Xây dựng bộ dữ liệu Hỏi đáp cho Tiếng Việt về COVID-19*

Thái Minh Triết[1], Chu Hà Thảo Ngân[2], Võ Tuấn Anh[3], Lưu Thanh Sơn[5], and Nguyễn Gia Tuấn Anh[4]

[1] Trường Đại học Công nghệ thông tin, Đại học quốc gia Thành phố Hồ Chí Minh
[2] Khoa Khoa học và Kỹ thuật Thông tin
{[1]19522397,[2]19521882,[3]19521226}@gm.uit.edu.vn
{[4]sonlt,[5]anhngt}@uit.edu.vn

Tóm tắt nội dung Tính đến ngày 21 - 06 - 2021, dịch COVID-19 đã xuất hiện tại 40 tỉnh thành của Việt Nam với hơn 10.000 ca nhiễm. Trước tình hình này, Trung tâm kiểm soát & phòng ngừa dịch bệnh, Học viện Quân Y và các tổ chức y tế công cộng đã nhanh chóng biên soạn những bộ tài liệu hỏi đáp nhằm giải đáp thắc mắc của cộng đồng về COVID-19 và các vấn đề liên quan. Với mong muốn giúp người dân Việt Nam có thể dễ dàng tiếp cận thông tin về COVID-19 và các chính sách phòng chống dịch, chúng tôi tiến hành thu nhập các câu hỏi đáp Tiếng Việt từ website của các tổ chức y tế, trung tâm y tế công cộng và Cổng thông tin điện tử Chính phủ, sau đó tiền xử lý, đánh giá và đưa ra bộ dữ liệu sạch và hoàn chỉnh. Trong tương lai, bộ dữ liệu có thể được sử dụng cho việc xây dựng hệ thống Hỏi đáp tự động, qua đó giúp người dân có thể tiếp cận thông tin một cách nhanh chóng và thuận tiện hơn.

Keywords: COVID-19 · Thu thập dữ liệu · Q&A Dataset· Tidy Data.

1 Giới thiệu chung

Ngày 11 - 03 - 2020, tại cuộc họp báo tại Genève, Thụy Sĩ, Tổng Giám đốc WHO Tedros Adhanom Ghebreyesus đã chính thức tuyên bố sự bùng phát của COVID-19 là đại dịch toàn cầu, trong bối cảnh virus này đã lây lan ở nhiều quốc gia và vùng lãnh thổ trên khắp thế giới. Do đó, việc nắm rõ các thông tin cũng như các chính sách của chính quyền về đại dịch COVID-19 là vô cùng cần thiết, giúp người dân bớt hoang mang và trang bị những kiến thức, kỹ năng phù hợp nhằm giảm thiểu sự lây lan virus ra ngoài cộng đồng.

Trước tình hình số ca mắc mới COVID-19 trên thế giới không ngừng gia tăng, Trung tâm kiểm soát & phòng ngừa dịch bệnh Hoa Kỳ, Học viện Quân Y cùng nhiều trang thông tin y tế điện tử đã nhanh chóng soạn thảo các chuyên mục hỏi đáp và tài liệu hướng dẫn về COVID-19 nhằm nâng cao sự hiểu biết của cộng đồng về COVID-19 và hoạt động phòng chống dịch.

* Hướng dẫn bởi TS. Nguyễn Gia Tuấn Anh và CN. Lưu Thanh Sơn

Trong bối cảnh nội dung số phát triển, nhu cầu tìm kiếm và truy xuất thông tin ngày một tăng lên, vẫn còn nhiều người dân Việt Nam gặp khó khăn trong việc tra cứu thông tin hỏi đáp COVID-19 bằng tiếng Việt để giải đáp những thắc mắc về đại dịch và các chính sách liên quan. Với mong muốn giúp người dân tiếp cận thông tin một cách nhanh chóng, rõ ràng và chính xác, chúng tôi tiến hành xây dựng bộ dữ liệu Hỏi đáp cho Tiếng Việt về COVID-19, nhằm phục vụ cho việc tra cứu cũng như sử dụng cho các bài toán thuộc lĩnh vực Xử lý ngôn ngữ tự nhiên trong tương lai, đặc biệt là bài toán Hỏi đáp tự động.

