**Thu thập văn bản song ngữ tự động Việt - Hoa từ các website**

**Trịnh Văn Minh1, Lê Nhựt Quang2, Trần Thanh Phước3**

*NLP-KD Lab, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh city, VietNam*

*1Email:51403301@tdt.edu.vn*

*2Email:51403014@tdt.edu.vn*

*3Email: tranthanhphuoc@tdt.edu.vn*

**Tóm tắt: Ngữ liệu đơn - song ngữ là vô cùng cần thiết, phục vụ cho các bài toán xử lý ngôn ngữ tự nhiên, đặc biệt là dành cho dịch máy. Hoa-Việt là cặp ngôn ngữ nghèo tài nguyên, dữ liệu song ngữ Hoa-Việt hiện còn hạn chế. Vì vậy, trong bài báo này, chúng tôi đề xuất thu thập tự động các văn bản song ngữ Hoa-Việt từ các website song ngữ Hoa-Việt. Ngữ liệu văn bản song ngữ này là tiền đề để rút trích các cặp câu song ngữ ở những công trình tiếp theo. Hệ thống thu thập của chúng tôi được tiến hành trên 10 trang web song ngữ Việt – Hoa và bước đầu đã cho được kết quả đáng khích lệ. Hệ thống này có thể được triển khai để thu thập tự động cho nhiều cặp ngôn ngữ khác.**

1. **Giới thiệu**

Kho ngữ liệu song ngữ là bắt buộc đối với các hướng tiếp cận dịch máy hiện đại, như dịch máy dựa trên thống kê hay dịch máy dựa trên nơ ron. Tất cả các phương pháp dịch máy đều được ước tính bằng cách huấn luyện tối đa và có khả năng lặp đi lặp lại dựa trên dữ liệu lớn các văn bản song ngữ tự nhiên. Từ đó cho thấy dữ liệu đầu vào để cho dịch máy thống kê là rất quan trọng và chất lượng của dịch máy cũng phụ thuộc vào dữ liệu. Do đó, xây dựng một tập hợp các câu, từ song ngữ là một nhiệm vụ hết sức quan trọng trong việc xây dựng bất kỳ hệ thống dịch máy nào.

Để có được dữ liệu cho dịch máy, các nhà nghiên cứu ngôn ngữ học thường sử dụng một trong hai phương pháp đó là thu thập thủ công hoặc rút trích văn bản tự động từ các website. Phương pháp thủ công có độ chính xác cao, nhưng phải đòi hỏi có một đội ngũ chuyên gia về ngôn ngữ học và am hiểu về song ngữ. Phương pháp thủ công này thì tốn rất nhiều thời gian và chi phí.

Hiện nay với sự phát triển của các website song ngữ ngày càng phong phú cho nên phương pháp rút trích tự động cũng mang lại hiệu quả cao, tuy nhiên mức độ chính xác sẽ không cao bằng phương pháp thủ công, nhưng cũng có thể sử dụng trong nghiên cứu được. Với phương pháp tự động này sẽ tiết kiệm được thời gian và chi phí và số lượng từ ngữ thu được sẽ rất nhiều.

Trong bài báo này, mục đích của chúng tôi là bước đầu thu thập dữ liệu song ngữ Việt- Hoa từ các trang báo mạng có dữ liệu song ngữ của người Việt. Chúng tôi lựa chọn thu thập dữ liệu từ các trang báo mạng vì những trang báo này có nhiều thể loại như là: kinh tế, chính trị, văn hóa, giáo dục, thể thao, đời sống,… Vốn từ ngữ và nội dung của các trang báo này rất phong phú, đủ các lĩnh vực từ sang trọng cho đến đời sống dân dã, để các bài báo được thể hiện ở trên các trang báo mạng có uy tín như vậy thì cũng đã được thông qua bộ phận kiểm duyệt về nội dung để phù hợp với mọi tầng lớp, cũng như không vi phạm pháp luật. Như vậy các trang báo mạng là một nguồn dữ liệu dồi dào và an toàn,là tiền để cho dữ liệu song ngữ Việt-Hoa.

Ngoài ra, chúng tôi còn giới thiệu phương pháp thu thập dữ liệu dựa trên thư viện có tên (JSOUP) chúng tôi biết đến theo đường dẫn trong tài liệu tham khảo [3] để lấy các đoạn văn bản bên trong tên thẻ HTML[[1]](#footnote-1). Thư viện này là thư viện của Java, được tạo ra để làm việc với HTML. Thư viện này cung cấp một API[[2]](#footnote-2) rất tiện lợi để rút trích và thao tác dữ liệu, sử dụng các phương pháp tốt nhất của DOM[[3]](#footnote-3), CSS [[4]](#footnote-4)và jQuery.

Công dụng của thư viện này bao gồm: Rút trích, phân tích HTML từ đường dẫn, tìm và rút trích dữ liệu, thao tác các trên các phần tử, ngăn chặn các cuộc tấn công XSS[[5]](#footnote-5).

Trước tiên, chúng tôi thực hiện phương pháp tìm số phân trang lớn nhất để biết có tổng bao nhiêu phân trang có thể lấy và để lưu lại số phân trang hiện tại đang lấy trên tổng số phân trang. Từ đó, chúng tôi tiến hành cấu hình để lấy văn bản trong thẻ tiêu đề và thẻ nội dung. Sau nhiều lần cải tiến, chúng tôi đã làm cho tốc độ lấy dữ liệu tăng đáng kể (cụ thể trong phần 4 trình bày về thử nghiệm).

