đề nhỏ hơn bao gồm [phân loại tài liệu](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_t%C3%A0i_li%E1%BB%87u&action=edit&redlink=1) (text categorization, text classification), [gom cụm văn bản](https://vi.wikipedia.org/w/index.php?title=Gom_c%E1%BB%A5m_v%C4%83n_b%E1%BA%A3n&action=edit&redlink=1) (text clustering), [trích xuất thực thể](https://vi.wikipedia.org/w/index.php?title=Tr%C3%ADch_xu%E1%BA%A5t_th%E1%BB%B1c_th%E1%BB%83&action=edit&redlink=1) (concept/entity extraction), [phân tích tình cảm](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_t%C3%ACnh_c%E1%BA%A3m) (sentiment analysis), [tóm tắt tài liệu](https://vi.wikipedia.org/wiki/T%C3%B3m_t%E1%BA%AFt_t%C3%A0i_li%E1%BB%87u) (document summarization), và trích xuất quan hệ giữa các thực thể (entity relation modeling).

1. XÂY DỰNG ỨNG DỤNG VÀ THỰC NGHIỆM
   1. Các bước xây dựng bài toán

Dựa vào cơ sở lý thuyết của đề tài để xây dựng hệ thống theo các bước như sau:

Bước 1: Cào dữ liệu bằng beatyfulsoup, newpaper3k

Bước 2: Làm sạch dữ liệu: stopword, xóa dấu cách, tokenzie

Bước 3: Train, test, predict model SVM

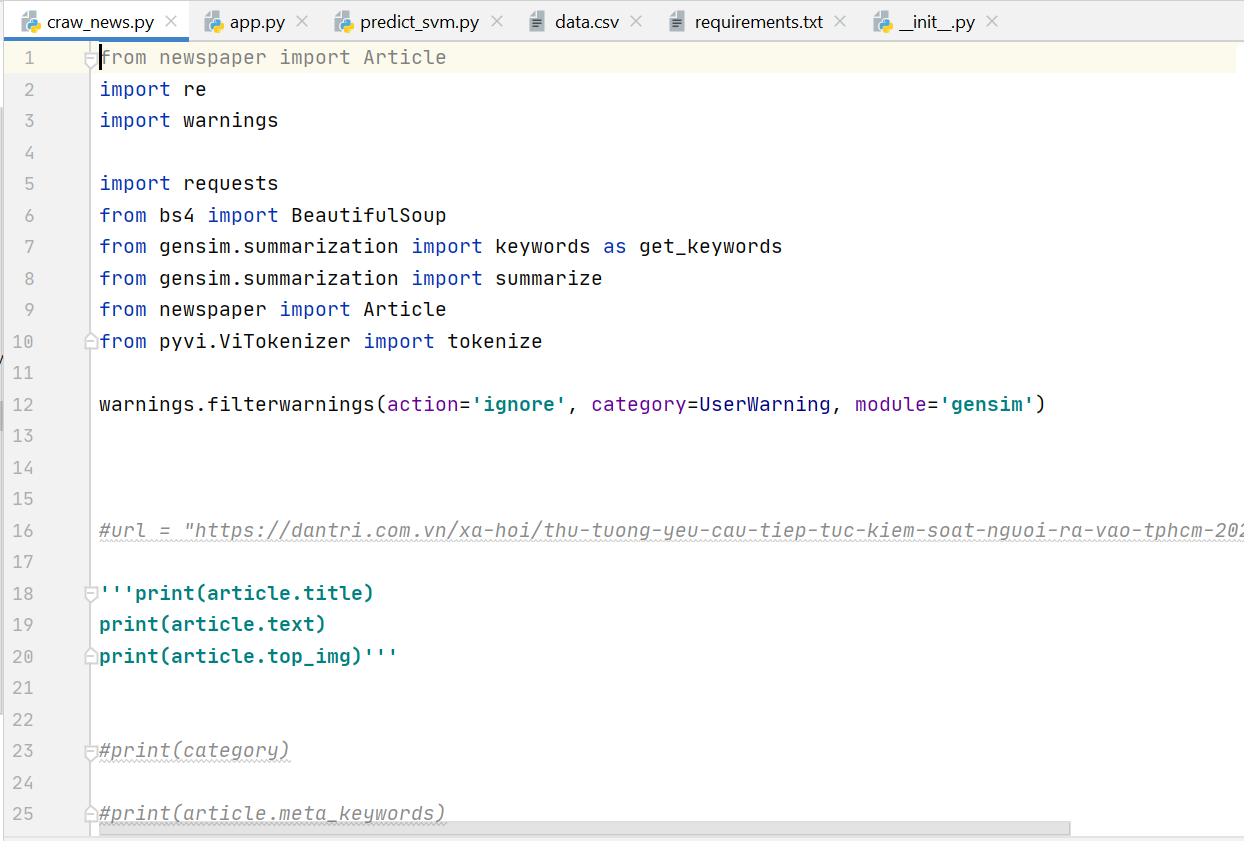
Bước 4: ktinker python lập trình ứng dụng

