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 Lưu 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 viên 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 để giúp cho cộng đồng có thể tra cứu thông tin về COVID-19 và các chính sách của chính phủ trong bối cảnh dịch bệnh còn đang diễn biến phức tạp. Trong tương lai, bộ dữ liệu có thể được sử dụng cho việc xây dựng các mô hình 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

Sự bùng phát dịch bệnh COVID-19 được tuyên bố là tình trang y tế cộng đồng khẩn cấp, virus này đã lây lan nhiều quốc gia và nhiều vùng lãnh khổ, đây là tình trạng gây quan ngại quốc tế. Việc cộng đồng nắm được thông tin và diễn biến của COVID-19 đóng vai trò quan trọng, giúp cộng đồng hiểu rõ về dịch bệnh, giảm đi sự lo lắng cũng như có kiến thức cho cộng đồng hành động ngăn ngừa sự lan rộng của dịch bệnh.

Trước tình hình này, Tổ chức Y tế Thế giới WHO, Trung tâm kiểm soát & phòng ngừa dịch bệnh CDC, Bộ Y tế, Học viện Quân Y và nhiều Cổng thông tin y tế điện tử đã nhanh chống biên soạn tài liệu hướng dẫn kỹ thuật và giáo dục cộng đồng cho hoạt động phòng chống COVID-19. Tuy trong bối cảnh nội dung số phát triển hiện nay, nhu cầu tìm kiếm, tra cứu và truy xuất thông tin

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

chính xác, đầy đủ và kịp thời là hơn bao giờ hết, nhưng phần lớn người dân Việt Nam còn gặp nhiều khó khăn để có thể tiếp cận những nguồn tin chính thống tiếng Việt để giải đáp những thắc mắc của mình về đại dịch và các chính sách liên quan. Với mong muốn cung cấp thông tin cho người dân 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 để phục vụ cho bài toán tra cứu cũng như bài toán Hỏi đáp tự động trong tương lai.

Thách thức đặt ra là xây dựng một bộ dữ liệu 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 nhập

3 Tiền xử lý dữ liệu

4 Bộ dữ liệu tidy data

Sau khi quá trình thu nhập từ các nguồn ở mục và thực hiện tiền xử lí ở mục 3, kết quả thu được là bộ dữ liệu hoàn chỉnh về các câu hỏi và trả lời sử dụng ngôn ngữ tiếng Việt liên quan tới dịch COVID-19 gồm 620 điểm dữ liệu. Bảng 1 thể hiện ví dụ các điểm dữ liệu được trích ra từ bộ dữ liệu và bảng 2 biểu diễn codebook của bộ dữ liệu.

Bảng 1: Ví dụ các điểm dữ liệu của bộ dữ liệu

id	Question	Answer
QA1	Tại sao bệnh này lại được gọi là .. COVID-19?	Vào ngày 11 tháng 2 năm 2020, Tổ Chức Y Tế Thế Giới ...
QA607	Covid-19 ... truyền qua nước uống không?	Covid-19 không thể lây truyền ... qua nước uống.
QA339	Nên duy trì chế độ ăn...phòng chống Covid-19?	Không có chế độ ăn ... riêng với Covid-19
QA127	... xử lý rác thải là trang bị bảo hộ cá nhân ...?	Thải bỏ PPE vào thùng rác. Rác thải tại ... khử trùng.

QA122	Khẩu trang vải có giống ... (PPE) không?	Không, khẩu trang vải không phải là ... (PPE)...