Chúng tôi đã tiến hành thu thập dữ liệu trên một số trang báo, với nguồn dữ liệu thu thập được chúng tôi trình bày trong phần 4 (Thử nghiệm). Chúng tôi thu thập về và lưu dưới dạng tập tin văn bản, để dễ dàng xử lý và lưu trữ. Ngoài ra, chúng tôi còn phát triển ứng dụng để tiện lợi cho thu thập thêm dữ liệu từ các website khác sau này và lấy dữ liệu mới từ các website đã thu thập.

Phần còn lại của bài báo này theo cấu trúc như sau: phần 2 trình bày một số công trình liên quan, phần 3 trình bày chi tiết về phương pháp lấy dữ liệu, phần 4 trình bày về các dữ liệu chúng tôi đã thử nghiệm. Cuối cùng, phần 5 là Hướng phát triển và lời kết**.**

1. **Công trình liên quan**

Trong phần này, chúng tôi tập trung vào việc khảo sát các phương pháp thu thập dữ liệu từ web. Để biết thêm các phương pháp thu thập dữ liệu và thu thập dữ liệu tự động như vậy có hợp pháp và các thư viện mã nguồn mở hỗ trợ lấy dữ liệu từ web.

Trong tài liệu tham khảo [1] có phần nghiên cứu về các kỹ thuật thu thập dữ liệu tự động có hợp pháp trên các quốc gia trên thế giới. Kết quả là chưa có bất kỳ thủ tục tố tụng pháp lý liên quan đến rút trích dữ liệu tự động. Tuy nhiên, ở nước Đức đã có quyết định vế quyền sở hữu dữ liệu trực tuyến và ngăn chặn tình trạng thu thập dữ liệu tự động.

Nhưng hầu hết về luật pháp trên thế giới, chỉ cần không dùng các thủ thuật để thu thập tài khoản, mật khẩu, thông tin cá nhân, hoặc các tệp dữ liệu bị cấm tải xuống và không gây thiệt hại cho chủ sở hữu dữ liệu trước khi trả phí thì đều hợp pháp.

Với nghiên cứu ở [1] thì chúng tôi rút trích văn bản tự động từ các trang báo mạng là hoàn toàn hợp pháp, vì dữ liệu báo mạng là cho mọi người cùng đọc và được để ở chế độ công khai. Ngoài ra, chúng tôi không dùng bất kỳ các thủ thuật nào khác để lấy thông tin cá nhân của người dùng hoặc gây thiệt hại cho chủ sở hữu.

Trong tài liệu tham khảo [2] nghiên cứu về các phương pháp rút trích và kỹ thuật khai thác văn bản, đã có phần nói về các thư viện mã nguồn mở để rút trích dữ liệu, và đưa ra các một số mã nguồn mở trong đó có JSOUP mà chúng tôi đang sử dụng. Giulio Barcaroli [2] cũng đưa ra một số khó khăn khi sử dụng mã nguồn mở này đó là các trang web không thể truy cập đầy đủ và không hoàn toàn dựa trên văn bản chuẩn HTML.

Từ nghiên cứu [2] chúng tôi biết được những khó khăn mình gặp phải và nên tùy thuộc vào công nghệ sử dụng cho trang báo đó, chúng tôi có các cách lấy dữ liệu khác nhau.

1. **Phương pháp lấy dữ liệu**

Để thuận lợi cho việc lấy dữ liệu từ nhiều nguồn báo mạng khác nhau, để mang lại nguồn dữ liệu thử nghiệm lớn khi cần thiết, chúng tôi đã tạo ra một ứng dụng để lấy dữ liệu dựa trên mã nguồn mở JSOUP, với mã nguồn mở này hỗ trợ cho chúng tôi lấy nội dung bên trong thẻ HTML khi chúng tôi truy cập vào trang web đó. Ví dụ, với một trang báo chúng tôi chỉ lấy tiêu đề và nội dung bên trong trang báo đó, không lấy các phần khác, khi sử dụng mã nguồn mở này, chúng tôi đưa đầu vào là tên thẻ chứa nội dung cần lấy trong HTML với đường dẫn của trang báo đó, thư viện sẽ trả về cho chúng tôi nội dung bên trong thẻ đó.

Chúng tôi lựa chọn các trang báo có dữ liệu song ngữ để làm đầu vào, để có được dữ liệu lớn và môi trường thử nghiệm chính xác điều đầu tiên là phải tìm các trang báo uy tín, có độ tin cậy cao. Sau đó tiến hành lấy tất cả các bài báo (không bao gồm hình ảnh và phim ảnh).

Do số lượng các bài báo quá nhiều trên một trang, để làm giảm không gian tìm kiếm các bài báo có cùng một nội dung thì ngay từ lúc lấy dữ liệu này phải phân loại dữ liệu ra. Ví dụ: trang vietnamplus.vn có dữ liệu là ngôn ngữ tiếng việt và ngôn ngữ tiếng trung, thì chúng tôi sẽ thu thập dữ liệu được phân loại theo chuyên mục (chính trị, kinh tế, văn hóa,…).

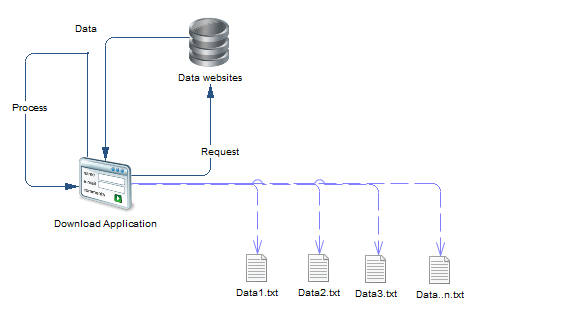
Với dữ liệu song ngữ Việt Nam và Trung Quốc, do các trang báo là của người việt và chủ yếu để cho người Việt Nam sử dụng nên việc các dữ liệu tiếng Việt thường nhiều hơn dữ liệu tiếng Trung, dữ liệu tiếng việt thường được viết các tin mới nhiều hơn dữ liệu tiếng Trung. Do vậy dữ liệu mẫu ban đầu từ các trang báo mạng đã không cân bằng và bị nhiễu, vì vậy phải dùng nhiều phương pháp sàng lọc để tìm ra các cặp câu, từ song ngữ.

