TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**Xây dựng tập dữ liệu cặp câu song ngữ Anh-Việt cho bài toán dịch máy**

*Người hướng dẫn*: **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **BÙI DUY KHÁNH - 51303082**

Lớp **: 13050303**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2018**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**Xây dựng tập dữ liệu cặp câu song ngữ Anh-Việt cho bài toán dịch máy**

*Người hướng dẫn*: **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **BÙI DUY KHÁNH - 51303082**

Lớp **: 13050303**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2018**

LỜI CẢM ƠN

Em xin chân thành cảm ơn thầy Nguyễn Chí Thiện đã tận tâm hướng dẫn em qua từng buổi gặp mặt/báo cáo quá trình. Qua đồ án này em học được nhiều vấn đề về bài toán dịch máy cũng như có cơ hội tìm hiểu sâu thêm về các ngôn ngữ lập trình như Python, Java, và các kiến thức giải thuật mới. Một lần nữa, em xin chân thành cảm ơn thầy.

Vì thời gian ngắn và kiến thức của em vẫn còn hạn chế. Do vậy, không tránh khỏi những thiếu sót, em rất mong nhận được những ý kiến đóng góp của thầy và các giảng viên khác để kiến thức của em trong lĩnh vực này được hoàn thiện hơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi và được sự hướng dẫn của TS Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Bùi Duy Khánh*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

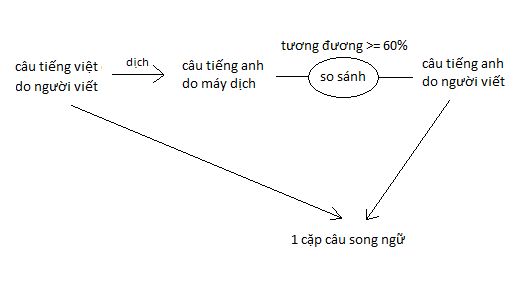
TÓM TẮT

Đồ án tìm hiểu về cách xây dựng một tập dữ liệu các cặp câu song ngữ Anh – Việt, từ hai file:

1. file văn bản tiếng Việt do người viết.
2. file văn bản tiếng Anh do người viết.

Trong Đồ án này, tác giả chọn file văn bản đó là một câu truyện, input đầu vào là 2 file text (1 tiếng Việt và 1 tiếng Anh của cùng một câu truyện), sau đó xử lý:

* Lấy một câu tiếng Việt do người viết dịch thành một câu tiếng Anh, ta được một câu tiếng Anh do máy dịch.
* Lấy câu tiếng Anh do máy dịch đó so sánh với từng câu tiếng Anh do người viết.
* Nếu độ tương đương của câu tiếng Anh đó máy dịch và câu tiếng Anh do người viết từ 60% trở lên, ta ghi nhận lại câu tiếng Việt (gốc của câu tiếng Anh do máy dịch đang so sánh) và câu tiếng Anh do người viết đó là 1 cặp câu song ngữ.



MỤC LỤC

[LỜI CẢM ƠN](#_Toc387692905) i

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN](#_Toc387692906) iii

[TÓM TẮT](#_Toc387692907) iv

[MỤC LỤC 1](#_Toc387692908)

[MỞ ĐẦU 2](#_Toc387692909)

[1. Lý do chọn đề tài 2](#_Toc387692911)

[2. Mục tiêu nghiên cứu 2](#_Toc387692911)

[3. Đối tượng 3](#_Toc387692911)

[4. Phương pháp 3](#_Toc387692911)

[5. Ý nghĩa của đồ án 3](#_Toc387692911)

[6. Cấu trú báo cáo đồ án 3](#_Toc387692911)

[CHƯƠNG 1 – NGHIÊN CỨU TỔNG QUAN 4](#_Toc387692910)

[1. Tập dữ liệu song ngữ 4](#_Toc387692911)

[2. Tìm hiểu thêm về một số kho dữ liệu song ngữ khác 4](#_Toc387692911)

[3. Một số kỹ thuật sử dụng để xây dựng tập dữ liệu cặp câu song ngữ 5](#_Toc387692911)

[4. Vai trò của Xử lý ngôn ngữ tự nhiên 8](#_Toc387692911)

[CHƯƠNG 2 – GIẢI PHÁP XÂY DỰNG TẬP DỮ LIỆU CẶP CÂU SONG NGỮ. 10](#_Toc387692917)

[1. Giới thiệu 10](#_Toc387692911)

[2. Mô hình tổng thể 10](#_Toc387692911)

[3. Xây dựng tập dữ liệu song ngữ 11](#_Toc387692911)

[4. Chọn nguồn dữ liệu song ngữ 11](#_Toc387692911)

[5. Xây dựng cấu trúc tập dữ liệu cặp câu song ngữ 12](#_Toc387692911)

[CHƯƠNG 3 – PHÁT TRIỂN ỨNG DỤNG 14](#_Toc387692920)