Thách thức đặt ra là xây dựng một bộ dữ liệu càn phải có tính nhất quán, chính xác và kịp thời các thông tin về vấn đề COVID-19 hiện nay. Từ kết quả đạt được, bộ dữ liệu chúng tôi đã xây dựng nhằm mục đích cung cấp thông tin về COVID-19 và chính sách liên quan đến cộng đồng một cách rõ ràng, mang tính hành động trong công tác ngăn ngừa, phát hiện sớm và kiểm soát COVID-19 trong cộng đồng.

Trong bài báo cáo này, chúng tôi tập trung xây dựng bộ dữ liệu từ những câu hỏi đáp thường gặp về COVID-19 trên ngôn ngữ tiếng Việt. Cấu trúc bài báo cáo được trình bày như sau: Ở mục 2, chúng tôi trình bày chi tiết phương pháp thu nhập và quá trình xây dựng bộ dữ liệu. Ở mục 3, chúng tôi tiếp cận bộ dữ liệu và thực hiện tiền xử lý dữ liệu. Kết quả chúng tôi thu được là bộ dữ liệu tidy data sẽ được trình bày ở mục 4. Và cuối cùng ở mục 5 là kết luận và hướng phát triển.

2 Phương pháp thu thập

2.1 Nguồn thu thập

Bộ dữ liệu được thu thập từ các trang hỏi đáp của những nguồn sau đây:

Trung tâm kiểm soát và phòng ngừa dịch bệnh Hoa Kỳ

- **Website:** `https://vietnamese.cdc.gov`
- **Giới thiệu:** Trung tâm kiểm soát và phòng ngừa dịch bệnh Hoa Kỳ (tiếng Anh: Centers for Disease Control and Prevention, viết tắt là CDC) là một cơ quan thuộc Bộ Y tế và Dịch vụ Nhân sinh Hoa Kỳ có trụ sở tại quận DeKalb, tiểu bang Georgia, Hoa Kỳ. Nhiệm vụ của CDC tập trung vào việc phát triển và ứng dụng hệ thống phòng ngừa và kiểm soát bệnh tật (nhất là bệnh truyền nhiễm), sức khỏe môi trường, y tế nghề nghiệp, giúp nâng cao sức khỏe người dân Hoa Kỳ. Trong bối cảnh dịch bệnh diễn biến phức tạp, CDC đã soạn thảo trang thông tin về COVID-19 nhằm giải đáp câu hỏi thường gặp và nâng cao nhận thức của người dân về đại dịch.
- **Nội dung thu thập:** Các nội dung giải đáp và khuyến cáo mà CDC đưa ra xoay quanh chủ đề về đại dịch , bao gồm các thông tin cơ bản về COVID-19, xét nghiệm, vắc xin, du lịch, tổ chức tang lễ,... [7]. Những nội dung này đều hỗ trợ phiên bản Tiếng Việt nên phù hợp để chúng tôi xem xét đưa vào bộ dữ liệu.

Hệ thống tiêm chủng VNVC

- **Website:** `https://vnvc.vn`
- **Giới thiệu:** Hệ thống tiêm chủng VNVC (thuộc Công ty Cổ phần Vacxin Việt Nam) chính thức đi vào hoạt động từ tháng 6 năm 2017. Hệ thống tiêm chủng VNVC ra đời nhằm cung cấp cho người Việt những loại vắc xin có chất lượng tốt nhất cùng với hệ thống phòng tiêm chủng an toàn, hiện đại và cao cấp [2]. Khi dịch COVID-19 bắt đầu bùng phát ở khắp nơi trên thế giới, VNVC đã nhanh chóng tổng hợp và giải đáp 101 câu hỏi thường gặp về virus corona chủng mới.
- **Nội dung thu thập**: Các thông tin hỏi đáp cơ bản về COVID-19, sự lây lan của COVID-19, triệu chứng, cách phòng ngừa, truy vết, vệ sinh - khử trùng, xét nghiệm và vắc xin [8].