Mỗi một trang báo mạng song ngữ có các đặc điểm khác nhau, thậm chí cùng một trang nhưng cấu trúc giữa tiếng Việt và tiếng Trung khác nhau, hoặc là cùng một ngôn ngữ trên một trang nhưng các danh mục (kinh tế, xã hội, …) lại khác nhau.

Ví dụ như trên cùng một trang nhưng danh mục kinh tế để dữ liệu kiểu phân trang và lưu trữ dữ liệu theo cách truyền thống (mỗi loại danh mục thông tin có hiển thị số trang cuối cùng), nhưng cũng chính trang đó, danh mục xã hội lại không phân trang và dùng cách truyền dữ liệu bằng ajax (người dùng sẽ đọc các tin mới nhất ở trên cùng, nếu người dùng muốn xem tin cũ hơn phải kéo xuống, server lúc này sẽ yêu cầu các tin tức cũ hơn để trả về cho người dùng).

Với các trang báo khác nhau, tùy vị trí đặt server, số tiền mà chủ trang báo chi trả cho server cũng mang lại tốc độ truy cập và lấy dữ liệu khác nhau. Như các khó khăn nêu ra ở trên, chúng tôi phải xử lý hết để có nguồn dữ liệu song ngữ lớn và phục vụ cho các bước tác ra dữ liệu song ngữ hoàn chỉnh.

Hình 1 mô tả sơ lược phương pháp lấy dữ liệu. Chúng tôi chọn ngôn ngữ lập trình java để hiện thực các phương pháp lấy dữ liệu từ báo mạng. Ngoài ra, nhờ sử dụng thêm thư viện JSOUP để phân tích, tách cấu trúc của một trang HTML ra thành các thẻ do đội ngũ phát triển trang báo đó tạo ra, từ đó lấy nội dung bên trong theo tên thẻ mà chúng tôi muốn.



Hình 1: Cấu trúc của ứng dụng download.

Sau các bước chọn trang báo để lấy dữ liệu và lựa chọn nội dung bên trong các trang .Chúng tôi hiện thực theo các bước sau:

* Bước 1: Xác định tiêu đề và nội dung nằm trong thẻ nào của danh mục cần lấy.

Chúng tôi sử dụng thư viện JSOUP để kiểm tra xem, ứng với danh mục này, thì tiêu đề và nội dung nằm ở thẻ có tên là gì, và tên các thẻ cha của nó. Từ đó chúng tôi có thể lấy dữ liệu bên trong thẻ.

* Bước 2- Tìm số phân trang lớn nhất trong danh mục của trang báo cần lấy.

Bước này rất quan trọng, khi có số phân trang lớn nhất chúng tôi sẽ lấy từ phân trang thứ nhất cho đến phân trang cuối cùng để đảm bảo các bài báo đều được lấy về. Do dữ liệu của các mục báo rất nhiều, có danh mục của một trang lên đến 200,000 bài báo nên không phải lúc nào chúng tôi cũng có thời gian để lấy trọn vẹn từ phân trang đầu tiên cho đến phân trang cuối cùng (thời gian lấy dữ liệu của một trang có thể lên đến vài ngày), cũng như phòng ngừa rủi ro trong lúc đang lấy dữ liệu gặp các hiện tượng bị ngắt kết nối nên chúng tôi sẽ lưu lại trang đang lấy hiện tại trên tổng số trang để biết chúng tôi đã lấy được bao nhiêu phần trăm dữ liệu.

Chúng tôi hiện thực cách lấy số phân trang lớn nhất theo đoạn mã giả dưới đây:

**Algorithm:getMaxPageNumber(String pageName)**

\*/ pageName là đường dẫn mà chúng tôi muốn lấy.

\*/ result và factor là những con số ngẫu nhiên để bắt đầu.

\*/checkValidLink(result) là hàm kiểm tra với số phân trang đó có tồn tại đường dẫn hay không

1: Integer result = 1000, factor = 1000;

2: Boolean previous = true, reachInvalidLink = false;

3: **while(true)**

4: **if(checkValidLink(result))**

5: **if(reachInvalidLink = true)**

6: factor = factor /2;

7: result = result + factor;

8: **else**

9: result = result + factor;

10: previous = true;

11: **else**

12 factor = factor/2;

13: result = result - factor;

14: previous = true; reachInvalidLink = true;

15: **if(factor <= 5)**

16: break;

17: **if(factor >=5000 || factor <=0)**

20: return -1;

21: **end while**

23: **if(pre = true)**{

\*/Nếu chạy 10000 lần không có kết quả thì ngừng

24: **for( i = result; i<10000; i++){**

25: **if(checkValidLink(i))**

24: return i;

25: **end for**

26: **else**

27: **for(i = result; i>=1; i--)**

28 **if(checkValidLink(i))**

24: return i;

25: **end for**

26: return -1;

Về cơ bản, ban đầu chúng tôi không biết số phân trang lớn nhất là bao nhiêu. Nên chúng tôi cho một con số phân trang ngẫu nhiên để kiểm tra, nếu phân trang đó tồn tại, tiếp tục kiểm tra phân trang lớn hơn, và nếu không tồn tại, chúng tôi sẽ kiểm tra các số phân trang nhỏ hơn cho đến khi thấy phân trang đó có tồn tại và không có phân trang lớn hơn tồn tại.

