Jobs Recommendation System

Phạm Đức Thể[1,2*], Trần Thành Luân[1,2], Trần Nhật Nam[1,2] and Nguyễn Văn Kiệt[1,2]

[1]Khoa Học & Kĩ Thuật Thông Tin, Trường Đại Học Công Nghệ Thông Tin, TP.HCM, Việt Nam.
[2]Đại Học Quốc Gia TP.HCM, Việt Nam.

*Corresponding author(s). E-mail(s): 19522253@gm.uit.edu.vn;
Contributing authors: 19521810@gm.uit.edu.vn;
19521872@gm.uit.edu.vn; kietnv@uit.edu.vn;

Tóm tắt nội dung

Ngày nay, với sự phát triển và tăng nhanh của dữ liệu trên các trang web, việc tìm kiếm một thông tin hay một sản phẩm mong muốn trên các nền tảng này là việc không hề dễ dàng, nó làm cho người dùng mất nhiều thời gian để tìm kiếm các thông tin, sản phẩm phù hợp với nhu cầu của họ. Lĩnh vực việc làm cũng không nằm ngoài trong số đó, nhu cầu tuyển dụng, tìm việc làm ngày càng lớn và việc tìm kiếm việc làm thông qua các trang web là một điều tất yếu. Nhưng để cho các ứng viên có được một vị trí công việc mà họ mong muốn là rất khó khăn. Việc xây dựng được một *hệ khuyến nghị công việc (jobs recommendation system)* tốt giúp cho ứng cử viên và nhà tuyển dụng dễ dàng hơn trong việc kết nối với nhau. Từ đó, giúp cho việc tìm kiếm thông tin công việc trở nên dễ dàng và tối ưu hóa trải nghiệm người dùng trên các trang web tìm kiếm việc làm trực tuyến. Trong báo cáo này, chúng tôi thực hiện xây dựng bộ dữ liệu *Vietnamese Jobs Dataset*. Chúng tôi tiến hành xử lý dữ liệu để tạo ra các tập dữ liệu công việc và ứng viên phù hợp với phương pháp *content-based filtering*. Sau đó, chúng tôi so sánh kết quả đạt được thông qua Accuracy. Kết quả tốt nhất mà chúng tôi đạt được là 90% với phương pháp BERT trên thuộc tính Industry-Industry.

Keywords: Jobs Recommendation System, Vietnamese Jobs Dataset, Content-based filtering, TF-IDF, BERT, Word2Vec.

1 Giới Thiệu

Hệ thống Khuyến nghị (Recommendation Systems) được ứng dụng rất thành công trong dự đoán sở thích/thói quen của người dùng dựa vào sở thích/thói quen của họ trong quá khứ. Hệ thống khuyến nghị đang được ứng dụng trong rất nhiều lĩnh vực khác nhau như thương mại điện tử (hỗ trợ bán hàng trực tuyến), giải trí (gợi ý phim ảnh, bài hát,...), giáo dục đào tạo (gợi ý nguồn tài nguyên học tập, nghiên cứu,...). Chính vì khả năng ứng dụng rộng rãi của nó, hệ thống khuyến nghị mở ra nhiều tiềm năng trong nghiên cứu cũng như trong xây dựng các hệ thống thực tế, đặc biệt là các hệ thống hỗ trợ người dùng ra quyết định.

Có hai thực thể chính trong một recommendation system là *user* và *item*. *User* là *người dùng*; *item* là *sản phẩm*, ví dụ như các bộ phim, bài hát, cuốn sách, clip, công việc hoặc cũng có thể là các người dùng khác trong bài toán gợi ý kết bạn. Mục đích chính của các recommender system là dự đoán mức độ quan tâm của một người dùng tới một sản phẩm nào đó, qua đó có chiến lược recommendation phù hợp. Các recommendation system thường được chia thành hai nhóm lớn[1]:

1. *Content-based system*: khuyến nghị dựa trên đặc tính của item bằng cách sử dụng các đặt trưng mô tả của item để khuyến nghị các item khác tương tự như những gì người dùng yêu thích, dựa trên các hành động trước đây của họ hoặc phản hồi rõ rang trong hệ thống.
2. *Collaborative filtering*: hệ thống khuyến nghị các sản phẩm dựa trên sự tương đồng giữa các người dùng hoặc sản phẩm. Có thể hiểu rằng ở nhóm này một sản phẩm được khuyến nghị tới một người dùng dựa trên những người dùng có hành vi tương tự.

Với sự phát triển mạnh mẽ của kinh tế và internet, nhu cầu tuyển dụng cũng như tìm việc trực tuyến của con người ngày càng tăng. Vấn đề đặt ra trên các trang web tìm kiếm việc làm trực tuyến hiện nay là việc tìm kiếm thông tin công việc mong muốn ngày càng trở nên khó khăn và mất nhiều thời gian. Để giúp người dùng tìm thấy công việc phù hợp với kỹ năng, mức lương và vị trí công việc của họ, chúng tôi đã xây dựng một *hệ khuyến nghị việc làm (jobs recommendation system)* bằng cách sử dụng các phương pháp khuyến nghị khác nhau. Từ đó, giúp cho việc tìm kiếm thông tin công việc trở nên dễ dàng và tối ưu hóa trải nghiệm người dùng trên các trang web tìm kiếm việc làm trực tuyến.

Trong báo cáo này, trước tiên chúng tôi trình bày về quy trình thu thập và tạo ra bộ dữ liệu *Vietnamese Jobs Dataset* để sử dụng cho bài toán *Jobs Recommendation System*. Ở Phần 2 chúng tôi sẽ giới thiệu các công trình liên quan. Chúng tôi tiến hành xử lý dữ liệu trên hai bộ dữ liệu công việc và ứng viên (Phần 3). Tiếp theo, hướng tiếp cận bài toán được mô tả chi tiết trong Phần 4. Đối với Phần 5, chúng tôi tiến hành thực nghiệm và phân tích kết quả của các phương pháp khuyến nghị trên tập dữ liệu hiện có. Cuối cùng, chúng tôi rút ra kết luận ở Phần 6.