Học viện Quân y Việt Nam

- **Website:** `http://hocvienquany.vn`
- **Giới thiệu:** Trước tình hình dịch bệnh diễn biến phức tạp tại Việt Nam, Tiểu ban kỹ thuật của Ban Chỉ đạo Phòng chống dịch viêm phổi cấp do chủng mới của virus Corona (Covid-19) trong Quân đội đã thống nhất đề nghị Học viện Quân y chủ trì biên soạn gấp các tài liệu hướng dẫn kỹ thuật và giáo dục cộng đồng để cung cấp ngay cho hoạt động phòng chống dịch Covid-19. Với tinh thần trách nhiệm cao, Học viện Quân y đã tập hợp các nhà khoa học, chuyên gia trong các chuyên ngành liên quan đến phòng chống dịch bệnh của Học viện tổng hợp, biên soạn và biên tập sổ tay "100 câu hỏi - đáp về dịch bệnh Covid-19". Các thông tin khoa học này giúp người đọc có thêm hiểu biết đầy đủ hơn về dịch bệnh, từ đó thực hiện đúng hơn và tuân thủ tốt hơn các biện pháp phòng chống dịch vừa để bảo vệ bản thân vừa bảo vệ cộng đồng [11].
- **Nội dung thu thập:** Bao gồm các thông tin hỏi đáp chung về dịch bệnh COVID-19, tác nhân gây bệnh, đề kháng chống virus, triệu chứng bệnh và biện pháp phòng bệnh.

Bộ Giáo dục và Đào tạo

- **Website:** `https://moet.gov.vn/`
- **Giới thiệu:** Để đảm bảo các điều kiện an toàn cho học sinh, sinh viên, học viên trở lại trường học tập sau thời gian tạm nghỉ học để phòng, chống dịch Covid-19, Bộ Giáo dục và Đào tạo phối hợp với Học viện Quân y biên soạn tài liệu "100 câu hỏi - đáp về phòng, chống dịch bệnh Covid-19 trong các cơ sở giáo dục" nhằm cung cấp cho cha mẹ học sinh (CMHS), các em học sinh, sinh viên (HSSV) và cán bộ nhân viên ngành Giáo dục những kiến thức cơ bản về dịch Covid-19, hướng dẫn thực hiện cho đúng và hiệu quả các biện pháp phòng chống dịch do các cơ quan chức năng triển khai, đồng thời có thái độ đúng mực không chủ quan nhưng cũng không quá lo sợ trước dịch bệnh [12].

– **Nội dung thu thập:** Bao gồm các thông tin hỏi đáp chung về dịch bệnh COVID-19, hỏi đáp chung cho cha mẹ học sinh, học sinh - sinh viên, giáo viên và cán bộ giáo dục. Một số câu hỏi đáp trong tài liệu này có sự trùng lặp với các câu hỏi đáp trong tài liệu "100 câu hỏi - đáp về dịch bệnh Covid-19" do Học viện quân y soạn thảo, và sẽ được lọc bỏ ở bước tiền xử lý dữ liệu.

Giải đáp chính sách online - Cổng thông tin điện tử Chính phủ

– **Website:** `https://chinhsachonline.chinhphu.vn`
– **Giới thiệu:** Giải đáp chính sách online là trang thông tin điện tử thuộc Cổng thông tin điện tử Chính phủ Việt Nam. Là nơi các Bộ, Ban, Ngành giải đáp trực tuyến những thắc mắc của người dân về các chính sách và chỉ thị của Chính phủ trong các lĩnh vực y tế, giáo dục, đầu tư,... Trước tình hình dịch bệnh phức tạp, Cổng thông tin điện tử Chính phủ đã bổ sung thêm chuyên mục Hỏi đáp về COVID-19 nhằm tạo điều kiện cho người dân được bày tỏ những thắc mắc của mình [9].
– **Nội dung thu thập:** Bao gồm các thông tin hỏi đáp về COVID-19 và những ảnh hưởng của đại dịch trong các lĩnh vực Giáo dục - Đào tạo, Y tế, Bảo hiểm xã hội, Tài chính - Ngân hàng,...