* Bước 3- Lấy tiêu đề và nội dung của trang báo và lưu xuống tập tin. Tuy nhiên, trong quá trình lấy sẽ xảy ra rất nhiều trục trặc, có thể trang báo đó tồn tại mà không có nội dung, hoặc thời gian để truy cập vào trang báo đó quá lâu.

Có một số trang tồn tại nhưng chỉ toàn là hình ảnh kèm với tên của hình ảnh đó, như vậy số lượng dữ liệu về ngôn ngữ trong danh mục đó rất ít. Với việc lấy dữ liệu tự động này, sẽ có những phân trang mà không tồn tại dữ liệu khi chúng tôi bắt đầu lấy từ phân trang đầu tiên cho đến phân trang lớn nhất. Như vậy chúng tôi sẽ bỏ qua hết những rủi ro, trục trặc nêu trên để tiếp tục cho đến khi lấy xong hoàn toàn dữ liệu của trang đó.

1. **Thử nghiệm**
   1. **Công cụ thử nghiệm**

Chúng tôi tự tạo ra ra ứng dụng để thử nghiệm dựa trên mã nguồn mở JSOUP trên ngôn ngữ lập trình Java. Với mã nguồn mở này hỗ trợ cho chúng tôi lấy nội dung bên trong thẻ HTML, từ đó chúng tôi có thể cấu hình riêng cho mỗi trang để lấy nội dung do mỗi trang web có các kiểu cấu hình khác nhau, về phương pháp chúng tôi đã trình bày ở *phần 3.Phương pháp lấy dữ liệu.*

* 1. **Đặc điểm các trang báo**

Mỗi một trang web có các đặc điểm khác nhau, nhưng về số lượng dữ liệu thu thập được thì hầu hết các trang báo mạng đều có số lượng bài báo tiếng Việt nhiều hơn bài báo tiếng Trung, do các trang báo này là của người Việt Nam thành lập. Tốc độ cụ thể khi thu thập chúng tôi đã thống kê trong trong phần *4.3 Kết quả thử nghiệm*. Vì vậy chúng tôi thống kê các đặc điểm của những trang web theo bảng 1 bên dưới.

|  |  |
| --- | --- |
| Địa chỉ trang web | Đặc điểm riêng |
| http://www.vietnamplus.vn | - Cách lấy số phân trang: Số phân trang lớn nhất nằm ở cuối mỗi nội dung. Chỉ cần chạy từ phân trang đầu tiên đến phân trang cuối cùng lấy dữ liệu.  - Trong quá trình lấy, gặp một số trang báo có tồn tại, có thể truy cập đến, nhưng hệ thống đang bảo trì.  - Số lượng bài báo nhiều.  - Tốc độ khi thu thập nhanh, nhưng không ổn định, chậm dần về sau. |
| http://baothainguyen.org.vn | - Cách lấy số phân trang: Số phân trang lớn nhất nằm ở cuối mỗi nội dung. Chỉ cần chạy từ phân trang đầu tiên đến phân trang cuối cùng lấy dữ liệu.  - Dữ liệu tiếng trung không có phân loại nội dung và dữ liệu quá ít.  - Tốc độ lấy rất chậm, số lượng bài báo không nhiều. |
| http://www.sggp.org.vn | - Cách lấy số phân trang: Kiểm tra từng phân trang xem có hợp lệ, nếu hợp lệ là số phân trang trong link nhỏ hơn hoặc bằng số phân trang lớn nhất. Nếu số phân trang đang kiểm tra lớn hơn số phân trang lớn nhất sẽ bị trở về phân trang đầu tiên. Dùng thuật toán ở phần 3.  - Tốc độ thu thập nhanh, ổn định trong quá trình. |
| http://www.tapchicongsan.org.vn | - Cách lấy số phân trang: Bằng thuật toán ở phần 3.  - Dữ liệu tiếng Trung quá ít.  - Tốc độ thu thập trung bình. |
| http://thoidai.com.vn | - Cách lấy số phân trang: Với tiếng Việt có hiện số trang lớn nhất ở cuối nội nội dung. Nhưng tiếng Trung thì phương pháp lấy như trang http://www.sggp.org.vn, chỉ khác số phân trang không hợp lệ là không có danh sách bài viết, và phải kiểm tra phân trang đó có trống hay không.  - Dữ liệu tiếng Trung và tiếng Việt khá tương đồng.  - Tốc độ thu thập trung bình khá. |
| http://baobinhduong.vn | - Cách lấy số phân trang: Số phân trang lớn nhất không có, tuy nhiên khi lấy vượt số phân trang lớn nhất, trang web tự trả về số phân trang lớn nhất.  - Tốc độ thu thập dữ liệu trung bình. |
| http://www.nhandan.com.vn | - Cách lấy số phân trang: Bằng thuật toán ở phần 3.  - Dữ liệu tiếng Trung chạy với tốc độ trung bình. Với tiếng Việt dữ liệu nhiều và tốc độ thu thập rất chậm. Tốc độ chậm dần về sau. |

Bảng 1: Đặc điểm các trang báo mạng đã thu thập dữ liệu.

* 1. **Kết quả thử nghiệm**

Chúng tôi tiến hành thử nghiệm trên hệ điều hành Ubuntu 16.04 LTS 64 bits với một số thông tin cấu hình như RAM 16GB, Intel Core i5-7400 CPU, đồ họa GeForce GT 730/PCIe/SSE2 và kết nối mạng ổn định với tốc độ tải xuống 4.1 MB/s và tốc độ tải lên 4.7 MB/s.