2 Công Trình Liên Quan

2.1 Khuyến nghị tự động

Ba công nghệ chính được sử dụng trong các hệ thống khuyến nghị truyền thống: khuyến nghị dựa trên nội dung[2], khuyến nghị dựa trên lọc cộng tác[3] và khuyến nghị dựa trên lọc kết hợp[4]. Khuyến nghị dựa trên nội dung, phương pháp này xem xét các hồ sơ thông tin trước đó của người dùng để khuyến nghị nội dung tương tự cho người dùng. Khuyến nghị dựa trên lọc cộng tác nhằm mục đích tìm những người dùng tương tự khác để đưa ra khuyến nghị cho người dùng mục tiêu. Có hai loại khuyến nghị dựa trên lọc cộng tác: khuyến nghị dựa trên *item* và khuyến nghị dựa trên *user*. Khuyến nghị dựa trên item sẽ tính toán mức độ tương đồng giữa các item liên quan đến người dùng và các item khác để khuyến nghị. Khuyến nghị dựa trên user sẽ tính toán sự giống nhau giữa các người dùng để đưa ra khuyến nghị. Các hệ thống khuyến nghị dựa trên lọc kết hợp sẽ sử dụng hai hoặc nhiều loại kỹ thuật khuyến nghị khác nhau để tạo ra các khuyến nghị tốt hơn.

2.2 Phương pháp phân cụm văn bản

Phân cụm văn bản là quá trình nhóm các tập văn bản có các tính chất tương tự nhau trong một tập dữ liệu vào các cụm sao cho các văn bản trong cùng một cụm có các tính chất tương đồng nhau. Hai phương pháp chính là:

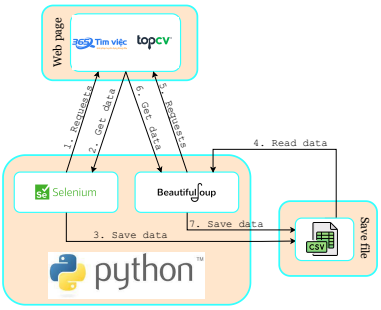
- **Word2Vec**[5]: Word2vec là một nhóm các mô hình có liên quan được sử dụng để tạo word embeddings. Các mô hình này là two-layer neural networks, được huấn luyện để tái tạo lại ngữ cảnh ngôn ngữ của các từ. Word2vec lấy đầu vào là một kho văn bản lớn và tạo ra một không gian vector, thường có vài trăm chiều, với mỗi từ duy nhất trong kho văn bản được gán một vector tương ứng trong không gian. Hai từ được coi là càng tương tự nhau nếu khoảng cách giữa các vector tương ứng của chúng là càng thấp.
- **K-mean**[6]: là một thuật toán phổ biến và đơn giản để thực hiện. Đầu tiên chúng sẽ tạo ra các điểm trung tâm ngẫu nhiên. Sau đó gán mỗi điểm trong tập dữ liệu vào trung tâm gần nó nhất. Tiếp theo, chúng sẽ cập nhật lại trung tâm và tiếp tục lặp lại các bước đã kể trên. Điều kiện dừng của thuật toán là khi các trung tâm không thay đổi trong 2 vòng lặp kế tiếp nhau. Tuy nhiên, việc đạt được 1 kết quả hoàn hảo là rất khó và rất tốn thời gian, vậy nên thường người ta sẽ cho dừng thuật toán khi đạt được 1 kết quả gần đúng và chấp nhận được.

3 Bộ Dữ Liệu

3.1 Thu thập dữ liệu

Bộ dữ liệu sử dụng trong báo cáo này có tên là *Vietnamese Jobs Dataset* được chúng tôi thu thập từ các trang web tìm việc trực tuyến. Sử dụng ngôn ngữ lập trình Python kết hợp với hai framework được hỗ trợ mạnh mẽ cho việc cào

dữ liệu là Selenium và BeautifulSoup để thu thập thông tin về công việc và ứng viên. Chi tiết về quy trình thu thập dữ liệu được trình bày ở Hình 1. Bộ dữ liệu được thu thập gồm hai bộ là: bộ dữ liệu *công việc (jobs)* và bộ dữ liệu *ứng viên (users)*. Bộ dữ liệu công việc được chúng tôi thu tập trên hai trang web topcv.vn và timviec365.vn gồm 14,634 công việc với 18 thuộc tính. Bộ dữ ứng viên được chúng tôi thu tập trên trang web timviec365.vn gồm 14,868 ứng viên với 17 thuộc tính. Thông tin chi tiết về các thuộc tính của bộ dữ liệu được thể hiện trong Bảng 1, những thuộc tính có màu giống nhau được xem như có ý nghĩa tương đồng với nhau ở hai bộ dữ liệu.



Hình 1 Quy trình crawl data.

	Bộ dữ liệu công việc		Bộ dữ liệu ứng viên	
Index	**Thuộc tính**	**Ý nghĩa**	**Thuộc tính**	**Ý nghĩa**
0	Industry	Ngành nghề	Industry	Ngành nghề
1	Job Title	Tiêu đề tuyển dụng	Desired Job	Công việc mong muốn
2	Gender	Yêu cầu giới tính	Gender	Giới tính ứng viên
3	Salary	Mức lương	Desired Salary	Mức lương mong muốn
4	Years of Experience	Kinh nghiệm	Work Experience	Kinh nghiệm làm việc
5	Job Address	Địa chỉ làm việc	Workplace Desired	Nơi làm việc mong muốn
6	Job Description	Mô tả công việc	Current Place of Residence	Nơi ở hiện tại
7	Job Requirements	Yêu cầu công viêc	Province	Tên Tỉnh/TP
8	Benefits	Quyền lợi	District	Tên Quận/Huyện
9	Company Address	Địa chỉ công ty	URL User	URL của ứng viên
10	Job Type	Hình thức làm việc	User Name	Tên ứng viên
11	Name Company	Tên công ty	Marriage	Tình trạng hôn nhân
12	Number Cadidate	Số lượng tuyển	Date of Birth	Ngày sinh
13	Career Level	Cấp bậc	Target	Mục tiêu nghề nghiệp
14	Company Size	Quy mô công ty	Skills	Kĩ năng
15	Company Overview	Tổng quan công ty	Degree	Bằng cấp
16	URL Job	URL của công việc	UserID	ID của ứng viên
17	Submission Deadline	Hạn nộp hồ sơ		

Bảng 1 Thông tin chi tiết của các thuộc tính.