[1. Giai đoạn thu thập dữ liệu 14](#_Toc387692911)

[2. Giai đoạn xử lý dữ liệu đầu vào 14](#_Toc387692911)

[3. Giai đoạn phân tích, ghép các cặp câu song ngữ 15](#_Toc387692911)

[4. Lựa chọn độ tương đương phù hợp 16](#_Toc387692911)

[5. Code và thiết kế giao diện 20](#_Toc387692911)

[CHƯƠNG 4 – KHAI THÁC TẬP DỮ LIỆU SONG NGỮ 24](#_Toc387692920)

[CHƯƠNG 5 – KẾT QUẢ ĐẠT ĐƯỢC 26](#_Toc387692920)

[CHƯƠNG 6 – DEMO ỨNG DỤNG VÀ HƯỚNG DẪN SỬ DỤNG 27](#_Toc387692920)

MỞ ĐẦU

1. Lý do chọn đề tài.

Với sự phát triển của xã hội thế giới và Internet toàn cầu hiện nay, lượng thông tin để khai thác là vô cùng lớn bao gồm những thông tin đã và đang được tạo ra hằng ngày, và để những thông tin đó mang tín linh hoạt, dễ dàng tiếp cận trên nhiều quốc gia thì phần lớn các thông tin đều là tiếng Anh – ngôn ngữ quốc tế được sữ dụng rộng rãi nhất trên thế giới hiện nay. Nhưng điều đó cũng làm cho một số người gặp khó khăn trong việc tiếp cận thông tin hay nói cách khác là nó trở thành một rào cản ngôn ngữ với họ, và đó là một vấn đề rất nghiêm trọng đối với xã hội.

Để giải quyết vấn đề đó, trên thế giới đã có những nghiên cứu về xử lý ngôn ngữ tự nhiên, các nghiên cứu đó đòi hỏi một kho tài liệu song ngữ rất lớn. Hiện nay trên thế giới có rất nhiều kho tài liệu song ngữ nhưng hầu như là ngôn ngữ của các quốc gia lớn như là: Anh – Pháp, Anh – Hoa,.Tuy nhiên, đối với tiếng việt hiện nay thì có rất ít kho hoặc kho có rất ít dữ liệu đáp ứng cho người sử dụng.

Đễ góp phần giải quyết vần đề trên, tôi xin chọn đề tài: **“*Xây dựng tập dữ liệu cặp câu song ngữ Anh-Việt cho bài toán dịch máy”.***

1. Mục tiêu nghiên cứu.

Xây dựng kho dữ liệu chứa các cặp câu song ngữ Anh – Việt từ các nguồn tài liệu khác nhau như: web, từ điển, sách, văn bản,…và tìm hiểu thêm các nguồn tài liệu song ngữ khác.

Thu thập các nguồn dữ liệu song ngữ Anh – Việt, nghiên cứu các giải pháp xây dựng tập dữ liệu cặp câu song ngữ Anh – Việt, để tạo ra một cơ sở dữ liệu cho các công việc như: học tiếng Anh, dịch tự động, nghiên cứu xử lý ngôn ngữ tự nhiên,…

1. Đối tượng.

Các nguồn tài liệu có thể xây dựng tập dữ liệu cặp câu song ngữ, các cơ sở dữ liệu của kho dữ liệu cặp câu song ngữ.

1. Phương pháp.

Thu thập các nguồn tài liệu song ngữ, sử dụng các kỹ thuật tách câu từ một văn bản, câu truyện, bộ truyện,…

Tìm hiểu cách xây dựng một tập dữ liệu cặp câu song ngữ Anh – Việt.

1. Ý nghĩa của đề tài.

Tập dữ liệu song ngữ cặp câu song ngữ Anh - Việt là tài nguyên có giá trị trong việc góp phần tạo ra được một sơ sở dữ liệu phục vụ cho việc học tiếng Anh, dịch tự động, nghiên cứu xử lý ngôn ngữ tự nhiên,…

1. Cấu trúc báo cáo đồ án.

Báo cáo đồ án được tổ chức thành 3 chương.

Chương 1. Nghiên cứu tổng quan. Khái niệm tập dữ liệu cặp câu song ngữ và ứng dụng của nó, tìm hiểu thêm về một số kho dữ liệu song ngữ khác, tìm hiểu một số thuật toán về xử lý ngôn ngữ tự nhiên,…

Chương 2. Giải pháp xây dựng tập dữ liệu cặp câu song ngữ.

Chương 3. Phát triển ứng dụng. Kết quả xây dựng tập dữ liệu cặp câu song ngữ.

**CHƯƠNG 1 – NGHIÊN CỨU TỔNG QUAN**

1. Tập dữ liệu song ngữ.

*Khái niệm:*

Tập dữ liệu song ngữ là một tập các cặp văn bản song ngữ (báo, truyện,..), trong đó mỗi ngôn ngữ là bản dịch của ngôn ngữ kia.