Tất cả các trang web báo mạng mà chúng tôi thử nghiệm đều không tương đồng hoàn toàn về nội dung. Trong các bảng ở phần này, những con số **0** xuất hiện trong bảng là không có nội dung tương ứng để thu thập dữ liệu.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.vietnamplus.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 49,031 | 231 | 11,631 | 43.5 |
| Chính trị | 12,808 | 68.6 | 15,287 | 62.1 |
| Thể thao | 24,472 | 108 | 766 | 2.94 |
| Văn hóa | 15,868 | 77.4 | 5,727 | 21.4 |
| Công nghệ | 22,721 | 94 | 928 | 3.56 |
| Xã hội | 37,717 | 184 | 2,545 | 9,85 |
| Thế giới | 114,432 | 456 | 960 | 4.01 |
| **Đời sống** | 11,707 | 55,5 | **0** | **0** |
| **Khoa học** | 8,896 | 37.6 | **0** | **0** |
| **Chuyện lạ** | 3,686 | 5.32 | **0** | **0** |
| **Môi trường** | **0** | **0** | 1,147 | 4.35 |

Bảng 2: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://www.vietnamplus.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.vietnamplus.vn | Số danh mục[[6]](#footnote-6) | Số lượng tập tin thu được[[7]](#footnote-7) | Tổng dung lượng  (MB) | Tốc độ tải[[8]](#footnote-8)  (tập tin/phút) |
| Tiếng Trung | 9 | 46,863 | 181 | 434 |
| Tiếng Việt | 10 | 301,422 | 1351 | 273 |

Bảng 3: Tổng kết quả thu thập dữ liệu song ngữ trang http://www.vietnamplus.vn

Từ kết quả thống kê ở bảng 2 và bảng 3 cho thấy dữ liệu song ngữ của trang http://www.vietnamplus.vn là khá nhiều cả tiếng Trung và tiếng Việt. Với số lượng tập tin tiếng Việt nhiều gấp hơn 6 lần tiếng Trung, và số lượng tập tin thu được nhiều gấp hơn 7 lần. Số lượng nội dung tiếng Việt có mà tiếng Trung không có là 3, và tiếng Trung có mà tiếng Việt không có là 1 nội dung.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://baobinhduong.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 13,874 | 78.9 | 6,317 | 24.4 |
| Chính trị | 18,516 | 77 | 2,711 | 5.10 |
| Sức khỏe | 60 | 0.3 | 1,021 | 1.48 |
| Văn hóa | 5,439 | 26,8 | 2,114 | 8.10 |
| Công nghệ | 60 | 0.3 | 496 | 1.92 |
| Xã hội | 15,701 | 80.9 | 5,567 | 20.9 |
| Quốc tế | 8,257 | 37.1 | 1,553 | 5.96 |
| Môi trường | 1,979 | 9,78 | 324 | 1.25 |
| **Chính sách** | **0** | **0** | 335 | 1.26 |
| **Bất động sản** | **0** | **0** | 97 | 0.39 |
| **Biển đảo** | **0** | **0** | 110 | 0.4 |
| **Du lịch** | **0** | **0** | 968 | 3.71 |
| **Việt Nam và Thế Giới** | **0** | **0** | 2,751 | 10.8 |
| **Bạn đọc** | 2,139 | 12.9 | **0** | **0** |
| **Doanh nghiệp** | 933 | 5.64 | **0** | **0** |
| **Gia đình** | 2,755 | 13.7 | **0** | **0** |
| **Hồ sơ - Tư liệu** | 1,057 | 10.2 | **0** | **0** |
| **Lao động** | 2,394 | 12.5 | **0** | **0** |
| **Ô tô - Xe máy** | 60 | 0.3 | **0** | **0** |
| **Phân tích** | 2,089 | 11.1 | **0** | **0** |
| **Pháp luật** | 4,837 | 24.1 | **0** | **0** |
| **Quốc phòng - An ninh** | 2,302 | 12.1 | **0** | **0** |
| **Thể thao** | 8,192 | 40.6 | **0** | **0** |

Bảng 4: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://baobinhduong.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://baobinhduong.vn | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 13 | 24,344 | 39.3 | 286 |
| Tiếng Việt | 18 | 90,644 | 492 | 192 |

Bảng 5: Tổng kết quả thu thập dữ liệu song ngữ trang http://baobinhduong.vn

Từ kết quả thống kê ở bảng 4 và bảng 5 cho thấy dữ liệu song ngữ của trang http://baobinhduong.vn chênh lệch rất nhiều, Số lượng tập tin thu được của tiếng Việt nhiều hơn xấp xỉ 4 lần, nhưng tổng dung lượng của tiếng Việt lại nhiều hơn gấp 12 lần. Do đó, chúng tôi dự đoán nội dung bên trong các tập tin tiếng Trung cùng một về một vấn đề sẽ ít hơn rất nhiều so với tiếng Việt. Số lượng nội dung tiếng Việt có mà tiếng Trung không có là 10, nhưng tiếng Trung có mà tiếng Việt không có là 5 nội dung.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.sggp.org.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 2,153 | 13 | 418 | 1.39 |
| Chính trị | 2,445 | 17.8 | 943 | 2.78 |
| Công nghệ | 1,470 | 7.96 | 182 | 0.6 |
| Sức khỏe | 1,470 | 7.53 | 296 | 1.17 |
| Thế giới | 1,466 | 6.28 | 950 | 3.64 |
| Giáo dục | 1,466 | 8.64 | 357 | 1.28 |
| Pháp luật | 1,465 | 6.58 | 402 | 1.22 |
| Bạn đọc | 2,156 | 12.6 | 246 | 1.20 |
| **Đời sống công nghệ** | 284 | 1.31 | **0** | **0** |
| **Văn hóa** | 1,939 | 11.2 | **0** | **0** |
| **Xã hội** | 1,464 | 8.12 | **0** | **0** |
| **Hoạt động người Hoa** | **0** | **0** | 300 | 1.01 |
| **Giải trí** | **0** | **0** | 433 | 1.31 |
| **Thể thao** | **0** | **0** | 329 | 0.9 |
| **Du lịch** | **0** | **0** | 90 | 0.3 |

