

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

Thái Minh Triết¹, Chu Hà Thảo Ngân², Võ Tuấn Anh³, Lưu Thanh Sơn⁵, and
Nguyễn Lưu Tuấn Anh⁴

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

{¹19522397, ²19521882, ³19521226}@gm.uit.edu.vn

{⁴sonlt, ⁵anhngt}@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 tiến 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 trạng y tế cộng đồng khẩn cấp, virus này đã lây lan nhiều quốc gia và nhiều vùng lãnh thổ, 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 chóng 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 của bài toán chúng tôi 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

5 Kết luận và hướng phát triển

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.

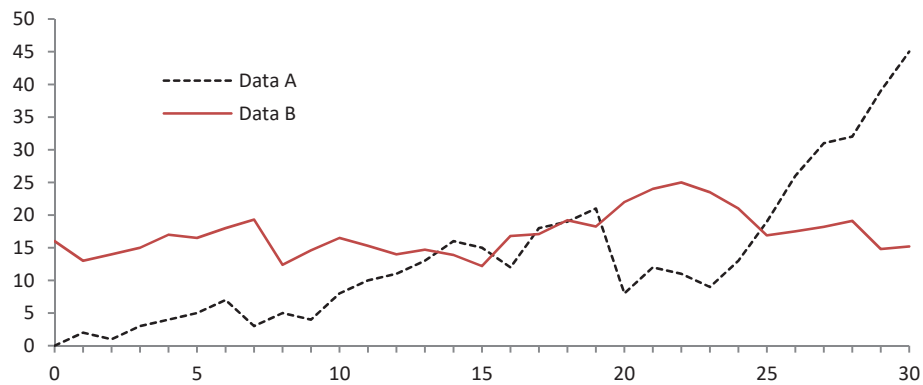
Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels. Displayed equations are centered and set on a separate line.

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

Bảng 1. 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	<i>Lowest Level Heading.</i> Text follows	10 point, italic



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.

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.

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 [1–3], [1, 3–5].

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/10.1007/1234567890>
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