Hệ thống Y tế Vinmec

– **Website:** `https://vinmec.com`
– **Giới thiệu:** Vinmec là hệ thống y tế phi lợi nhuận do Tập đoàn Vingroup đầu tư phát triển từ năm 2012. Tới nay, Vinmec trở thành thương hiệu y tế tư nhân hàng đầu Việt Nam với đầy đủ các cấu phần tương tự các hệ thống y tế hàng đầu thế giới: Bệnh viện đa khoa - Phòng khám - Chuỗi các viện nghiên cứu chuyên sâu - Hệ thống đào tạo chuyên nghiệp [6]. Trong bối cảnh dịch COVID-19 đang hoành hành tại Việt Nam, đội ngũ chuyên gia, bác sĩ đầu ngành của Vinmec đã tổng hợp và đăng tải các câu hỏi đáp thường gặp về đại dịch nhằm nâng cao nhận thức của người dân và cộng đồng [10].
– **Nội dung thu thập:** Gồm các câu hỏi đáp xoay quanh vấn đề về nguồn gốc COVID-19, cách thức lây nhiễm, triệu chứng, cách phòng bệnh, xét nghiệm, du lịch,...

2.2 Ngôn ngữ lập trình và môi trường lập trình

Ngôn ngữ lập trình: Python

– **Ưu điểm:**
  • Cú pháp ngắn gọn, dễ hiểu, dễ sử dụng,
  • Chạy được trên nhiều hệ điều hành,
  • Nhiều thư viện hỗ trợ cho việc cào dữ liệu web như BeautifulSoup, Scrapy,...
  • Nhiều thư viện hỗ trợ xử lý dữ liệu văn bản.
– **Nhược điểm:**
  • Tốc độ xử lý của Python không nhanh bằng JAVA và C++.
  • Không có vòng lặp **do. . . while** và **switch. . . .case**.
– **Thư viện sử dụng:** *BeautifulSoup*, *pycpdf*, *request, re*

Môi trường lập trình: Google Colaboratory

- **Ưu điểm:**
  - Tích hợp sẵn các thư viện Python phổ biến
  - Cung cấp tài nguyên về bộ nhớ và không gian lưu trữ.
  - Hỗ trợ GPU để tăng tốc độ xử lý và tính toán.
  - Miễn phí (khi không sử dụng GPU)
  - Hỗ trợ kết nối với Google Drive
- **Nhược điểm:**
  - Cần kết nối Internet để có thể sử dụng.
  - Tốn thêm phí để sử dụng GPU không giới hạn.

2.3 Cách thức thu thập

Cách thức thu thập và chuyển đổi dữ liệu từ dạng thô về dạng tidy được thực hiện như sau:

- **Bước 1:** Thu thập câu hỏi và câu trả lời từ dữ liệu thô (Raw Data)

  Đối với dữ liệu dạng XML:
  - **Bước 1.1:** Duyệt qua tuần tự các đường dẫn đến trang hỏi đáp của các nguồn thu thập,
  - **Bước 1.2:** Sử dụng hàm *get()* của thư viện *request* để lấy phản hồi từ đường dẫn của trang đang xét.
  - **Bước 1.3:** Sử dụng hàm *BeautifulSoup()* của thư viện *bs4* để lấy nội dung HTML từ phản hồi.
  - **Bước 1.4:** Sử dụng phương thức *find_ all()* để tìm các tag HTML chứa câu hỏi và câu trả lời từ mã nguồn HTML thu được ở Bước 1.3.
  - **Bước 1.5:** Lấy nội dung dạng văn bản bằng cách gọi phương thức *text* của tag tìm được. Câu hỏi và câu trả lời được lưu tạm vào cấu trúc list trong Python.