* 1. Dữ liệu
     1. Thu thập dữ liệu

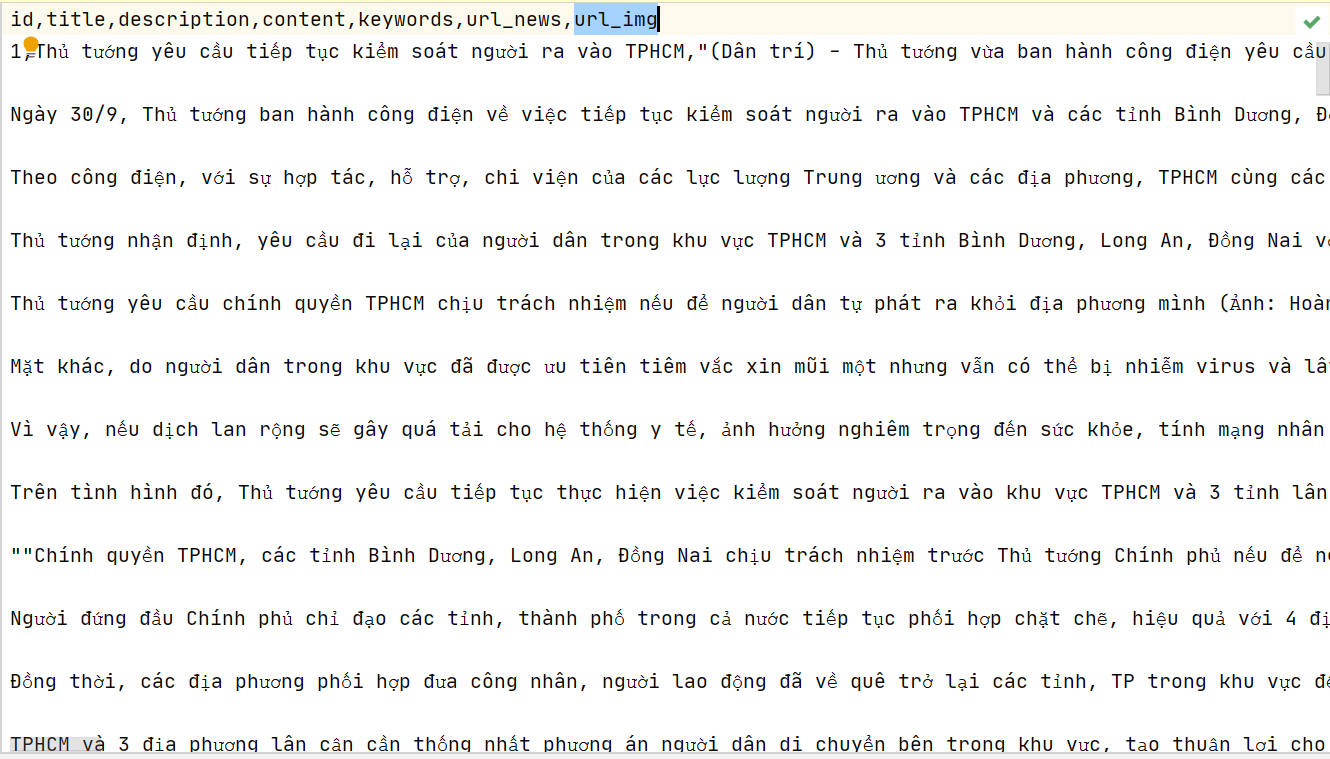
|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Giải thích |
| 1 | id | Number | Số thứ tự |
| 2 | title | Text | Tiêu đề bài báo |
| 3 | description | Text | Mô tả bài báo |
| 4 | content | Text | Nội dung bài báo |
| 5 | keywords | Text | Từ khóa của bài báo |
| 6 | url\_news | Text | Link bài báo |
| 7 | url\_img | Text | Link hình ảnh |

