ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHUYẾN NGHỊ TIN TỨC TRÊN VNEXPRESS

Sinh viên thực hiên

18521493 - Trần Nhật Tiến

18520568 - Kiên Tiến Đạt

18521239- Nguyễn Tấn Phong

GVHD : Huỳnh Văn Tín

TP. HỒ CHÍ MINH – 12/2021

1. GIỚI THIỆU

Với sự tiến bộ của công nghệ truyền thông tương tác, Internet đã trở thành một nguồn tin tức chính do tính khả dụng 24/7 cùng với sự cập nhật tức thời và miễn phí. Đặc biệt hiện nay đang thời buổi dịch bệnh, nên nhu cầu đọc tin của mọi người càng tăng cao do đó việc gợi ý ra các bài báo khác, dựa trên những bài báo mà người dùng đã đọc là hết sức cần thiết để giúp người dùng tiếp nhận được những thông tin liên quan và chính xác. Nhiều nguồn tin tức phổ biến ở Việt Nam như VnExpress, Báo Mới, Đời sống và Pháp luật, Vietnamnet,… Trong đồ án này nhóm em làm về bài toán Khuyến nghị tin tức từ báo Vnexpress. Vnexpress là 1 trong các trang báo mạng uy tín, báo điện tử nhanh nhất Việt Nam với khả năng cập nhật liên tục các tin tức sự kiện về chính trị, xã hội, văn hóa, kinh tế, thể thao, âm nhạc, thời tiết,.. mang đến cho người đọc những nội dung chính xác và có giá trị thiết thực.

Xác định bài toán :

- **Input** : Nội dung tóm tắt (abstract) và tiêu đề (title) của một bài báo mà người dùng đọc.
- **Output** : Danh sách top các tiêu đề của bài báo khuyến nghi dành cho người dùng

2. BỘ DỮ LIỆU

2.1. Tạo bộ dữ liệu

Để xây dựng một hệ khuyến nghị cho bài toán khuyến nghị tin tức, chúng em đã tiến hành xây dựng bộ dữ liệu Vnexpress News. Bộ dữ liệu được thu thập từ trang vnexpress.net từ đầu tháng 10 đến tháng 12 từ 9 chủ đề : giáo dục, đời sống, khoa học, giải trí, thể thao, sức khoẻ, kinh doanh, thời sự, số hoá. Kết quả thu được 13531 bài báo bao gồm đường dẫn url và id của từng bài báo. Sau đó chúng em tiến hành thu thập thông tin của các bài báo báo và ID của những người dùng (users) đã bình luận trên bài báo đó thu được 6315 bài báo và 24152 người dung đã bình luận . Thông tin của mỗi bào báo bao gồm ID của bài báo (NewsID), tiêu đề của bài báo (NewsTitle), nội dung tóm tắt của bài báo (NewsAbstract) và chủ đề của bài báo (Topic). Về thông tin của người bình luận, chúng em chỉ thu thập ID của người bình luận về bài báo. Mỗi dòng trong cột UsersID là danh sách của từng người dung bình luận được ngăn cách với nhau

bởi khoảng trắng, ví dụ : 1043446290 1070564564 1058227852 1002629270. Bảng dưới đây mô tả thông tin của bộ dữ liệu:

Thông tin	**Nội dung**
Tên bộ dữ liệu	**Vnexpress News**
Nguồn thu thập	https://www.vnexpress.net
Số dòng	6315
Số thuộc tính	5
Thông tin các thuộc tính	- NewsID : Số định danh của bài báo - NewsTitle : Tiêu đề của bài báo - NewsAbstract : Nội dung tóm tắt của bài báo - UserID : danh sách id của những user bình luận trong bài báo - Topic : chủ đề của bài báo