  Đối với dữ liệu dạng PDF:
  - **Bước 1.1:** Sử dụng hàm *PDF()* của thư viện *pycpdf* để đọc và lấy nội dung từ file pdf.
  - **Bước 1.2:** Duyệt qua tuần tự từng trang trong đối tượng *pycpdf.PDF* thu được ở Bước 1.1.
  - **Bước 1.3:** Sử dụng các hàm xử lý chuỗi thông dụng trong Python như *split()*, *sub()*, *remove()*,... để xử lý, nhận diện và tách các câu hỏi và câu trả lời tương ứng ở mỗi trang.
  - **Bước 1.4:** Lưu câu hỏi và câu trả lời vào cấu trúc list trong Python.
- **Bước 2:** Gộp các list câu hỏi và câu trả lời tương ứng từ các nguồn, sau đó lưu dưới dạng cấu trúc dataframe hỗ trợ bởi thư viện *pandas*. Dataframe gồm 3 thuộc tính: mã định danh câu hỏi đáp, câu hỏi và câu trả lời.
- **Bước 3:** Xuất bộ dữ liệu hoàn chỉnh dưới dạng file .csv sử dụng phương thức *to_ csv()* của dataframe.

3 Tiền xử lý dữ liệu

Sau khi thu thập và đưa về dữ liệu dạng tidy data, chúng tôi tiến hành các bước tiền xử lý sau để làm sạch dữ liệu đã thu thập.

3.1 Loại bỏ các ký tự thừa, các chỉ mục và cụm từ không liên quan

Bảng 1. Ví dụ về lời giải đáp trước và sau khi tiền xử lý.

Trước khi tiền xử lý	**Sau khi tiền xử lý**
Bộ Giáo dục và Đào tạo 100 câu hỏi - đáp về phòng, chống dịch bệnh Covid-19 trong các cơ sở giáo dục 36 Sốt là biểu hiện của rất nhiều bệnh có thể có hoặc không liên quan đến bệnh Covid-19...	Sốt là biểu hiện của rất nhiều bệnh có thể có hoặc không liên quan đến bệnh Covid-19...
Trả lời: Chúng tôi không có thông tin từ các báo cáo khoa học nào được công bố về mức độ dễ mắc của phụ nữ mang thai...	Chúng tôi không có thông tin từ các báo cáo khoa học nào được công bố về mức độ dễ mắc của phụ nữ mang thai...

Khi thực hiện thu thập dữ liệu từ các trang web và tài liệu pdf, có một số chỉ mục, số thứ tự hoặc các cụm từ không liên quan vô tình được thu thập. Những cụm từ và ký tự thừa này đôi khi khiến cho các câu hỏi đáp không rõ ràng và không truyền tải hết được ý nghĩa nội dung. Do đó, cách tốt nhất là xác định và loại bỏ chúng để bộ dữ liệu trở nên sạch hơn.

3.2 Loại bỏ các câu hỏi đáp bị trùng lặp

Bảng 2. Ví dụ về câu hỏi đáp bị trùng lặp

Câu hỏi	**Trả lời**
Hiện nay các biện pháp chính để điều trị bệnh do Covid-19 là gì?	Do chưa có thuốc điều trị đặc hiệu, nên việc điều trị hỗ trợ nâng đỡ thể trạng, sức đề kháng và điều trị triệu chứng là chủ yếu. Cần theo dõi và phát hiện sớm, xử lý kịp thời các ca bệnh nặng, nguy kịch như suy hô hấp hoặc suy các tạng khác.
Các biện pháp chính để điều trị bệnh Covid-19 là gì?	Hiện nay, do chưa có thuốc điều trị đặc hiệu nên việc điều trị chủ yếu là hỗ trợ nâng đỡ thể trạng, sức đề kháng và điều trịtriệu chứng.