Bộ dữ liệu được thu thập từ tháng 11/2022 được sử dụng riêng cho mục đích học tập và nghiên cứu môn học Phân Tích Dữ Liệu Truyền Thông Xã Hội. Hình 2 thể hiện ví dụ về các điểm dữ liệu thô trong bộ dữ liệu.

	URL User	UserID	User Name	Industry	Desired Job	Workplace Desired	Current Place of Residence	Province	District	Desired Salary	Gender	Marriage	Date of Birth	Target	Skills	Degree	Work Experience
0	https://timviec365.vn /ung-vien/nguyen-quan-uv9...	964496	Nguyễn Quân	[Vận tải - Lái xe, <...	Nhân viên lái xe bằng B2	['Hà Nội']	Quận Ba Đình, Hà Nội	Hà Nội	Quận Ba Đình	Thỏa thuận	Nam	Đã có gia đình	03/01/1986	Với kinh nghiệm 10 năm trong lĩnh vực lái xe t...	- Bằng lái xe hạng C, D, E và có kinh nghiệm 5...	Đơn vị đào tạo: Trường trung cấp nghề	Trên 5 năm
1	https://timviec365.vn /ung-vien/nguyen-thi-hong...	970215	Nguyễn Thị Hồng Nhung	[Sinh viên làm thêm]	nhân viên tư vấn bán hàng	['Hà Nội']	Nguyên Xá, Phường Minh Khai, QUận Bắc Từ Liêm,...	Hà Nội	Quận Bắc Từ Liêm	Thỏa thuận	NaN	NaN	NaN	NaN	NaN	NaN	Chưa có kinh nghiệm
2	https://timviec365.vn /ung-vien/le-hong-phu-uv9...	978880	Lê Hồng Phú	[Vận tải - Lái xe]	Nhân viên lái xe bằng B2	['Hồ Chí Minh']	Quận 10, hồ chí minh	Hồ Chí Minh	Quận 10	Thỏa thuận	Nam	Đã có gia đình	11/11/1994	NaN	NaN	NaN	1 năm
3	https://timviec365.vn /ung-vien/nguyen-chau-uv9...	986206	Nguyễn Châu	[Vận tải - Lái xe, <...	Nhân viên lái xe bằng D	['Bình Dương']	Thành phố Thuận An, Bình Dương	Bình Dương	Thành phố Thuận An	Thỏa thuận	Nam	Đã có gia đình	03/01/1978	Gắn bó lâu dài với công việc cũng như doanh ng...	Có chí tiến thủ đối với công việc\nCó khả năng...	Đơn vị đào tạo: Trường trung cấp nghề	Trên 10 năm
4	https://timviec365.vn /ung-vien/dang-tran-cong-...	985789	Đặng Trần Công	[Marketing - PR, <sp...	HSE ENGINEER	['Khánh Hòa']	Hòa Phú	Phú Yên	Huyện Tây Hoà	Thỏa thuận	NaN	NaN	19/01/1996	NaN	NaN	NaN	1 năm

(a) 5 điểm dữ liệu thô đầu tiên của bộ dữ liệu ứng viên.

	URL Job	Job Title	Name Company	Company Overview	Company Size	Company Address	Job Description	Job Requirements	Benefits	Job Address	Job Type	Gender	Number Cadidate	Career Level	Years of Experience	Salary	Submission Deadline	Industry
0	https://www.topcv.vn /viec-lam/sale-admin-websi...	Sale Admin Website	Công ty Cổ phần Belie	DHC Việt Nam trực thuộc Công ty CP Belie – Nhà...	100-499 nhân viên	HN: Tầng 9, Tòa IDMC Mỹ Đình, 18 Tôn Thất Thuy...	Thường xuyên cập nhật các thay đổi liên quan t...	1. Học vấn, kiến thức và trình độ chuyên môn:T...	Được xét duyệt đánh giá định kỳ 2 lần/năm.Làm ...	['- Hà Nội: 18 Tôn Thất Thuyết, Nam Từ Liêm']	Toàn thời gian	Không yêu cầu	1 người	Nhân viên	1 năm	9-11 triệu	05/01/2023	['Hành chính / Văn phòng', 'Kinh doanh / Bán h...
1	https://www.topcv.vn /viec-lam/thuc-tap-sinh-la...	Thực Tập Sinh Lập Trình (No-Code, Low-Code Pla...	Công Ty Cổ Phần Global Online Branding	Thành viên thuộc Tập đoàn G-Group:Công ty Cổ p...	100-499 nhân viên	Lầu 6, 235 Lý Thường Kiệt ,Phường 6 ,Quận Tân ...	Phối hợp với bộ phận lập trình No-code, Low-co...	Sinh viên năm 3, năm 4 hoặc sinh viên cao học ...	Làm việc trong môi trường văn hóa Nhật Bản chu...	['- Hồ Chí Minh: 235 Lý Thường Kiệt, Phường 6,...	Toàn thời gian	Không yêu cầu	2 người	Thực tập sinh	Không yêu cầu kinh nghiệm	Tới 1 triệu	01/01/2023	['Công nghệ thông tin', 'IT phần mềm', 'Việc l...
2	https://www.topcv.vn /viec-lam/hr-business-part...	HR Business Partner	Công Ty Melody Logistics	Công Ty TNHH Melody Logistics được biết đến là...	100-499 nhân viên	Số 01 Nguyễn Văn Đậu Phường 5, Quận Phú Nhuận,...	Xây dựng và thực hiện chiến lược nhân sự phù h...	Tốt nghiệp đại học các chuyên ngành liên quan ...	Lương thỏa thuận theo năng lực;Làm việc trong ...	['- Hồ Chí Minh: Phú Nhuận']	Toàn thời gian	Không yêu cầu	1 người	Trưởng/Phó phòng	Trên 5 năm	Thoả thuận	28/12/2022	['Nhân sự', 'Quản lý điều hành', 'Xuất nhập kh...
3	https://www.topcv.vn /viec-lam/general-manager/...	General Manager	CÔNG TY TNHH DV HOLDINGS	Khách sạn Danaciti By Dabi là khách sạn 4 sao,...	25-99 nhân viên	80 Đông Kinh Nghĩa Thục, Sơn Trà, Đà Nẵng	1. Lập và triển khai kế hoạch kinh doanh-Định...	- Tốt nghiệp đại học trở lên chuyên ngành quản...	- Tiền lương, thưởng và các khoản trợ cấp khác...	['- Đà Nẵng: 80 Đông Kinh Nghĩa Thục, Sơn Trà']	Toàn thời gian	Không yêu cầu	1 người	Giám đốc	3 năm	Thoả thuận	31/12/2022	['Khách sạn / Nhà hàng', 'Thực phẩm / Đồ uống'...
4	https://www.topcv.vn /viec-lam/le-tan-gymasaste...	Lễ Tân Gymasaster Quận 12	CÔNG TY TNHH TM XNK NGUỒN SỐNG VIỆT	Công ty TNHH TM XNK Nguồn Sống Việt cung cấ...	100-499 nhân viên	235 đường 9A, khu dân cư Trung Sơn, huyện Bi...	- Trực quầy Lễ tân- Tư vấn bán hàng gói member...	- Giới tính: Ưu tiên Nữ từ 22 - 35 tuổi- Kinh ...	Thu nhập = Lương cơ bản + Thưởng KPIs + Hoa hồ...	['- Hồ Chí Minh: Gymaster 99 Phan Văn Hớn, Qu...	Toàn thời gian	Không yêu cầu	2 người	Nhân viên	Không yêu cầu kinh nghiệm	6-12 triệu	06/01/2023	['Hành chính / Văn phòng', 'Dịch vụ khách hàng...