Bảng 6: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://www.sggp.org.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.sggp.org.vn | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 12 | 4,956 | 16.8 | 251 |
| Tiếng Việt | 11 | 17,778 | 101 | 456 |

Bảng 7: Tổng kết quả thu thập dữ liệu song ngữ trang http://www.sggp.org.vn

Từ kết quả thống kê ở bảng 6 và bảng 7 cho thấy dữ liệu song ngữ của trang http://www.sggp.org.vn không tương đồng cao. Với tổng số tập tin thu được của tiếng Việt nhiều gấp hơn 3 lần và tổng dung lượng nhiều gấp 6 lần. Số lượng nội dung tiếng Việt có mà tiếng Trung không có là 3 và tiếng Trung có mà tiếng Việt không có là 4 nội dung.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.tapchicongsan.org.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 455 | 8.26 | 33 | 0.5 |
| Chính trị | 556 | 10.9 | 44 | 0.7 |
| Xã hội | 124 | 2.10 | 30 | 0.4 |
| **Tin tức** | **0** | **0** | 183 | 1.21 |
| **An ninh** | 103 | 1.39 | **0** | **0** |
| **Bình luận** | 958 | 13.1 | **0** | **0** |
| **Đối ngoại** | 213 | 3.48 | **0** | **0** |
| **Nghiên cứu** | 1,234 | 27.7 | **0** | **0** |
| **Sinh hoạt-Tư tưởng** | 147 | 1.19 | **0** | **0** |
| **Thế giới** | 2,157 | 29.1 | **0** | **0** |
| **Thông tin luận lý** | 1,457 | 12.9 | **0** | **0** |

Bảng 8: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://www.tapchicongsan.org.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.tapchicongsan.org.vn | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 5 | 328 | 3.44 | 434 |
| Tiếng Việt | 10 | 7,404 | 110 | 495 |

Bảng 9: Tổng kết quả thu thập dữ liệu song ngữ trang http://www.tapchicongsan.org.vn

Từ kết quả thống kê ở bảng 8 và bảng 9 cho thấy dữ liệu song ngữ của trang http://www.tapchicongsan.org.vn chênh lệch rất nhiều. Với tổng số tập tin thu được của tiếng Việt nhiều hơn gấp 22 lần và tổng dung lượng nhiều hơn gấp gần 32 lần. Dữ liệu song ngữ chỉ có 3 nội dung ở trang web này. Với dữ liệu như vậy, trang báo này hoàn toàn không phù hợp để phân tích tìm ra dữ liệu song ngữ.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://thoidai.com.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 3,794 | 22 | 2,468 | 9.24 |
| Thế giới | 5,460 | 27.8 | 816 | 2.92 |
| Gia đình việt | 473 | 3.57 | 612 | 2.30 |
| Bạn năm châu | 1074 | 6.26 | 2,212 | 8,42 |
| Văn hóa | 4,975 | 27.7 | 3,194 | 11.5 |
| **Yêu quê hương** | **0** | **0** | 915 | 3.50 |
| **Chính trị** | **0** | **0** | 4,521 | 18.2 |
| **Xã hội** | **0** | **0** | 2,096 | 7.75 |
| **Mối quan hệ Việt Nam** | **0** | **0** | 623 | 2.47 |
| **An toàn giao thông** | 40 | 0.2 | **0** | **0** |
| **Bạn đọc** | 192 | 1.50 | **0** | **0** |
| **Giáo dục** | 1,186 | 6.96 | **0** | **0** |
| **Môi trường** | 573 | 2.91 | **0** | **0** |
| **Nhân ái** | 50 | 0.2 | **0** | **0** |
| **Nông thôn** | 414 | 2.65 | **0** | **0** |
| **Pháp luật** | 1,478 | 8.85 | **0** | **0** |
| **Sức khỏe** | 2,106 | 10.4 | **0** | **0** |
| **Tấm lòng bạn bè** | 160 | 0.8 | **0** | **0** |
| **Thể thao** | 2,040 | 11.6 | **0** | **0** |
| **Thời sự** | 5,234 | 32.9 | **0** | **0** |

Bảng 10: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://thoidai.com.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://thoidai.com.vn | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 9 | 17,502 | 66.4 | 129 |
| Tiếng Việt | 16 | 29,249 | 166 | 104 |