Tuy bộ dữ liệu được thu thập từ nhiều nguồn khác nhau, nhưng đều xoay quanh chủ đề về dịch bệnh, do đó khó tránh khỏi sự trùng lặp, không nhất quán.

Phần lớn các nội dung bị trùng lặp nằm ở 2 quyển tài liệu do Học viện Quân y soạn thảo.

Để đảm bảo tính nhất quán cho bộ dữ liệu, chúng tôi tiến hành duyệt qua các câu hỏi đáp và loại bỏ nếu chúng bị trùng lặp với những câu hỏi đáp đã có trong bộ dữ liệu.

3.3 Loại bỏ các câu hỏi đáp không phù hợp

Bảng 3. Ví dụ về một số câu hỏi đáp không phù hợp

Câu hỏi	**Trả lời**
Có những tài liệu nào về vấn đề chôn cất cựu chiến binh Người Mỹ Da Đỏ/Người Alaska Bản Địa?	Vợ/chồng và gia đình Cựu Chiến Binh Người Mỹ Da Đỏ/Người Alaska Bản Địa có thể đủ điều kiện được hỗ trợ chôn cất...
"Luồng không khí có định hướng" nghĩa là gì? Chúng ta nên sử dụng luồng không khí này ở đâu và như thế nào?	Luồng khí có định hướng là một khái niệm thông gió bảo vệ trong đó luồng không khí chuyển động theo hướng từ sạch đến kém sạch...
Mất bao lâu để hòa tan hàm lượng các loại hạt tạp gây lây nhiễm trong một căn phòng sau khi được tạo ra?	Khi những giọt nước lớn (từ 100 micrometer [μm] trở lên) sẽ lắng xuống các bề mặt xung quanh trong vài giây, những hạt nhỏ hơn có thể ngưng đọng trong không khí lâu hơn. Có thể mất vài phút để các hạt có kích thước 10 μm lắng xuống, trong lúc các hạt có kích thước từ 5 μm trở xuống có thể không lắng xuống trong nhiều giờ hoặc nhiều ngày...

Có một số câu hỏi đáp do CDC (Hoa Kỳ) soạn thảo chưa phù hợp với văn hóa, lối sống con người Việt Nam nên được chúng tôi loại bỏ, do các nội dung này dựa trên các chính sách áp dụng tại Hoa Kỳ và chỉ đơn thuần được dịch sang tiếng Việt.

Ngoài ra còn có những câu hỏi đáp mang nặng tính thuật ngữ, khó hiểu cũng bị loại bỏ ra khỏi bộ dữ liệu.

4 Bộ dữ liệu tidy data

Sau quá trình thu nhập dữ liệu từ các nguồn thông tin và thực hiện tiền xử lí, kết quả thu được là bộ dữ liệu hoàn chỉnh với 620 điểm dữ liệu và 3 thuộc tính, gồm các câu hỏi và trả lời bằng ngôn ngữ tiếng Việt liên quan tới các vấn đề xoay quanh đại dịch COVID-19. Dưới đây là ví dụ các điểm dữ liệu được trích ra từ bộ dữ liệu và codebook của bộ dữ liệu.