(b) 5 điểm dữ liệu thô đầu tiên của bộ dữ liệu công việc.

Hình 2 Ví dụ về các điểm dữ liệu thô.

3.2 Xử lý dữ liệu

Dữ liệu tốt đóng vai trò quan trọng trong việc tạo ra các mô hình dự báo có độ chính xác cao và tổng quát hóa. Để có được bộ dữ liệu tốt ứng với từng tác vụ thì ta có những cách xử lý dữ liệu khác nhau.

3.2.1 Bộ dữ liệu công việc

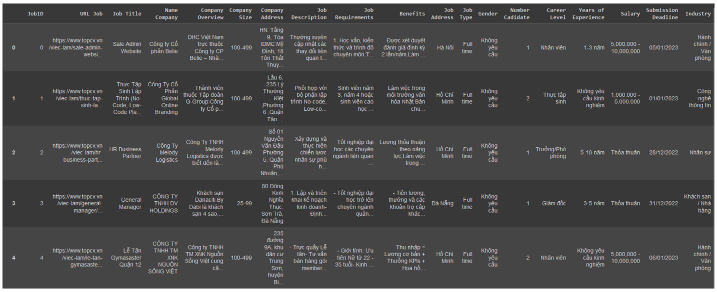
Bộ dữ liệu được xử lý qua hai phần chính: chuẩn hóa dữ liệu text và chuẩn hóa các thuộc tính. Ngoài ra, chúng tôi còn xử lý missing values, thêm và xóa các thuộc tính để có được bộ dữ liệu thích hợp.

- **Chuẩn hóa dữ liệu text:** Hầu hết các thuộc tính của bộ dữ liệu là text nên đều được xử lý qua bước này bằng cách xóa các kí tự đặt biệt và khoảng trắng dư thừa trong quá trình thu thập.
- **Chuẩn hóa các thuộc tính:** Trích xuất các giá trị số và gom nhóm lại thành một khoảng giá trị thích hợp đối với từng thuộc tính: *Company Size, Salary, Years of Experience*. Thuộc tính *Job Address*, chúng tôi loại bỏ đi địa chỉ cụ thể và chỉ dữ lại tên tỉnh/thành phố. Với thuộc tính *Number Cadidate*, chúng tôi trích xuất ra số và chuyển kiểu dữ liệu từ *object* sang *int64*. Thuộc tính *Industry* thường bao gồm nhiều ngành nghề/lĩnh vực khác nhau, chúng tôi chỉ lấy một ngành nghề/lĩnh vực đầu tiên. Chi tiết các thuộc tính chuẩn hóa và giá trị được trình bày tại Bảng 2.
- **Các xử lý khác:** Bộ dữ liệu có một số ít bị missing values, chúng tôi tiến hành xử lý bằng cách thay thế missing values bằng giá trị *"Đang cập nhật"*. Cuối cùng, chúng tôi thêm thuộc tính *JobID*, gán ID cho từng công việc để dễ dàng trong việc xây dựng mô hình và khuyến nghị công việc theo JobID.

Thuộc tính	Giá trị chuẩn hóa
Company Size	1-9; 10-24; 25-99; 100-499; 500-999; 1000-3000; 3000+.
Job Type	Part time; Full time; Remote; Others.
Salary	Thỏa thuận; 1,000,000 - 5,000,000; 5,000,000 - 10,000,000; 10,000,000 - 15,000,000; 15,000,000 - 20,000,000; 20,000,000 - 40,000,000; Trên 40,000,000.
Years of Experience	Không yêu cầu kinh nghiệm; Dưới 1 năm; 1-3 năm; 3-5 năm; 5-10 năm; Trên 10 năm.
Job Address	Hồ Chí Minh; Hà Nội; Đà Nẵng; Bình Dương;...
Number Cadidate	1;2;3;4;5;...
Industry	Nhân sự; Công nghệ thông tin; Văn phòng; Lái xe; ...

Bảng 2 Chuẩn hóa các thuộc tính của bộ dữ liệu công việc.

Hình 3 thể hiện ví dụ về các điểm dữ liệu đã được xử lý của bộ dữ liệu công việc.