Bảng 2: Codebook của bộ dữ liệu

Thông tin	Nội dung
Tên bộ dữ liệu	Covid-19_Q&A
Nguồn thu nhập và cách thu nhập	Bộ dữ liệu được thu nhập tự động qua các nguồn trang web - CDC: +https://vietnamese.cdc.gov/coronavirus/2019-ncov/faq.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/vaccines/faq.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/travelers/faqs.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/community/general-business-faq.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/community/retirement/faq.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/community/correction-detention/faq.html +ttps://vietnamese.cdc.gov/coronavirus/2019-ncov/community/large-events/event-planners-and-attendees-faq.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/liver-disease.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/community/tribal/faq-burial-practice.html +https://vietnamese.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/hiv.html

	+https://vietnamese.cdc.gov/coronavirus/2019-ncov/community/medicolegal-faq.html -VNVC: +https://vnvc.vn/cau-hoi-thuong-gap-covid-19/ -Cục Quân Y: +urlhttp://asttmoh.vn/wp-content/uploads/2020/03/100-cau-hoi-dap-ve-dich-Covid-19-Cuc-QY-Final.pdf -Bộ GDĐT: +https://moet.gov.vn/content/tintuc/Lists/News/Attachments/6507/1582974693976755.pdf - Cổng thông tin chính sách của chính phủ: +https://chinhsachonline.chinhphu.vn/Cau-hoi-chon-loc/Hoi-dap-ve-dich-Covid19/14/trang1.vgp -VinMec: +https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/dich-2019-ncov/tu-van-bac-si/hoi-dap-ve-dich-virus-corona-2019-phan-8-lay-nhiem-covid-19-o-tre-em-phu-nu-mang-thai -tre-so-sinh- va-ba-me-cho-con-bu/
Số thuộc tính	3
Thông tin các thuộc tính	id: định danh câu hỏi đáp Question: câu hỏi Answe: câu trả lời

5 Kết luận và hướng phát triển

Trong bài báo cáo, nhóm đã trình bày quá trình xây dựng bộ dữ liệu về câu hỏi đáp liên quan tới COVID-19 gồm 620 câu hỏi và câu trả lời.

Từ kết quả đã có, trong tương lai nhóm sẽ thực hiện phân tích và tiếp tục xây dựng bộ dữ liệu với mong muốn bộ dữ liệu kích thước lớn hơn, tăng thêm tính kịp thời, tin cậy và chính xác phù hợp với tình hình dịch bệch COVID-19 diễn biến phức tạp trên thế giới và Việt Nam đồng thời cung cấp, bổ sung thông tin về vấn đề vắc-xin COVID-19 đang được quan tâm hàng đầu hiện nay. Từ đó bộ dữ liệu có thể thích hợp cho những bài toán về ngôn ngữ tự nhiên sử dụng ngôn ngữ là tiếng Việt như xây dựng hệ thống hỏi đáp tự động, ...

5.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 3 gives a summary of all heading levels.

Bảng 3. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

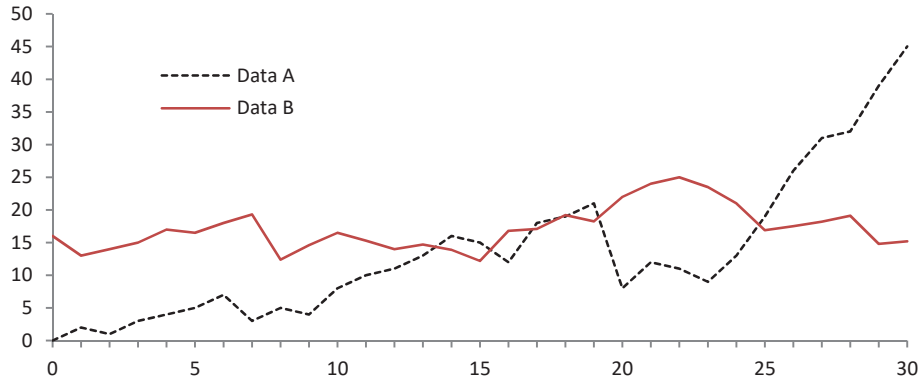
Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Chứng minh. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.



Hình 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1,2,3], [1,3,4,5].

Tài liệu

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, `http://www.springer.com/lncs`. Last accessed 4 Oct 2017