Bảng 11: Tổng kết quả thu thập dữ liệu song ngữ trang http://thoidai.com.vn

Từ kết quả thống kê ở bảng 10 và bảng 11 cho thấy dữ liệu song ngữ của trang http://thoidai.com.vn không bị chênh lệch quá nhiều về dung lượng, cũng như tập tin thu thập được số với các trang web khác đã thử nghiệm nhưng trang web này bị chênh lệch khá nhiều ở nội dung, cụ thể có 20 nội dung trong trang báo này nhưng chỉ có 5 nội dung là trùng khớp giữa tiếng Trung và tiếng Việt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.nhandan.com.vn | Tiếng Việt | | Tiếng Trung | |
| Số lượng tập tin thu được | Tổng dung lượng(MB) | Số lượng tập tin thu được | Tổng dung lượng(MB) |
| Kinh tế | 27,324 | 163 | 6,287 | 22.2 |
| Chính trị | 57,256 | 345 | 6,701 | 26.8 |
| Xã hội | 35,507 | 183 | 6,861 | 24.5 |
| Văn hóa | 18,663 | 111 | 1,762 | 6.14 |
| Thế giới | 34,258 | 154 | 4,143 | 15.2 |
| Thể thao | 8,567 | 39.4 | 1,532 | 5.02 |
| **Bạn đọc** | 5,136 | 25.3 | **0** | **0** |
| **Công nghệ** | 4,761 | 23.2 | **0** | **0** |
| **Giáo dục** | 8,126 | 47.5 | **0** | **0** |
| **Khoa học** | 6,652 | 39.3 | **0** | **0** |
| **Pháp luật** | 17,438 | 114 | **0** | **0** |
| **Sức khỏe** | 7,436 | 39.9 | **0** | **0** |
| **Asean** | **0** | **0** | 252 | 1.01 |
| **Tài liệu** | **0** | **0** | 244 | 1.33 |
| **Cầu nối hữu nghị** | **0** | **0** | 5,940 | 22.2 |
| **Hoạt động chính phủ** | **0** | **0** | 2,320 | 9.3 |
| **Nội bộ** | **0** | **0** | 287 | 1.01 |
| **Đất nước Việt Nam** | **0** | **0** | 77 | 0.3 |
| **Du lịch** | **0** | **0** | 1,690 | 5.87 |
| **Cửa sổ Việt Nam** | **0** | **0** | 202 | 0.8 |

Bảng 12: Kết quả thu thập dữ liệu song ngữ đã phân loại theo nội dung trang http://www.nhandan.com.vn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.nhandan.com.vn | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 14 | 38,298 | 142 | 103 |
| Tiếng Việt | 12 | 231,133 | 1280 | 110 |

Bảng 13: Tổng kết quả thu thập dữ liệu song ngữ trang http://www.nhandan.com.vn

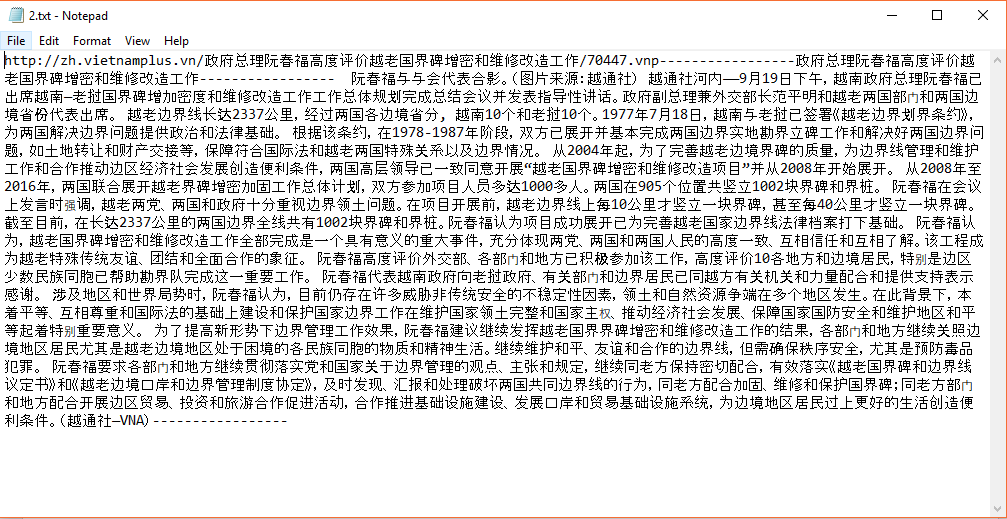
Từ kết quả thống kê ở bảng 12 và bảng 13 cho thấy dữ liệu song ngữ của trang http://www.nhandan.com.vn bị chênh lệch nhiều về số lượng tập tin thu được, tổng dung lượng của các tập tin, các danh mục mang cùng một nội dung trùng lặp là 6 trên tổng 20. Dữ liệu tiếng Trung của trang báo này cũng khá nhiều ở phần song ngữ tuy rằng không nhiều bằng tiếng Việt, nhưng đây cũng là một trang web có tiềm năng để rút trích, phân tích sau này cho bài toán song ngữ.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://baothainguyen.org.vn/ | Số danh mục | Số lượng tập tin thu được | Tổng dung lượng  (MB) | Tốc độ tải  (tập tin/phút) |
| Tiếng Trung | 16 | 36,813 | 195 | 42 |
| Tiếng Việt | 1 | 96 | 0.3 | 47 |

