**ĐỀ XUẤT ĐỀ TÀI**

PHÂN LOẠI TIÊU ĐỀ BÀI BÁO THÀNH CÁC CHỦ ĐỀ CHO TRƯỚC

**GVHD**: TS. Nguyễn Nhật Quang

**Danh sách thành viên nhóm**:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Họ và tên | MSSV | Email |
| 1 | Nguyễn Trung Kiên | 20194088 | Kien.nt194088@sis.hust.edu.vn |
| 2 | Hoàng Ngọc Lâm | 20194089 | Lam.hn194089@ sis.hust.edu.vn |
| 3 | Đào Nguyễn Tiến Huy | 20194077 | Huy.dnt194077@ sis.hust.edu.vn |
| 4 | Mai Văn Tiến | 20194183 | Tien.mv194183@ sis.hust.edu.vn |

# **Mô tả bài toán thực tế**

Phân loại văn bản là bài toán phổ biến trong xử lý ngôn ngữ tự nhiên. Đối với phân loại văn bản tiếng Việt, sẽ khó khăn hơn đôi chút so với phân loại văn bản tiếng Anh. Trong đề tài này nhóm sẽ đi giải quyết bài toán phân loại văn bản bằng tiếng Việt – phân loại tiêu đề bài báo thành các chủ đề cho trước.

Ứng dụng của bài toán phổ biến, chúng ta có thể sử dụng mô hình này để phân loại văn bản tin tức tiếng Việt cho một trang báo điện tử. Mỗi khi một bài báo được đăng, chương trình giúp tự động xác định được bài báo đó nằm trong chủ đề nào. Các chủ đề gồm có: thời sự, đời sống, giải trí, khoa học, kinh doanh, pháp luật, sức khỏe, thể thao, du lịch, giáo dục (10 chủ đề).

# **Giải thuật học máy đề xuất**

Sau một thời gian tìm hiểu, nhóm dự định sử dụng thuật toán Naive Bayes hoặc Neural network đơn giản (chỉ gồm 1, 2 hidden layers).

# **Thông tin về input, output của hệ thống ML sẽ được cài đặt và cách biểu diễn các ví dụ học**

* Input của hệ thống sẽ là tiêu đề của bài báo.
* Output là chủ đề được phân loại.
* Learning examples là câu tiếng việt ngắn, sẽ được xử lý thành các vectors chứa thông tin stop words.

# **Datasets**

Dữ liệu là tiêu đề của các bài báo trên trang vnexpress.net được nhóm tự crawl bằng các packages Selenium, Beautiful soup... Số lượng data examples sẽ gồm khoảng 10 000 examples.

Ví dụ:

* Thời sự:

+ Chuyên gia đề nghị bỏ cách ly với khách quốc tế ở vùng xanh.

+ Nước lũ gây ngập nhiều nơi ở Quảng Nam, Thừa Thiên Huế.

+ Lội kênh hái rau muống dại mưu sinh ở thành phố HCM.

+ Áp lực khi thiếu nhà ở cho công nhân.

* Đời sống:

+ Cậu bé người rừng đi học.

+ Tặng đất miễn phí cho người dân.

+ Tủ, giường gắn trên trần nhà.

+ Cô gái Thụy Điển trả ơn Việt Nam sau khi tìm được mẹ.

* Giải trí:

### + [Sao Việt theo đuổi mốt áo cài một cúc](https://vnexpress.net/sao-viet-theo-duoi-mot-ao-cai-mot-cuc-4375459.html).

### + ['Mỹ nhân gợi cảm nhất thế giới' mặc đầm Công Trí](https://vnexpress.net/my-nhan-goi-cam-nhat-the-gioi-mac-dam-cong-tri-4376017.html).

### + [Tài xế có lỗi trong cái chết của sao 'Thủy Hử'](https://vnexpress.net/tai-xe-co-loi-trong-cai-chet-cua-sao-thuy-hu-4376160.html).

### + [Phim Việt đầu tiên công bố lịch ra rạp](https://vnexpress.net/phim-viet-dau-tien-cong-bo-lich-ra-rap-4376101.html).

* Khoa học:

+ Ý tưởng gắn động cơ vào Nam cực để đẩy trái đất.

+ Lỗ hổng dài 100 km ở băng biển Bắc cực.

+ Khối ngọc tím 2000 năm trong cống nước.

+ Quá khứ của thiên hà sắp bị dải Ngân hà nuốt chửng.

* Kinh doanh:

+ Doanh nghiệp gặp khó khi lên kế hoạch phục hồi sản xuất.

+ Sếp doanh nghiệp Việt Nam nhận lương ít hơn nước ngoài một tỉ đồng.

+ Xuất khẩu hải sản 9 tháng đạt 2,4 tỷ USD.

+ Mỹ thiếu bí ngô chơi Halloween.

* Pháp luật:

+ Xét sử trực tuyến chỉ nên áp dụng với những vụ án ít nghiêm trọng.

+ Cựu trung tá công an ba lần mượn tay côn đồ trả thù cấp dưới.

+ Mua súng để giải quyết mâu thuẫn.

* Sức khỏe:

+ Người khám bệnh tình dục tăng sau nới giãn cách.

+ Phát hiện ung thư ruột thừa từ triệu chứng đau bụng.

+ Nên ăn hoa quả lúc đói hay no.

* Thể thao:

+ Trọng tài cho điểm Thu Nhi như thế nào?

+ Ronaldo: “Tôi sẽ buộc miệng những kẻ chỉ trích”.

+ CĐV đề nghị loại Norwich khỏi ngoại hạng anh.

* Du lịch:

+ Một lần đến thăm nhà siêu sao Messi ở Argentina.

+ Tour liên tỉnh đầu tiên khởi hành từ Hà Nội.

+ Việt Nam nhận giải điểm đến hành đầu châu Á.

+ Ẩm thực Địa Trung Hải “thú vị và duyên dáng”.

* Đời sống:

+ Cậu bé người rừng đi học.

+ Đặt tên 12 con bằng 4 chữ cái.

+ Bí quyết thành công của người đàn ông không chân.

* Giáo dục:

+ Bé 2 tuổi bị bạn đánh trong lớp.

+ Trung Quốc thông qua luật cắt giảm bài tập về nhà.

+ Thủ khoa học viện cảnh sát nhân dân được phong hàm vượt cấp.

+ Trường học loay hoay đủ cách để dạy môn tích hợp lớp 6.

# **Kế hoạch thực hiện**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên nhiệm vụ** | **Người tham gia** | **Ngày bắt đầu** | **Ngày kết thúc** |
| Viết đề xuất đề tài | Nguyễn Trung Kiên | 22/10/2021 | 24/10/2021 |
| Crawl data | Mai Văn Tiến | 25/10/2021 |  |
| Tìm hiểu kĩ về thuật toán Naive Bayes và NN sau đó chốt thuật toán sẽ sử dụng | Nguyễn Trung Kiên, Hoàng Ngọc Lâm, Đào Nguyễn Tiến Huy, Mai Văn Tiến | 25/10/2021 |  |
| Training data, data preprocess, Feature engineering | Nguyễn Trung Kiên | 1/11/2021 |  |
| Training model | Nguyễn Trung Kiên, Hoàng Ngọc Lâm, Đào Nguyễn Tiến Huy, Mai Văn Tiến | unknow |  |
| Evaluation model | nt | unknow |  |
| Test model | nt | unknow |  |
| Viết báo cáo | Đào Nguyễn Tiến Huy | unknow |  |
| Thuyết trình/ bảo vệ bài tập lớn | Hoàng Ngọc Lâm | unknow |  |