Bảng 4. Ví dụ về các điểm dữ liệu trong bộ dữ liệu

id	Question	Answer
QA1	Tại sao bệnh này lại được gọi là bệnh vi-rút corona 2019, COVID-19?	Vào ngày 11 tháng 2 năm 2020, Tổ Chức Y Tế Thế Giới đã công bố tên chính thức cho căn bệnh gây ra đại dịch do vi-rút corona mới 2019 gây ra, được phát hiện lần đầu tại Vũ Hán, Trung Quốc. Tên mới của bệnh này là bệnh vi-rút corona 2019, viết tắt là COVID-19. Trong chữ COVID-19, 'CO' viết tắt của từ 'corona,' 'VI' viết tắt của từ 'vi-rút,' và 'D' là bệnh. Trước đó, căn bệnh này được gọi là "vi-rút corona mới 2019" hoặc "nCoV-2019".
QA122	Khẩu trang vải có giống như trang bị bảo hộ cá nhân (PPE) không?	Không, khẩu trang vải không phải là trang bị bảo hộ cá nhân (PPE). Các loại khẩu trang này không phải là mặt nạ và không phải là vật dụng thay thế thích hợp cho các trang bị bảo hộ đó tại nơi làm việc có khuyến cáo hoặc bắt buộc sử dụng mặt nạ để bảo vệ đường hô hấp.
QA339	Nên duy trì chế độ ăn như thế nào để tăng sức đề kháng phòng chống Covid-19?	Không có chế độ ăn đặc hiệu để tăng sức đề kháng riêng với Covid-19 Nên duy trì chế độ ăn hợp lý, đủ chất đinh dưỡng, có thể bổ sung vitamin để tăng sức đề kháng chung Do chưa loại trừ khả năng lây qua thức ăn nên thực hiện ""ăn chín uống sôi"" Tuyệt đối không ăn đồ ăn sống như tiết canh, thịt sống, đặc biệt là tiết canh, thịt sống của động vật hoang dã
QA437	Tôi đã khai báo y tế online trên trang khai báo y tế của Bộ Y tế, khi đã khai báo xong thì tôi có cần phải đến UBND xã để khai báo lại không?	Bộ Y tế trả lời vấn đề này như sau:Theo hướng dẫn của Bộ Y tế thì ông chỉ cần khai báo y tế online trên website: tokhaiyte.vn để phục vụ việc giám sát, truy vết nếu có nghi ngờ mắc Covid-19. Tuy nhiên, do việc áp dụng công nghệ thông tin tại các địa phương có thể chưa được triển khai thống nhất nên đề nghị ông liên hệ với chính quyền địa phương nơi đến để thực hiện.
QA607	Covid-19 có thể lây truyền qua nước uống không?	Covid-19 không thể lây truyền qua đường nước uống được. Virus được phát tán tới những người tiếp xúc gần thông qua giọt bắn, nhưng mà không qua nước uống.

Bảng 5. Codebook của bộ dữ liệu

Thông tin	**Nội dung**
Tên bộ dữ liệu	Bộ dữ liệu Hỏi đáp về COVID-19 cho Tiếng Việt (Vietnamese COVID-19 Question And Answering Dataset)

Nguồn thu thập	Từ 6 nguồn:
	1. CDC Hoa Kỳ:
	`https://vietnamese.cdc.gov/coronavirus/2019-ncov/faq.html`
	2. VNVC:
	`https://vnvc.vn/cau-hoi-thuong-gap-covid-19/`
	3. Học viện quân y, Cục Quân y:
	`http://asttmoh.vn/wp-content/uploads/2020/03/100-cau-hoi-dap-ve-dich-Covid-19-Cuc-QY-Final.pdf`
	4. Bộ Giáo dục & Đào tạo:
	`https://moet.gov.vn/content/tintuc/Lists/News/Attachments/6507/1582974693976755.pdf`
	5. Giải đáp chính sách online - Cổng thông tin điện tử Chính phủ:
	`https://chinhsachonline.chinhphu.vn/Cau-hoi-chon-loc/Hoi-dap-ve-dich-COVID19/14.vgp`
	6. Vinmec:
	`https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/dich-2019-ncov/tu-van-bac-si/`

Cách thức thu thập	Xác định, xử lý và trích xuất các câu hỏi và câu trả lời tương ứng từ mã nguồn HTML của các trang web và tài liệu pdf.
Số điểm dữ liệu	620
Số thuộc tính	3
Thông tin thuộc tính	**id:** Mã định danh câu hỏi đáp **Question:** Nội dung câu hỏi **Answer:** Nội dung câu trả lời