	JobID	URL Job	Job Title	Name Company	Company Overview	Company Size	Company Address	Job Description	Job Requirements	Benefits	Job Address	Job Type	Gender	Number Cadidate	Career Level	Years of Experience	Salary	Submission Deadline	Industry
0	0	https://www.topcv.vn/viec-lam/sale-admin-websi...	Sale Admin Website	Công ty Cổ phần Belie	DHC Việt Nam trực thuộc Công ty CP Belie – Nhà...	100-499	HN: Tầng 9, Tòa IDMC Mỹ Đình, 18 Tôn Thất Thuy...	Thường xuyên cập nhật các thay đổi liên quan t...	1. Học vấn, kiến thức và trình độ chuyên môn:T...	Được xét duyệt đánh giá định kỳ 2 lần/năm.Làm ...	Hà Nội	Full time	Không yêu cầu	1	Nhân viên	1-3 năm	5,000,000 - 10,000,000	05/01/2023	Hành chính / Văn phòng
1	1	https://www.topcv.vn/viec-lam/thuc-tap-sinh-la...	Thực Tập Sinh Lập Trình (No-Code, Low-Code Pla...	Công Ty Cổ Phần Global Online Branding	Thành viên thuộc Tập đoàn G-Group:Công ty Cổ p...	100-499	Lầu 6, 235 Lý Thường Kiệt ,Phường 6 ,Quận Tân ...	Phối hợp với bộ phận lập trình No-code, Low-co...	Sinh viên năm 3, năm 4 hoặc sinh viên cao học ...	Làm việc trong môi trường văn hóa Nhật Bản chu...	Hồ Chí Minh	Full time	Không yêu cầu	2	Thực tập sinh	Không yêu cầu kinh nghiệm	1,000,000 - 5,000,000	01/01/2023	Công nghệ thông tin
2	2	https://www.topcv.vn/viec-lam/hr-business-part...	HR Business Partner	Công Ty Melody Logistics	Công Ty TNHH Melody Logistics được biết đến là...	100-499	Số 01 Nguyễn Văn Đậu Phường 5, Quận Phú Nhuận,...	Xây dựng và thực hiện chiến lược nhân sự phù h...	Tốt nghiệp đại học các chuyên ngành liên quan ...	Lương thỏa thuận theo năng lực;Làm việc trong ...	Hồ Chí Minh	Full time	Không yêu cầu	1	Trưởng/Phó phòng	5-10 năm	Thỏa thuận	28/12/2022	Nhân sự
3	3	https://www.topcv.vn/viec-lam/general-manager/...	General Manager	CÔNG TY TNHH DV HOLDINGS	Khách sạn Danaciti By Dabi là khách sạn 4 sao,...	25-99	80 Đông Kinh Nghĩa Thục, Sơn Trà, Đà Nẵng	1. Lập và triển khai kế hoạch kinh doanh-Định...	- Tốt nghiệp đại học trở lên chuyên ngành quản...	- Tiền lương, thưởng và các khoản trợ cấp khác...	Đà Nẵng	Full time	Không yêu cầu	1	Giám đốc	3-5 năm	Thỏa thuận	31/12/2022	Khách sạn / Nhà hàng
4	4	https://www.topcv.vn/viec-lam/le-tan-gymasaste...	Lễ Tân Gymasaster Quận 12	CÔNG TY TNHH TM XNK NGUỒN SỐNG VIỆT	Công ty TNHH TM XNK Nguồn Sống Việt cung cấ...	100-499	235 đường 9A, khu dân cư Trung Sơn, huyện Bi...	- Trực quầy Lễ tân- Tư vấn bán hàng gói member...	- Giới tính: Ưu tiên Nữ từ 22 - 35 tuổi- Kinh ...	Thu nhập = Lương cơ bản + Thưởng KPIs + Hoa hồ...	Hồ Chí Minh	Full time	Không yêu cầu	2	Nhân viên	Không yêu cầu kinh nghiệm	5,000,000 - 10,000,000	06/01/2023	Hành chính / Văn phòng

Hình 3 Bộ dữ liệu công việc sau khi đã xử lý.

3.2.2 Bộ dữ liệu ứng viên

Tương tự như bộ dữ liệu công việc, bộ dữ liệu ứng viên cũng được xử lý:

- **Chuẩn hóa dữ liệu text:** Xử lý tương tự như bộ dữ liệu công việc.
- **Chuẩn hóa các thuộc tính:** Trích xuất các giá trị số và gom nhóm lại thành một khoảng giá trị thích hợp đối với từng thuộc tính: *Desired Salary, Work Experience*. Thuộc tính *Industry* thường bao gồm nhiều ngành nghề/lĩnh vực khác nhau, chúng tôi chỉ lấy một ngành nghề/lĩnh vực đầu tiên. Chi tiết các thuộc tính chuẩn hóa và giá trị được trình bày tại Bảng 3.

Thuộc tính	Giá trị chuẩn hóa
Desired Salary	Thỏa thuận; 1,000,000 - 5,000,000; 5,000,000 - 10,000,000; 10,000,000 - 15,000,000; 15,000,000 - 20,000,000; 20,000,000 - 40,000,000; Trên 40,000,000
Work Experience	Không yêu cầu kinh nghiệm; Dưới 1 năm; 1-3 năm; 3-5 năm; 5-10 năm; Trên 10 năm
Industry	Nhân sự; Công nghệ thông tin; Văn phòng; Lái xe; ...

Bảng 3 Chuẩn hóa các thuộc tính của bộ dữ liệu công việc.

- **Các xử lý khác:** Vì những thuộc tính *Current Place of Residence, Province, District* bị missing values mà không thực sự cần thiết trong quá trình xây dựng mô hình và để giảm kích thước bộ dữ liệu nên chúng tôi tiến hành xóa bỏ các thuộc tính này. Với những thuộc tính quan trọng như *Target* và *Skills* nếu có missing values sẽ tiến hành xóa bỏ điểm dữ liệu đó (tức là xóa bỏ ứng viên). Còn lại với các thuộc tính khác bị missing values chúng tôi xử lý bằng cách thay thế giá trị missing values bằng giá trị *"Không xác định"*. Ngoài ra, chúng tôi thay thế thuộc tính *Date of Birth* thành thuộc tính *Age* để phù hợp hơn với bài toán. Hình 4 thể hiện ví dụ về các điểm dữ liệu đã được xử lý của bộ dữ liệu ứng viên.