Bảng 1.Thông tin về bộ dữ liệu

NewsID	NewsTitle	NewsAbstract	UserID	Topic
4404506	Người đàn ông 'nghiện' chế tạo máy bay mô hình...	Gần 100 chiếc máy bay mô hình gắn động cơ, tro...	1043446290 1070564564 1058227852 1002629270 10...	Đời sống
4404938	Lời khuyên vàng khi chọn rèm cửa - VnExpress Đ...	Rèm cửa tưởng là thứ rất nhỏ trong tổng thể, n...	1041608906	Đời sống
4404916	Truyện tranh siêu nhân có giá 2,6 triệu USD - ...	Quyển truyện tranh đời đầu về Siêu nhân, xuất ...	1074373109 1005291929 1005634914 1046496997	Đời sống
4403372	Căn hộ chia không gian bằng cánh tủ - VnExpres...	Nhờ hệ tủ đa năng, nữ gia chủ hay đón khách tớ...	1057748718 1061840527 1014687651 1003052895 10...	Đời sống
4404061	10 phong cách thiết kế nhà nổi bật năm 2021 - ...	10 xu hướng về cách bố trí không gian, lựa chọ...	1056461424 1065962404 1071937430 1062696358 10...	Đời sống
...	...	...	...	...
4401427	Cứu cô gái 18 tuổi bị đột quỵ - VnExpress Sức ...	Cô gái 18 tuổi ở Hải Phòng, cảm thấy đau đầu, ...	1074268383 1028835601 1059803421 1049576733 10...	Sức khỏe
4400786	Covid-19 Hà Nội: Test nhanh dương tính, nhiều ...	Người phụ nữ 70 tuổi, ở đường Lê Thanh Nghị, s...	1002942675 1070849195 1005001771 1054575285 10...	Sức khỏe
4400930	TP HCM kêu gọi nhà thuốc tư nhân chống dịch Co...	Sở Y tế TP HCM ngày 9/12 kêu gọi các nhà thuốc...	1012746042 1072119819 1003023118 1002657154 10...	Sức khỏe
4401344	Bệnh xá 30 năm tuổi trên đảo Trường Sa - VnExp...	Từ tổ quân y 3 người trên vùng cát sỏi san hô ...	1002679291 1069731719	Sức khỏe
4400290	Cách tập thở, tăng thể lực cho F0 tại nhà - Vn...	F0 tại nhà có thể thực hiện các bài tập thở, t...	1033920745	Sức khỏe

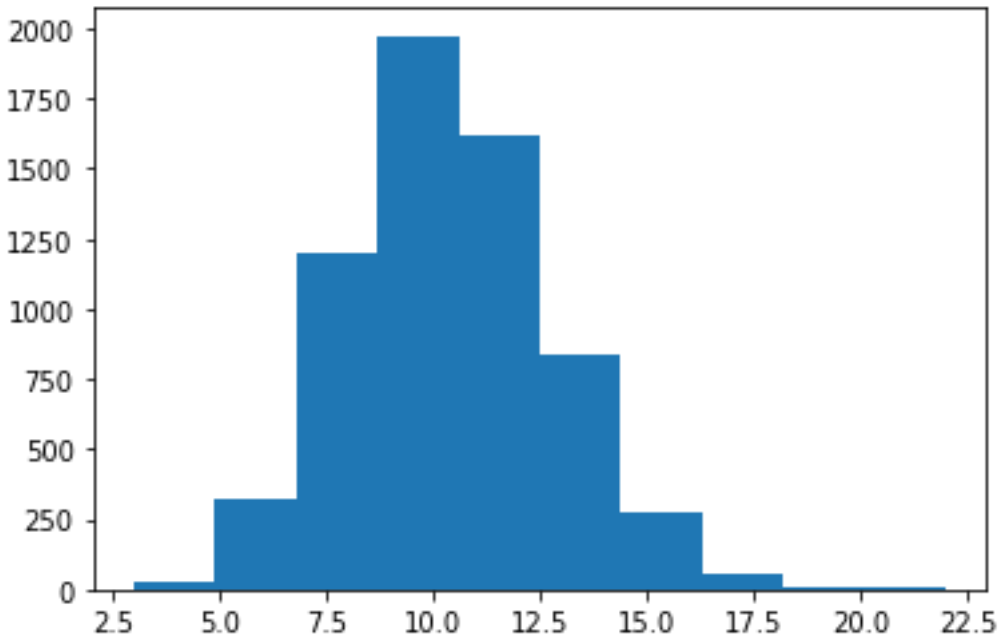
Hình 1. Bộ dữ liệu

Về cách thức thu thập bộ dữ liệu, chúng em sử dụng hai công cụ để thu thập là Scrapy kết hợp với Splash. Scrapy Framwork - một framwork mạnh về thu thập dữ liệu, lợi thế của scrapy ở việc nó hỗ trợ sẵn các hàm thư viện thuận tiện cho việc rút trích thông tin từ các trang web. Nhưng Scrapy có hạn chế là không thu thập được dữ liệu