Bảng 1: Mô tả dữ liệu của đề tài

Dữ liệu được lấy từ nguồn báo dantri.com.vn, 926 dòng tương dương với số bài báo mà đề tài đã lấy. Trong quá trình cào dữ liệu đề tài đã sử dụng thư viên hỗ trợ trong python để tiền hành cào dữ liệu xuống như: beatifulsoup, newpager3k. Và dữ liệu được lưu trữ trong file csv có tên data.csv. Sau đây là ảnh minh họa của qua trình cao dữ liệu:

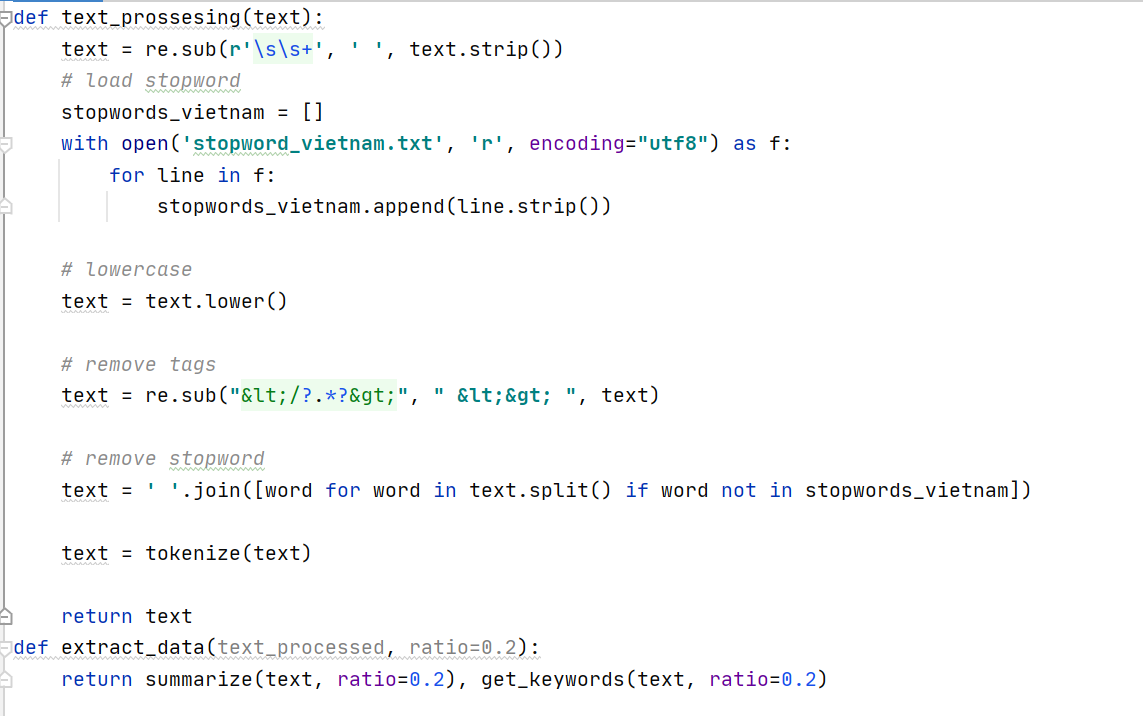


Hình 2: Code cào dữ liệu từ báo Dân Trí



Hình 3: File data.csv lưu trữ thông tin bài báo

* + 1. Xử lý và làm sạch dữ liệu dữ liệu
* Xử lý stopword cho dữ liệu. Sử dụng các stopword của ngôn ngữ Việt Nam để loại bỏ đi các từ không cần thiết trong dữ liệu. Trong quá trình sạch dữ liệu đã sử dụng thư viên tokenize để loại bỏ các từ không cần thiết. Sau đây là code của Xử lý stopword và làm sạch dữ liệu:



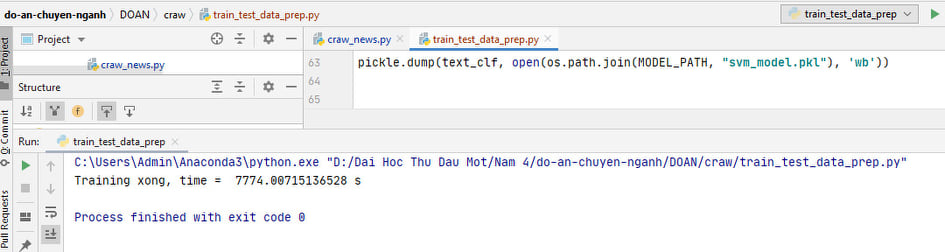
Hình 4: Xử lý stopword trong dữ liệu

* 1. Train, test, và predict mode SVM

Sau khi xử lý dữ liệu và làm sạch dữ liệu 🡪 có được dữ liệu để thực hiên quá trình train và test dữ liệu, dữ liệu được đem đi train theo tỷ lệ 8/2. Sau đây là code và kết quả của quá trình train và test dữ liệu:



Hình 5: train, test dữ liệu



Hình 6: Kết quả của quá trình train dữ liệu

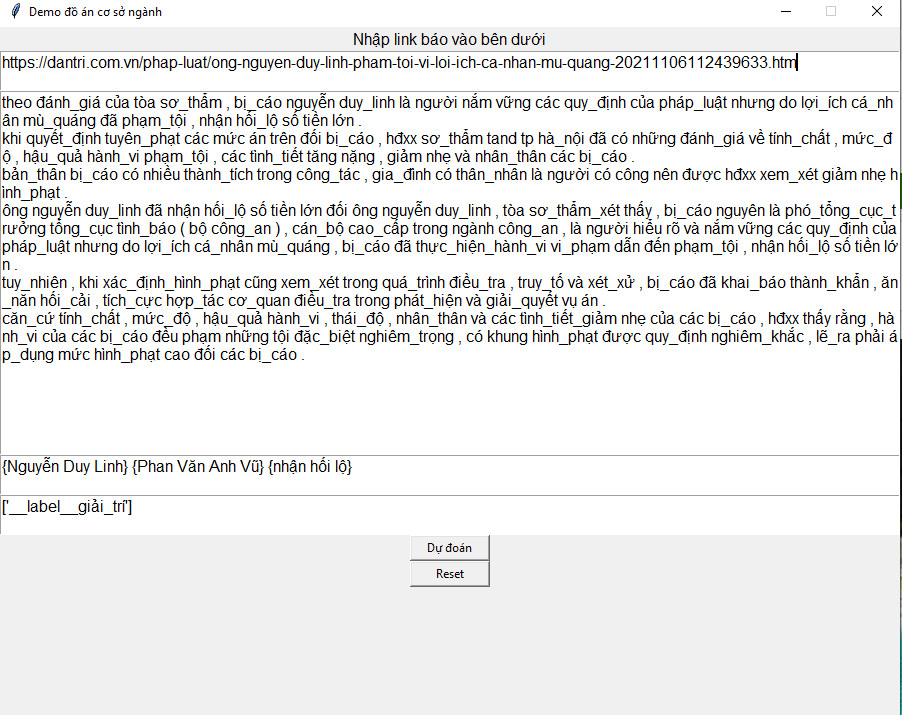
* 1. ktinker python lập trình ứng dụng

Sau khi quá trình train và test dữ liệu hoàn chỉnh và có được model thì tiếp theo là phần xây dựng giao diện. Giao diện ở đây được xây dựng đơn giản bằng cách sử dụng thư viên ktinker trong python. Thực hiện code giao diện đơn giản như sau:

app = Tk()  
app.resizable(**False**, **False**)  
app.geometry(**"1200x600"**)  
app.title(**"Demo đồ án cơ sở ngành"**)  
  
*# nhập link bài báo*lbl\_link\_news = Label(app, text=**"Nhập link báo vào bên dưới:"**, font=(**"Helvetica"**, 12))  
lbl\_link\_news.pack(pady=10, side= TOP, anchor=**"w"**)  
input\_link\_news = Text(app, height=1, width=100,font=(**"Helvetica"**, 12))  
input\_link\_news.pack()  
input\_link\_news.focus()  
  
*# tiêu đề bài báo*lbl\_title = Label(app, text=**"Tiêu đề:"**, font=(**"Helvetica"**, 12))  
lbl\_title.pack(pady=5, side= TOP, anchor=**"w"**)  
input\_title = Text(app, height=1, width=100,font=(**"Helvetica"**, 12))  
input\_title.pack()  
  
*# nội dung tóm tắt bài báo*lbl\_content = Label(app, text=**"Nội dung tóm tắt:"**, font=(**"Helvetica"**, 12))  
lbl\_content.pack(pady=5, side= TOP, anchor=**"w"**)  
input\_content = Text(app, height=12, width=100,font=(**"Helvetica"**, 12))  
input\_content.pack()  
  