	URL User	UserID	User Name	Industry	Desired Job	Workplace Desired	Desired Salary	Gender	Marriage	Age	Target	Skills	Degree	Work Experience
0	https://timviec365.vn/ung-vien/nguyen-quan-uv9...	964496	Nguyễn Quân	Vận tải - Lái xe	Nhân viên lái xe bằng B2	Hà Nội	Thỏa thuận	Nam	Đã có gia đình	36	Với kinh nghiệm 10 năm trong lĩnh vực lái xe t...	- Bằng lái xe hạng C, D, E và có kinh nghiệm 5...	Đơn vị đào tạo: Trường trung cấp nghề	5-10 năm
1	https://timviec365.vn/ung-vien/nguyen-chau-uv9...	986206	Nguyễn Châu	Vận tải - Lái xe	Nhân viên lái xe bằng D	Bình Dương	Thỏa thuận	Nam	Đã có gia đình	44	Gắn bó lâu dài với công việc cũng như doanh ng...	Có chí tiến thủ đối với công việc Có khả năng ...	Đơn vị đào tạo: Trường trung cấp nghề	Trên 10 năm
2	https://timviec365.vn/ung-vien/la-thi-hoang-uv...	769632	LÃ THỊ HOÀNG	Hành chính - Văn phòng	Nhân viên văn phòng, khu công nghiệp sử dụng t...	Bắc Giang	Thỏa thuận	Nữ	Độc thân	25	- Tìm kiếm một vị trí nhân viên hành chính và ...	- Am hiểu mọi quy trình về hành chính nhân sự ...	Mô tả công việc: CÔNG TY CỔ PHẦN DƯỢC PHẨM AN ...	1-3 năm
3	https://timviec365.vn/ung-vien/nguyen-van-bao-...	981314	Nguyễn Văn Bảo	Vận tải - Lái xe	Nhân viên lái xe hạng C	Hà Nội	Thỏa thuận	Nam	Đã có gia đình	33	Trở thành kỹ sư tốt trong công tác sửa chữa xe...	Có khả năng chu toàn công việc, siêng năng, tỉ...	Đơn vị đào tạo: Trung tâm dạy nghề	5-10 năm
4	https://timviec365.vn/ung-vien/nguyen-thi-tu-u...	962892	Nguyễn Thị Tứ	Quản lý đơn hàng	Nhân viên thu mua	Hồ Chí Minh	Thỏa thuận	Nữ	Độc thân	24	- Có một công việc có nhiều cơ hội thăng tiến,...	- Trung thực, cẩn thận, có óc sáng tạo - Kỹ...	Đơn vị đào tạo: Cao Đẳng Kinh Tế Kỹ Thuật Vina...	3-5 năm

Hình 4 Bộ dữ liệu ứng viên sau khi đã xử lý.

Sau quá trình xử lý dữ liệu, chúng tôi thu được bộ dữ liệu *Vietnamese Jobs Dataset* hoàn chỉnh gồm hai bộ: bộ dữ liệu công việc và bộ dữ liệu ứng viên. Bộ dữ liệu công việc có kích thước 14,634 dòng và 19 thuộc tính; bộ dữ liệu ứng viên có kích thước 3,983 dòng và 14 thuộc tính.

4 Hướng Tiếp Cận

4.1 Lý thuyết về phương pháp

Trong phần nay chúng tôi sẽ trình bày 3 phương pháp chuẩn hóa nội dung văn bản thành các vector để sử dụng cho phương pháp khuyến nghị dựa trên nội dụng.

4.1.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) là một thống kê số học nhằm phản ánh tầm quan trọng của một từ đối với một văn bản trong một tập hợp hay một ngữ liệu văn bản. Giá trị tf–idf tăng tỉ lệ thuận với số lần xuất hiện của một từ trong tài liệu và được bù đắp bởi số lượng tài liệu trong kho ngữ liệu có chứa từ, giúp điều chỉnh thực tế là một số từ xuất hiện nói chung thường xuyên hơn. TF-IDF có thể được tính toán như công thức 1:

$$TF - IDF(t, d) = tf(t, d) \times log\frac{|D|}{|d : t \subseteq d|} \quad (1)$$

Trong đó: t là một từ, d là một văn bản, tf(t,d) là tần suất xuất hiện của từ t trong văn bản d, $|D|$ là số lượng của tất cả các văn bản được quan sát.

4.1.2 Word2Vec

Word2vec là một kỹ thuật *xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP)*. Thuật toán Word2Vec sử dụng một mô hình *neural network* để học các liên kết từ (sự liên quan của từ) từ một kho ngữ liệu văn bản lớn. Sau khi được huấn luyện, mô hình có thể phát hiện các từ đồng nghĩa hoặc gợi ý các từ bổ sung cho một phần của câu. Word2Vec thể hiện cho mỗi từ riêng biệt với một danh sách cụ thể của các số được gọi là *vector*. Các vector được lựa chọn cẩn thận sao cho một hàm toán học đơn giản sẽ (độ tương tự cosin giữa các vector) cho biết mức độ của độ tương tự ngữ nghĩa giữa các từ được biểu diễn bằng các vector đó.

Word2Vec là một nhóm các mô hình có quan hệ với nhau được dùng để sản sinh các *word embedding*. Các mô hình này là các mạng thần kinh nông hai lớp, được huấn luyện để tái tạo lại ngữ cảnh ngữ nghĩa của các từ vựng. Word2Vec có dữ liệu đầu vào là một ngữ liệu văn bản lớn và đầu ra là một không gian vector, điển hình vài trăm chiều, với mỗi từ duy nhất trong corpus linguistics được gán cho một vector tương ứng trong không gian vector. Các vector từ được đặt trong không gian vector sao cho những từ chia sẻ chung ngữ cảnh trong *corpus (kho ngữ liệu)* có vị trí gần nhau (tính theo độ tương tự ngữ nghĩa) trong không gian.

Hai kiến trúc mô hình điển hình:

- **Skip-gram:** Dự đoán những từ ngữ cảnh nếu biết trước từ đích.
- **CBOW (Continuous Bag of Words):** Dựa vào những từ ngữ cảnh để dự đoán từ đích.

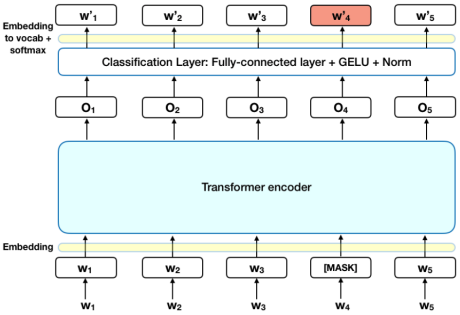
4.1.3 BERT

BERT (Bidirectional Encoder Representations from Transformers)[7] là một mô hình ngôn ngữ được tạo ra bởi Google AI. BERT được coi như là đột phá lớn trong machine learning bởi vì khả năng ứng dụng của nó vào nhiều bài toán NLP khác nhau: Question Answering, Natural Language Inference,... với kết quả rất tốt.