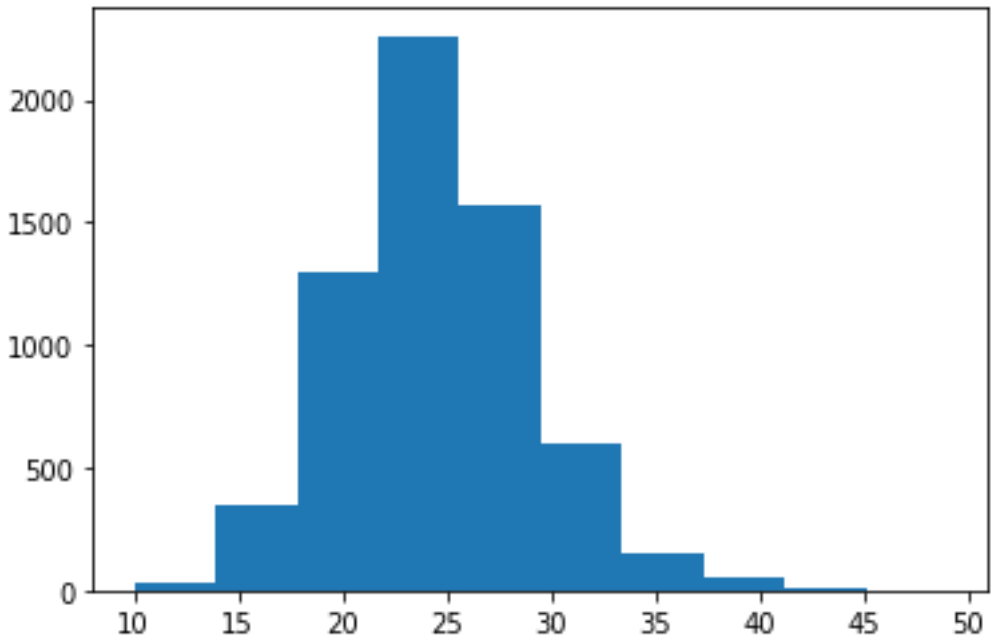
được viết bằng javascript. Chính vì vậy mà chúng em quyết định sử dụng Splash để render trang wed sau đó gửi dữ liệu về để Scrapy tiên hành thu thập dữ liệu.

2.2. Phân tích bộ dữ liệu.

Bộ dữ liệu bao gồm thông tin về 6315 bài bào và 24152 người dùng đã bình luận. Thống kê về số lượng từ ngữ của tiêu đề và nội dung tóm tắt của bài báo được trình bày trong hình 2 và hình 3 dưới đây.



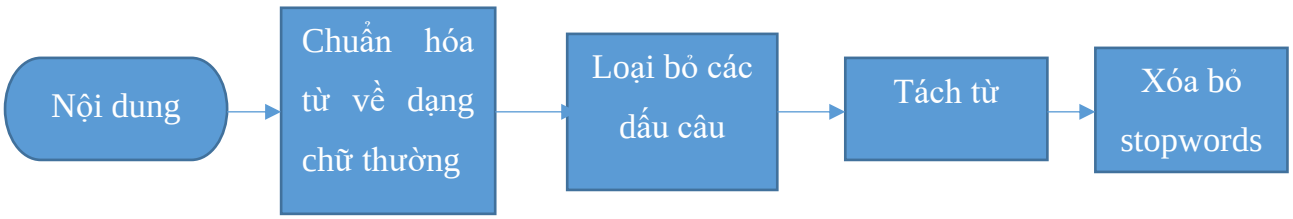
Hình 2 Thống kê độ dài tiêu đề bài báo



Hình 3. Thống kê về độ dài của nội dung tóm tắt của bài báo

Nhìn vào hai hình trên ta có thể thấy, có thể thấy các bài báo có tiêu đề trong khoảng đến 12 từ chiếm đa số bộ dữ liệu và nội dung tóm tắt của bài báo có độ dài từ 20 đến 30 từ chiếm đa số bộ dữ liệu.