5 Kết luận và hướng phát triển

5.1 Kết luận

Trong bài báo cáo này, nhóm chúng tôi đã trình bày quá trình xây dựng bộ dữ liệu về các câu hỏi đáp liên quan đến COVID-19, bao gồm các bước thu thập, chuyển đổi dữ liệu thô sang dữ liệu tidy và các bước tiền xử lý để làm sạch dữ

liệu. Bộ dữ liệu thu được gồm 620 điểm dữ liệu và 3 thuộc tính. Dù đảm bảo được tính kịp thời, đầy đủ, tin cậy và nhất quán, tuy nhiên trong bộ dữ liệu vẫn còn một số ít câu hỏi đáp chưa rõ ý nghĩa, dài dòng cần được điều chỉnh thêm để hoàn thiện hơn về mặt logic, đảm bảo hơn về tính chính xác và dễ hiểu.

5.2 Hướng phát triển

Từ kết quả đã có, trong tương lai nhóm chúng tôi sẽ thực hiện các điều chỉnh về mặt ngữ nghĩa, đồng thời bổ sung thông tin về vấn đề vắc-xin COVID-19 đang được quan tâm hàng đầu hiện nay. Qua đó, bộ dữ liệu được đảm bảo hơn tính kịp thời, tin cậy và chính xác để phù hợp với tình hình dịch bệch đang còn diễn biến phức tạp trên thế giới và tại Việt Nam.

Nhóm chúng tôi cũng sẽ tiếp tục xây dựng và mở rộng kích thước của bộ dữ liệu để sử dụng trong các bài toán về Xử lý Ngôn ngữ Tự nhiên trên ngôn ngữ tiếng Việt. Một số bài toán có thể sử dụng trên bộ dữ liệu như: Bài toán Đọc hiểu tự động (Machine Reading Comprehension hay MRC), Bài toán Hỏi đáp tự động (Automatic Question Answering), Bài toán Xây dựng Chatbot và Trợ lý ảo, Bài toán Gắn thẻ câu hỏi tự động (Automatic Question Tagging)...

Tài liệu

1. Trang chủ CDC Hoa Kỳ, https://vietnamese.cdc.gov.
2. Trang chủ Hệ thống tiêm chủng VNVC - Công ty cổ phần Vacxin Việt Nam, https://vnvc.vn/
3. Trang chủ Học viện Quân y, http://hocvienquany.vn/Portal/TrangChu.html
4. Trang chủ Bộ Giáo dục và Đào tạo, https://moet.gov.vn/
5. Trang chủ Giải đáp chính sách online - Cổng thông tin điện tử Chính phủ, https://chinhsachonline.chinhphu.vn
6. Trang chủ Bệnh viện Đa khoa Quốc tế Vinmec | Vinmec, https://vinmec.com
7. Câu hỏi thường gặp về vi-rút corona (COVID-19) | CDC, https://vietnamese.cdc.gov/coronavirus/2019-ncov/faq.html
8. [GIẢI ĐÁP] 101 CÂU HỎI THƯỜNG GẶP VỀ VIRUS CORONA (2019-NCOV), https://vnvc.vn/cau-hoi-thuong-gap-covid-19/
9. HỎI ĐÁP VỀ DỊCH COVID-19, https://chinhsachonline.chinhphu.vn/Cau-hoi-chon-loc/Hoi-dap-ve-dich-COVID19/14.vgp
10. Hỏi - đáp về dịch Covid - 19 | Vinmec, https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/dich-2019-ncov/tu-van-bac-si/
11. Học viện Quân y, Cục Quân y: 100 CÂU HỎI ĐÁP VỀ DỊCH BỆNH COVID-19. Hà Nội (17/02/2020)
12. Học viện Quân y, Bộ Giáo dục & Đào tạo: 100 CÂU HỎI ĐÁP VỀ PHÒNG, CHỐNG DỊCH BỆNH COVID-19 TRONG CÁC CƠ SỞ GIÁO DỤC. Hà Nội (24/02/2020)