BERT sử dụng Transformer là một mô hình sử dụng *cơ chế attention (attention mechanism)* học mối tương quan giữa các từ (hoặc 1 phần của từ) trong một văn bản. Transformer gồm có 2 phần chính: Encoder và Decoder, encoder thực hiện đọc dữ liệu đầu vào và decoder đưa ra dự đoán. Ở đây, BERT chỉ sử dụng Encoder.

Khác với các mô hình directional (các mô hình chỉ đọc dữ liệu theo 1 chiều duy nhất – trái → phải, phải → trái) đọc dữ liệu theo dạng tuần tự, Encoder đọc toàn bộ dữ liệu trong 1 lần, việc này làm cho BERT có khả năng huấn luyện dữ liệu theo cả hai chiều, qua đó mô hình có thể học được ngữ cảnh (context) của từ tốt hơn bằng cách sử dụng những từ xung quanh nó (phải & trái).

Hình 5 mô tả nguyên lý hoạt động của Encoder. Theo đó, input đầu vào là một chuỗi các token $w1, w2, \ldots$ được biểu diễn thành chuỗi các vector trước



Hình 5 Mô hình Encoder

khi đưa vào trong neural network. Output của mô hình là chuỗi ccs vector có kích thước đúng bằng kích thước input.

4.2 Phương pháp

Phương pháp chúng tôi sử dụng là khuyến nghị dựa trên nội dung. Sử dụng các thuộc tính cùng format của hai bộ dữ liệu công việc và ứng viên để đưa về cùng dạng vector nội dung rồi từ đó sử dụng độ tương đồng cosine và từ đó đưa ra top khuyến nghị cho các ứng viên. Các cặp thuộc tính chúng tôi coi là cùng format để sử dụng cho bài toán khuyến nghị dựa trên nội dung: *Job Requirements – Skill, Industry – Industry, Salary – Desired Salary, Job Title – Desired Job, Years of Experience – Work experience.*

Ngoài việc sử dụng đơn lẻ các thuộc tính chúng tôi còn kết hợp tất cả các thuộc tính để đưa về vector nội dụng. Sau khi có được vector nội dung từ các phương pháp ở phần 4.1 chúng tôi tiến thành sử dụng độ đo cosine để đánh giá sự tương đồng và sau đó đưa ra khuyến nghị.

5 Thực Nghiệm và Phân Tích Kết Quả

Sau quá trình xử lý dữ liệu, chúng tôi xây dựng một hệ khuyến nghị sử dụng các phương pháp được trình bày trong Phần 4 và đánh giá kết quả thông qua độ đo Accuracy để so sánh kết quả đạt được.

5.1 Gán nhãn và đánh giá

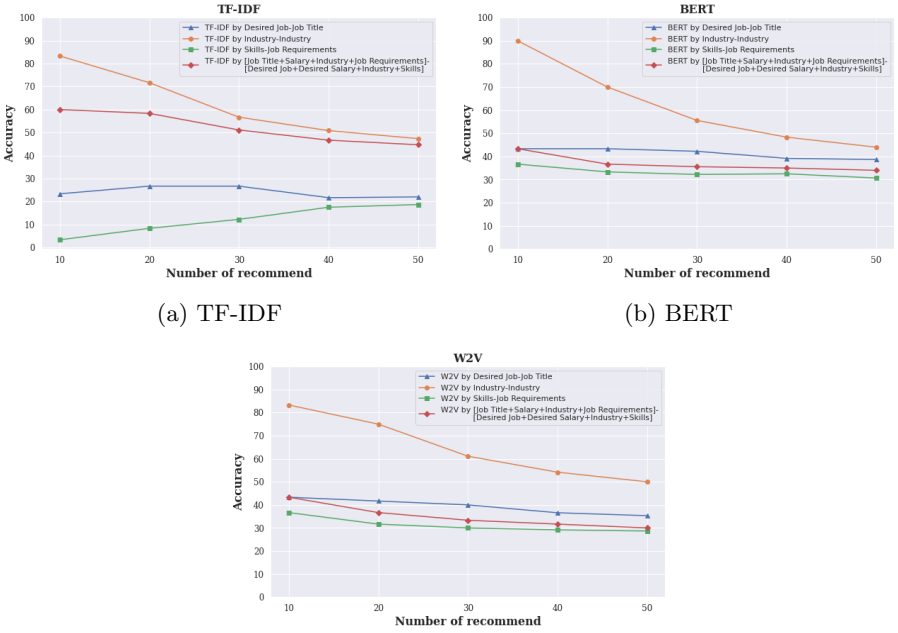
Với từng phương pháp, chúng tôi xây dựng 4 mô hình khác nhau sử dụng kết hợp các thuộc tính tương ứng với nhau ở hai bộ dữ liệu công việc và ứng viên. Với mỗi ứng viên sẽ được khuyến nghị một danh sách gồm 50 công việc. Sau đó, chúng tôi sẽ tiến hành gán nhãn "0" (tức là công việc không phù hợp với ứng viên) hoặc "1" (tức là công việc phù hợp với ứng viên) cho từng công việc một cách độc lập với nhau. Cuối cùng, chúng tôi sẽ đánh giá kết quả theo độ đo Accuracy của Top 10, 20, 30, 40 và 50 công việc được khuyến nghị. Đo độ *Accuracy (A)* được định nghĩa như sau:

$$A = \frac{t}{n} \tag{2}$$

trong đó, t là số lượng công việc được khuyến nghị đúng cho ứng viên (tức là số lượng nhãn "1"), n là số lượng điểm dữ liệu được sử dụng để đánh giá (tức là Top 10, 20, 30, 40 hoặc 50).

Vì quá trình gán nhãn dữ liệu và đánh giá tốn nhiều thời gian và công sức, nên chúng tôi chỉ thực hiện gán nhãn và đánh giá trên 5 ứng viên thuộc các công việc/ngành nghề khác nhau.

5.2 Kết quả thực nghiệm



(a) TF-IDF

(b) BERT

(c) W2V

Hình 6 Kết quả thực nghiệm.