2.3. Tiền xử lý dữ liệu



Hình 4. Quy trình tiền xử lý dữ liệu

- **Chuẩn hoá từ về dạng chữ thường**

Vì thứ quan trọng nhất là nội dung của tiêu đề, nội dung của tóm tắt, không quan trọng chữ hoa hay chữ thường, nên việc chuyển chữ hoa về chữ thường sẽ quy về 1 chuẩn chung và tăng độ chính xác cho mô hình.

- **Loại bỏ các dấu câu, các kí tự đặc biệt**

Các dấu câu, các kí tự đặc không có ý nghĩa về mặt nội dung, thâm chí nó còn khiến gây nhiễu cho dư liệu khi tiến hành embedding từ do đó nhóm chúng em sẽ loại bỏ hết các dấu câu và ký tự đặc biệt.

- **Tách từ**

Tách từ trong tiếng Việt không chỉ đơn giản là tách bởi khoảng trắng. Tiếng Việt rất đa dạng, gồm các từ đơn, từ ghép. Ví dụ, từ "học sinh" là 1 từ ghép, nếu tách từ không đúng, kết quả cho ra 2 từ "học" và "sinh", sẽ làm thay đổi nghĩa của từ "học sinh" ban đầu. Để tách từ chính xác, chúng em sử dụng thư viện pyvi để tách từ.

- **Xóa bỏ Stopwords**

Stopword thường là các từ nối như: của, để, nọ, là, có, được, những, mà... Đây thường là những từ giới từ, trợ từ có tần suất xuất hiện tương đối cao trong một văn bản thông thường, tuy nhiên không mang nhiều ý nghĩa. Vì vậy chúng em cũng loại bỏ các từ này bằng việc sử dụng bộ stopwords w2v dành cho Tiếng Việt.

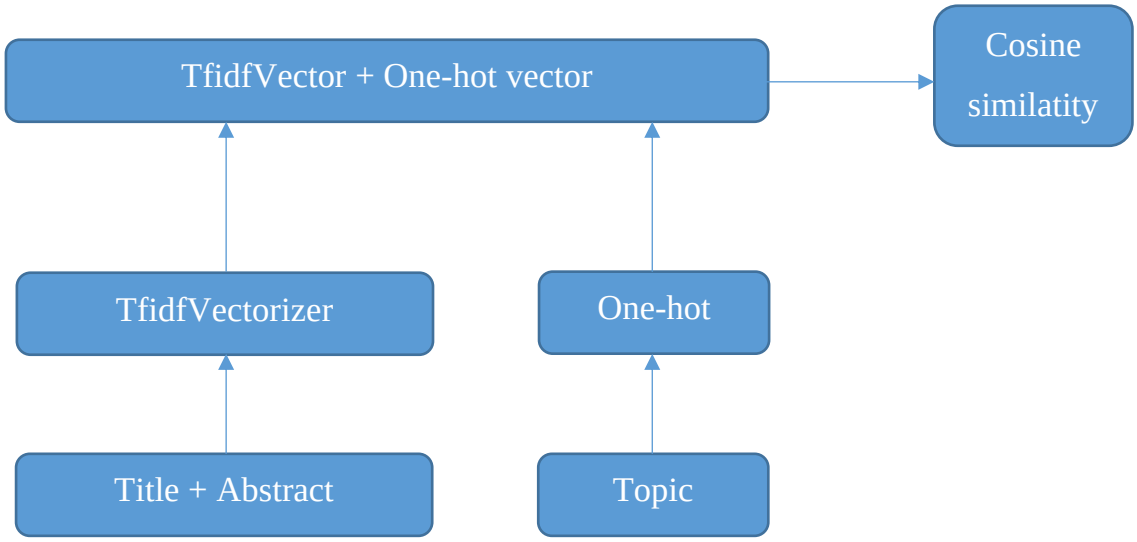
3. PHƯƠNG PHÁP

Trong đồ án này, chúng em sử dụng phương pháp lọc nội dung(Content-based filtering) với 3 phương pháp embedding từ là : TF-IDF, Word2vec, Fasttext và sử dụng thêm độ tương tự jaccard để tính toán mức độ tương tự của các bài báo dựa vào tiêu đề và nội dung tóm tắt của bài báo.