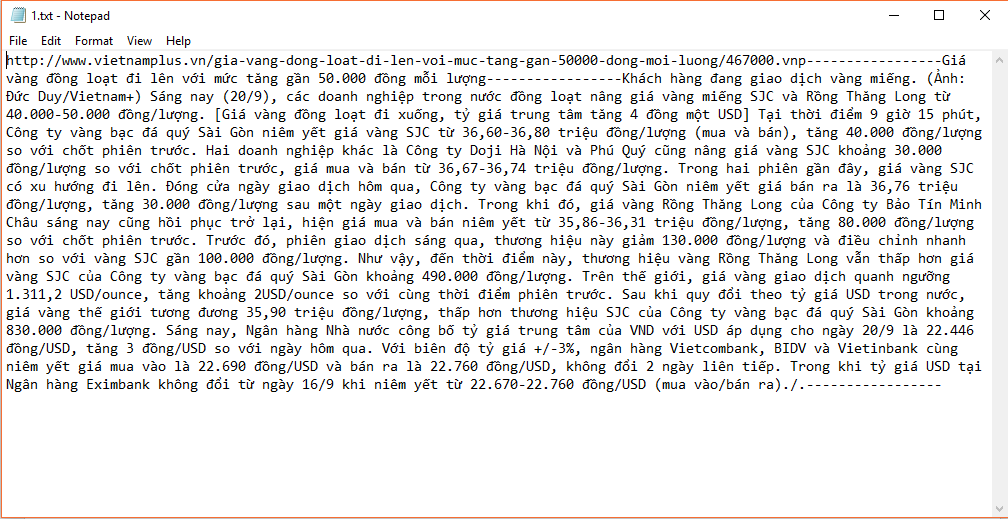
Bảng 14: Tổng kết quả thu thập dữ liệu song ngữ trang http://baothainguyen.org.vn

Từ kết quả ở bảng 14 cho thấy dữ liệu song ngữ của trang http://baothainguyen.org.vn là hoàn toàn không tương đồng. Do trang báo này chỉ có duy nhất một danh mục là chuyên mục tiếng Trung Quốc nên trang này không có bảng so sánh dữ liệu song ngữ đã phân loại. Trang báo này hoàn toàn không thích hợp là dữ liệu đầu vào của bài toán dữ liệu song ngữ.

Như hình 2 và hình 3, chúng tôi lưu dữ liệu bằng tập tin văn bản và lưu lại đường link, tiêu đề, và nội dung của bài báo, mỗi phần nội dung được phân cách bằng chuỗi “------------------------“, để sau này tiện cho quá trình phân tích, rút trích ra dữ liệu song ngữ.



Hình 2: Dữ liệu tiếng Trung sau khi thu thập.



Hình 3: Dữ liệu tiếng Việt sau khi thu thập.

1. **Kết luận**

Hiện tại chúng tôi đã thử nghiệm thành công trên 7 trang báo song ngữ. Để có một nguồn dữ liệu đầu vào an toàn cho dữ liệu song ngữ, chúng tôi sẽ thử nghiệm trên nhiều trang web có dữ liệu song ngữ hơn.Ngoài ra, chúng tôi sẽ đưa ra hướng dẫn tùy chỉnh cho cộng đồng, nếu ai muốn lấy dữ liệu từ một trang web bất kỳ họ có thể tự tùy chỉnh và lấy dữ liệu dựa nền ứng dụng chúng tôi đã tạo ra sẵn.

Trong bài báo này, chúng tôi đã đưa ra cách lấy thu thập dữ liệu dựa trên thư viện mã nguồn mở có tên là JSOUP. Việc lấy dữ liệu tự động này phù hợp với các nhóm nghiên cứu dịch máy mà không có chuyên gia về ngôn ngữ hoặc muốn làm giảm thời gian và chi phí, nhưng dữ liệu đầu vào sẽ không hoàn hảo như là một tập tin tiếng Việt và một bản dịch tập tin tiếng Trung, vấn đề dữ liệu không hoàn hảo là khó tránh khỏi.

Với kết quả thực nghiệm, chúng tôi dự định sẽ chọn ra một vài trang báo có các nội dung tiếng Trung và tiếng Việt không quá chênh lệch về số lượng tập tin và kích thước của nội dung để làm đầu vào cho việc phân tích, sàng lọc ra các tập tin, bài báo có nội dung tương đồng, từ đó chúng tôi có thể thu thập dữ liệu song ngữ.

**Tài liệu tham khảo**

[1] *Automatic data collection on the Internet (web scraping)* VERSION 18 May 2015

[2] Giulio Barcaroli, Alessandra Nurra, Marco Scarnò, Donato Summa Istituto Nazionale di Statistica Cineca, *Use of web scraping and text mining techniques in the Istat survey on “Information and Communication Technology in enterprises”*

[3] *https://jsoup.org/*

[4] Charmaine Bonifacio, Thomas E. Barchyn, Chris H. Hugenholtz, Stefan W. Kienzle, *CCDST: A free Canadian climate data scraping tool*

n, Stefan W. Kienzle

[5] Salim Khalil, Mohamed Fakir, *RCrawler: An R package for parallel web crawling and scraping*

[6] [LasseJohansson](http://www.sciencedirect.com/science/article/pii/S1364815214003478" \l "!), [VictorEpitropou](http://www.sciencedirect.com/science/article/pii/S1364815214003478" \l "!), KostasKaratzas, AriKarppinen, LeoWanner, StefanosVrochidis, AnastasiosBassoukos, JaakkoKukkonen, IoannisKompatsiaris, *Fusion of meteorological and air quality data extracted from the web for personalized environmental information services.*

1. HTML: Hypertext Markup Language [↑](#footnote-ref-1)
2. API: Application Program Interface [↑](#footnote-ref-2)
3. DOM: Document Object Model [↑](#footnote-ref-3)
4. CSS: Cascading Style Sheets [↑](#footnote-ref-4)
5. XSS: Cross-Site Scripting  [↑](#footnote-ref-5)
6. Các nội dung đã được phân loại do các nhà phát triển website. Ví dụ, như chính trị, kinh tế, văn hóa, giáo dục. [↑](#footnote-ref-6)
7. Mỗi tập tin tương tự như hình 1 hoặc hình 2 [↑](#footnote-ref-7)
8. Tốc độ tải phụ thuộc vào nhiều nhiều yếu tố như tốc độ mạng, tốc độ server, và thời gian duy trì tốc độ liên tục [↑](#footnote-ref-8)