*# keyword bài báo*lbl\_keyword = Label(app, text=**"Keyword:"**, font=(**"Helvetica"**, 12))  
lbl\_keyword.pack(pady=5, side= TOP, anchor=**"w"**)  
input\_keyword = Text(app, height=2, width=100,font=(**"Helvetica"**, 12))  
input\_keyword.pack()  
  
*# nhãn bài báo*lbl\_lable = Label(app, text=**"Nhãn dự đoán:"**, font=(**"Helvetica"**, 12))  
lbl\_lable.pack(pady=5, side= TOP, anchor=**"w"**)  
input\_lable = Text(app, height=2, width=100,font=(**"Helvetica"**, 12))  
input\_lable.pack()  
  
  
*# button dự đoán*btn\_predict = Button(app,height=1, width=10, text=**"Dự đoán"**, command=predict)  
btn\_predict.pack()  
  
*# button reset*btn\_reset = Button(app,height=1, width=10, text=**"Reset"**, command=reset)  
btn\_reset.pack()  
  
app.mainloop()

Tiếp theo là hàm dự đoán của button:

**def** predict():  
 link\_news = input\_link\_news.get(**"1.0"**,**'end-1c'**)  
 *# xóa dữ liệu cũ* input\_title.delete(**"1.0"**,**'end-1c'**)  
 input\_content.delete(**"1.0"**,**'end-1c'**)  
 input\_keyword.delete(**"1.0"**,**'end-1c'**)  
 input\_lable.delete(**"1.0"**,**'end-1c'**)  
  
 *# lấy dữ liệu bài báo* **try**:  
 \_title, \_des, \_keyword, \_text = get\_data\_news(link\_news)  
 \_sumaried\_text, \_key = extract\_data(\_text)  
 **except** Exception **as** erro:  
 reset()  
 messagebox.showwarning(**"Cảnh báo"**, **"Hãy nhập đúng đường dẫn!"**)  
  
 *# dự đoán nhãn bài báo* lable\_news = news\_predict(link\_news)  
  
 *# add dữ liệu lên form* input\_title.insert(**"1.0"**, \_title)  
 input\_content.insert(**"1.0"**,\_sumaried\_text)  
 input\_keyword.insert(**"1.0"**, \_keyword)  
 input\_lable.insert(**"1.0"**, lable\_news)  
  
 *# không cho sửa textbox* input\_title.config(state=**'disabled'**)  
 input\_content.config(state=**'disabled'**)  
 input\_lable.config(state=**'disabled'**)  
 input\_keyword.config(state=**'disabled'**)



Hình 7: Giao diện ứng dụng trích xuất thông tin bài báo

1. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CHO ĐỀ TÀI
   1. Kết luận

Trong đề tài “Xây dựng ứng dụng trích xuất thông tin bài báo bằng thuật toán SVM” chúng em đã nghiên cứu về quy trình các bước trong text mining và thuật toán SVM, trong thuật toán này chúng em đã nghiên cứu các kỹ thuật xây dựng nên ứng dụng trích xuất thông tin bài báo.

Từ việc cào dữ liệu từ báo Dân Trí và sau đó là quá trình xử lý và làm sạch dữ liệu, tiếp theo là đem dữ liệu đi train và test sau đó có được model để xây dựng ứng dụng. Đề tài đã xây dựng được quy trình trích xuất thông tin từ bài báo. Và xây dựng được ứng dụng trích xuất thông tin bài báo từ báo Dân Trí

* 1. Hướng phát triển

Trong thời gian tới chúng em sẽ cải tiến đề tài thêm nhằm tạo ra một ứng dụng dùng để trích xuất thông tin của bài báo, và có thể phát triển hơn không chỉ đơn giản là một ứng dụng mà còn có thể sử dụng trên website. Và có thể phát triển đề tài thành ứng dụng trích xuất thông tin cho tất cả các báo không chỉ là trên báo Dân Trí.

# TÀI LIỆU THAM KHẢO

[1] “Bài 19: Support Vector Machine”, trang học trực tuyến machinelearning cơ bản, đăng tải ngày: 9/8/2017.

[2] “Bài 20: Soft Margin Support Vector Machine”, trang học trực tuyến machinelearning cơ bản, đăng tải ngày: 13/8/2017.

[3] “Bài 21: Kernel Support Vector Machine”, trang học trực tuyến machinelearning cơ bản, đăng tải ngày: 22/8/2017.