3.1. TF-IDF

TF-IDF là một phương pháp embedding từ đơn giản. Trọng số TF-IDF thể hiện mức độ quan trọng của một từ trong văn bản. TF-IDF chuyển đổi dữ liệu từ dạng văn bản sang dạng không gian vector được sử dụng nhiều trong các bài toán xử lý ngôn ngữ

tự nhiên. Dưới đây là quy trình mà nhóm chúng em sử dụng TF-IDF trong bài toán khuyến nghị tin tức.



Hình 5. Quy trình áp dụng TF_IDF

Đầu tiên, chúng em lấy tiêu đề và nội dung tóm tắt đã qua tiền xử lý ghép lại với nhau rồi sử dụng TfidfVectorizer từ thư viện sklearn để tính toán trọng số TF-IDF cho từng từ rồi chuyển từng dữ liệu văn bản của từng bài báo sang dạng vector. Đối với chủ đề của từng bài báo, chúng em sử dụng phương pháp one-hot encoder để chuyển đổi chủ đề của từng bài báo sang dạng one-hot vector. Tiếp đến, đối với mỗi bài báo, chúng em nối hai vector TF-IDF và One-hot lại thành một vector duy nhất đại diện cho mỗi bài báo. Cuối cùng, chúng em sử dụng Cosine similatity để tính toán mức độ tương đồ của các vector với nhau.

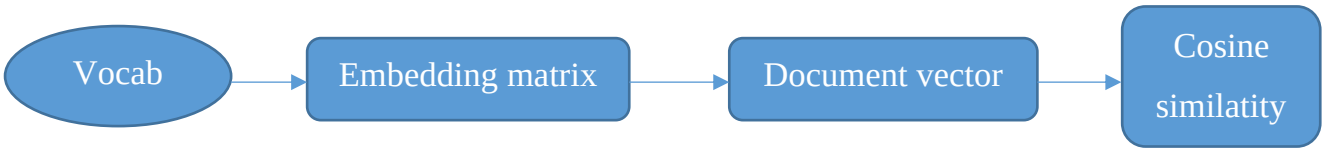
3.2. Word2vec

Word2Vec (Word to vector) là phương pháp được Google công bố vào năm 2013. Word2vec là một mô hình máy học không có giám sát được huấn luyện dựa trên một lượng lớn dữ liệu. Bằng cách thực hiện một tác vụ giả là dự đoán từ nào sẽ xuất trong một phạm vi các từ xung quanh (phạm vi này được gọi là cửa sổ ngữ cảnh – context

window). Về mặt toán học, Word2vec ánh xạ một tập từ vựng sang một không gian vectơ. 2 mô hình áp dụng Word2vec :

- Continuous Bag-of-Word Model : Phương pháp này lấy input là ngữ cảnh của mỗi từ và cố gắng dự đoán từ gần nhất tương ứng với ngữ cảnh đó.
- Skip-gram : Ngược lại với mô hình CBOW, mô hình Skip-gram lấy input là một từ và dự đoán những từ có liên quan.

Mô hình Pre-trained Word2vec cho tiếng Việt mà nhóm em sử dụng được huấn luyện dựa trên dữ liệu các bài báo từ baomoi.com của tác giả Vũ Xuân Sơn. Hình dưới đây là quy trình áp dụng Word2vec vào bài toán khuyến nghị tin tức.



Hình 6. Quy trình áp dụng Word2vec

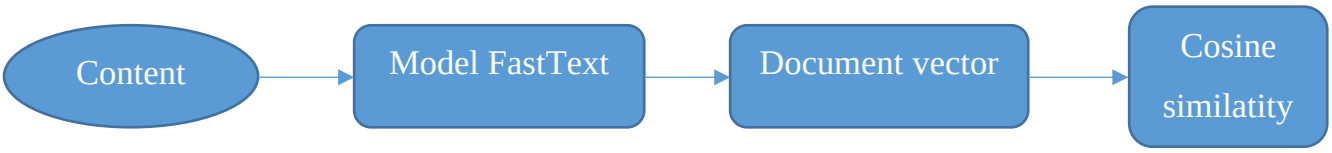
Đầu tiên, chúng khởi tạo bộ từ vựng cho mô hình Word2vec bằng việc tách từ từ tiêu đề và nội dung của từng bài báo, kết quả thu được một tập từ vựng (vocab) bao gồm 8032 từ. Từ bộ từ vựng đã được tạo, nhóm chúng em sử dụng mô hình pre-trained Word2vec để lấy ra những vector từ vựng để tạo thành mạ trận nhúng từ (Embedding matrix). Tiếp đến, chúng em ánh xạ bộ dữ liệu sang không gian vector với mỗi vector có số chiều bằng 300 bằng cách sử dụng Embedding matrix. Cuối cùng tương tự như phương pháp TF-IDF, chúng em cũng tính toán mức độ tương tự bằng Cosine Similarity.