Hệ khuyến nghị dựa trên nội dung được chúng tôi thực nghiệm trên ba phương pháp là TF-IDF, BERT và W2V. Với mỗi phương pháp, chúng tôi kết hợp các thuộc tính tương ứng với nhau ở hai bộ dữ liệu công việc và ứng viên. Qua đó, với mỗi phương pháp chúng tôi có được 4 mô hình. Hình 6 là kết quả trung bình của các thực nghiệm được đánh giá trên 5 ứng viên, ta có thể thấy rằng:

- **Về thuộc tính:** Mô hình sử dụng thuộc tính *Industry-Industry* cho kết quả cao nhất theo độ đo Accuracy ở cả 3 phương pháp và ở tất cả các

số lượng công việc được khuyến nghị. Vì vậy chúng tôi chọn thuộc tính *Industry-Industry* làm thuộc tính để xây dựng mô hình.

- **Về số lượng khuyến nghị:** Đa số các mô hình sẽ cho kết quả Accuracy cao nhất khi khuyến nghị 10 công việc đầu tiên (Accuracy $> 80\%$, tức là cứ 10 công việc được khuyến nghị sẽ có ít nhất 8 công việc có liên quan đến nghành nghề mà ứng viên đang tìm kiếm) và thấp nhất khi khuyến nghị 50 công việc đầu tiên (Accuracy $< 50\%$, tức là cứ 50 công việc được khuyến nghị sẽ có nhiều nhất 25 công việc có liên quan đến nghành nghề mà ứng viên đang tìm kiếm). Khi số lượng khuyến công việc được nghị càng tăng thì độ chính xác sẽ càng giảm. Vì vậy, với mỗi ứng viên chúng tôi sẽ khuyến nghị 10 công việc đầu tiên để giảm thiểu thời gian tìm kiếm và lựa chọn công việc phù hợp với ứng viên, cũng như để đảm bảo độ chính xác khi khuyến nghị danh sách công việc cho từng ứng viên.
- **Về phương pháp:** Chúng tôi thực nghiệm trên ba phương pháp là: TF-IDF, BERT và W2V. Với các thuộc tính và số lượng khuyến nghị được chọn như trên, kết quả Accuracy thu được là: $A_{TF-IDF} = 83.33\%$, $A_{BERT} = 90\%$ và $A_{W2V} = 83.33\%$. Từ kết quả này, ta có thể thấy phương pháp BERT cho kết quả cao nhất trong số ba phương pháp được chúng tôi thực nghiệm.

6 Kết Luận

Trong báo cáo này, chúng tôi đã thu thập, xây dựng, xử lí và trình bày bộ dữ liệu *Vietnamese Jobs Dataset*, một bộ dữ mới cho bài toán khuyến nghị công việc trực tuyến. Bộ dữ liệu gồm 2 tập dữ liệu công việc và ứng viên. Tập công việc với 14,634 công việc và 19 thuộc tính chứa nội dung tuyển dụng của các công ty khác nhau trên cả nước. Tập ứng viên với 3,983 ứng viên và 14 thuộc tính chứa hồ sơ thông tin của ứng viên. Dữ liệu được xử lý và phân tích để thu được bộ dữ liệu sạch từ dữ liệu thô ban đầu để phù hợp với từng phương pháp khuyến nghị dựa trên nội dung. Hiện tại, với phương pháp *Content-based filtering* chúng tôi đã cài đặt thành công mô hình content-based có sử dụng các phương pháp TF-IDF, BERT và Word2Vec để tạo ra các vector văn bản. Kết quả tốt nhất mà chúng tôi đạt được là ở phương pháp BERT với độ chính xác *Accuracy* là 90% với thuộc tính *Industry-Industry*. Chúng tôi đã xử lý, phân tích được bộ dữ liệu từ dữ liệu thô ban đầu; đề xuất ra được độ đo hợp lý trong bài và xây dựng được mô hình khuyến nghị đưa ra kết quả. Nhưng vẫn chưa giải quyết triệt để được yêu cầu bài toán như dữ liệu ứng viên chưa đa dạng. Nó đặt ra một thách thức cho các nhóm nghiên cứu sau về việc cải thiện kết quả cho bài toán.

Hướng phát triển trong tương lai:

- **Bộ dữ liệu:** Thu thập thêm dữ liệu từ các trang web tìm kiếm việc làm trực tuyến, cùng với đó là thu thập thêm các thuộc tính mới như: công việc mà người dùng đã nộp hồ sơ, các công việc mà người dùng đã quan tâm, thống nhất dữ liệu giữa nhà tuyển dụng và người tìm việc làm,... để cho ra bộ dữ liệu đầy đủ thông tin và chính xác hơn. Ngoài ra, cần tăng cường dữ

liệu trong nhiều ngành nghề khác nhau. Đưa ra một phương pháp mới để tận dụng thuộc tính mô tả và yêu cầu công việc.

- **Mô hình:** Áp dụng các phương pháp, kỹ thuật khuyến nghị khác như [8]: *Knowledge-Based Recommender Systems, Demographic Recommender Systems, Hybrid and Ensemble-Based Recommender Systems, ...* để cải thiện kết quả dự đoán tốt hơn nữa.

Tài liệu

[1] Vũ Hữu, Tiệp. "Machine Learning cơ bản." (2022).

[2] Lops, P., de Gemmis, M., Semeraro, G. (2011). Hệ thống đề xuất dựa trên nội dung: Hiện đại và xu hướng. Trong: Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds) Sổ tay Hệ thống Đề xuất. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-85820-3_3

[3] Schafer, JB, Frankowski, D., Herlocker, J., Sen, S. (2007). Hệ thống đề xuất lọc cộng tác. Trong: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Bài giảng về Khoa học Máy tính, tập 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_9

[4] Burke, Robin. "Hybrid web recommender systems." The adaptive web (2007): 377-408.

[5] CHURCH, K. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162. doi:10.1017/S1351324916000334

[6] K. Krishna and M. Narasimha Murty, "Genetic K-means algorithm," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 29, no. 3, pp. 433-439, June 1999, doi: 10.1109/3477.764879.

[7] Khattak, Faiza & Jeblee, Serena & Pou-Prom, Chloe & Abdalla, Mohamed & Meaney, Christopher & Rudzicz, Frank. (2019). A survey of word embeddings for clinical text.

[8] Aggarwal, Charu C. Recommender systems. Vol. 1. Cham: Springer International Publishing, 2016.