3.3. FastText

FastText là thư viện mã nguồn mở do **Facebook** tạo ra năm 2016, để khắc phục điểm hạn chế của word2vec là chỉ lấy được những vector của từ có trong tập từ vựng, fasttext hỗ trợ việc huấn luyện phép nhúng từ và phân loại văn bản. FastText tương tự như Word2vec nhưng khác ở chỗ FastText huấn luyện dựa trên các n-gram. Ví dụ từ “lạnh” sẽ được tách thành [“<lạnh>”,“lạ”, “lạn”, “ạnh”, “nh>”]. Vector của từ

"lạnh" sẽ là tổng của các vector ["<lạnh>","lạ", "lạn", "ạnh", "nh>"] nhân với vector của từ "lạnh". Việc huấn luyện n-gram đã khắc phục điểm hạn chế của word2vec, FastText có thể lấy được vector của những từ không xuất hiện trong bộ từ vựng bằng việc ghép nối các n-gram lại với nhau để tạo thành những từ mới.

Về quy trình ứng dụng FastText vào bài toán khuyến nghị tin tức cũng tương tự như Word2vec tuy nhiên thư viện của FastText có hỗ trợ việc lấy vector cho cả một văn bản nên bước tạo tập từ vựng và ma trận nhúng từ được bỏ qua. Hình dưới là quy trình áp dụng FastText vào bài toán khuyến nghị tin tức.



Hình 6. Quy trình áp dụng FastText

4. ĐÁNH GIÁ

Để đánh giá độ hiệu quả của hệ khuyến nghị, chúng em sử dụng 3 độ đo để đánh giá đó Precision, Recall cho top 10 tin tức mà hệ khuyến nghị cho users. Chính vì vậy mà chúng em đã lọc ra những người dùng đánh giá từ 10 tin bài trở lên làm tập kiểm thử với 500 dòng dữ liệu.

	Precision@10	**Recall@10**
TF-IDF	46.26	46.87
Word2vec	55.68	46.21
FastText	45.54	46.31

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong báo cáo này, chúng em đã trình bày về quá trình thu thu bộ dữ Vnexpress News cho bài toán khuyến nghị tin tức. Nhóm chúng em đã tiến hành thử nghiệm bộ dữ liệu trên các phương pháp embedding khác nhau. Kết quả thu được cho thấy phương pháp Word2vec embedding đem lại hiệu quả tốt nhất. Tuy nhiện, hiệu quả của các phương pháp đem lại không cao.

Trong tương lại, để có thể cải thiện được độ hiệu quả của hệ khuyến nghị, chúng em hy vọng có thể mở rộng thêm bộ dữ liệu nghiệm thêm các phương pháp embedding mới đồng thời thu thập thêm dữ liệu.

6. TÀI LIỆU THAM KHẢO

1. https://towardsdatascience.com/calculating-document-similarities-using-bert-and-other-models-b2c1a29c9630
2. Hung, Phan Duy. "Application of Customized Term Frequency-Inverse Document Frequency for Vietnamese Document Classification in Place of Lemmatization." *International Conference on Intelligent Computing & Optimization*. Springer, Cham, 2020.
3. A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, FastText.zip: Compressing text classification models

Source code :

https://colab.research.google.com/drive/17C5F3kXE45zqQFj4IjttAUvlER3ZfHb6?usp=sharing

Link dataset :

https://drive.google.com/file/d/1CS_v0guQAD_IzsUhbf0x1L6xZUxqfLV/view?usp=sharing