*Ứng dụng:*

So sánh các điểm tương đồng và khác biệt giữa các ngôn ngữ.

Làm nguồn dữ liệu, tài liệu sư phạm phong phú, thiết kế giáo trình dạy học ngoại ngữ.

Giúp phiên dịch để tìm ra sự tương đương giữa ngôn ngữ nguồn và đích, để phát triển các chiến lược dịch thuật.

1. Tìm hiểu thêm về một số kho dữ liệu song ngữ khác.
2. British National Corpus (BNC)

Kho ngữ liệu 100.000.000 từ được lấy từ các mẫu văn bản từ nhiều nguồn. Phần ngôn ngữ viết của BNC (90%) được lấy từ các tờ báo, các tạp chí,... Phần ngôn ngữ nói (10%) bao gồm phiên âm chữ viết của các cuộcc hội thoại không chính thức và ngôn ngữ nói.

1. Canadian Hansard Corpus (Anh – Pháp)

Kho ngữ liệu với khoảng 90 triệu từ Anh – Pháp, là ngữ liệu song song nổi tiếng được trích từ các văn bản của Quốc hội Canada, đã được xuất bản bằng ngôn ngữ chính thức tại Canada là tiếng Anh và tiếng Pháp.

1. JENAAD Japanese-English Parallel Corpus (Anh-Nhật)

Kho ngữ liệu Japanese - English News Article Alignment Data (JENAAD) chứa hơn 150.000 cặp câu. Nguồn gốc của kho ngữ liệu được lấy từ Yomiuri Shimbun, một trong những tạp chí quốc gia của Nhật Bản, và tờ báo tiếng Anh Daily Yomiuri.

1. PKU 863 (Anh - Trung) của Đại học Bắc Kinh

Kho ngữ liệu song song Anh - Trung PKU trong Dự án 863 của Viện Ngôn ngữ học Tính toán của Trường đại học Peking. Kho ngữ liệu gồm có hơn 200.000 liên kết những cặp câu được lấy từ những văn bản song ngữ có chất lượng (3.066.435 từ tiếng Anh và tiếng Trung Quốc), bao gồm nhiều thể loại và lĩnh vực.

1. Một số kỹ thuật sử dụng để xây dựng tập dữ liệu cặp câu song ngữ.
2. Thu thập dữ liệu.

Các kho dữ liệu song ngữ hiện nay thường được chọn lọc từ nhiều nguồn khác nhau như: sách báo, website, truyện, từ điển,…

* *Nguồn từ điển*: ở mỗi mục trong từ điển, thường có ví dụ, hướng dẫn sử dụng từ đó bằng tiếng Anh cũng được dịch chính xác sang tiếng Việt.
* *Nguồn Internet*: đây là nguồn dữ liệu khổng lồ, nguồn dữ liệu này có lợi thế là chúng đã tồn tại sẵn dưới dạng điện tử, nhưng chỉ có một số ít các trang Web song ngữ là đáp ứng được đúng tiêu chuẩn.
* *Nguồn sách*: các sách dạy tiếng Anh, mẫu câu tiếng Anh, sách song ngữ,…

1. Xử lý ngôn ngữ tự nhiên

Sau khi thu thập dữ liệu, cần phải có một số công đoạn xử lý dữ liệu đầu vào phân tích, tách đoạn, tách câu,…

* *Xử lý đầu vào*: dữ liệu đầu vào sẽ được làm sạch, xoá những phần không cần thiết, nội dung từ các trang web sẽ được rút trích lưu dưới dạng text.
* *Tách đoạn*: Tách dữ liệu đầu vào từ một văn bản thành các đoạn, xem văn bản là một khối liên tục các đoạn.
* *Tách câu*: Trong văn bản tiếng Anh, tiếng Việt, thông thường người ta dùng dấu chấm (.), chấm than (!), chấm hỏi (?) và một số dấu chấm câu khác để nhận biết kết thúc câu. Tuy nhiên do tính nhập nhằng của dấu báo hiệu kết thúc câu nên việc phân định ranh giới không đơn giản. Ví dụ dấu chấm có thể biểu thị cho một dấu thập phân (1,234.567), một cụm từ viết tắt (Mr., Dr., GS., TS., …), kết thúc câu văn và một số trường hợp như địa chỉ trang web, email… …(www.udn.org hoặc abc@yahoo.com.vn). Dấu chấm hỏi hay dấu chấm thang có thể xuất hiện trong dấu ngoặc đơn, ngoặc kép hay ở cuối câu.
* *So sánh độ tương đồng chuỗi*: Hiện nay trong hầu hết ngôn ngữ lập trình chuỗi là kiểu khó so sánh độ tương đồng nhất. Độ tương đồng ngữ nghĩa giữa các câu đóng một vai trò quan trọng trong các nghiên cứu về xử lý văn bản. Nó được sử dụng như là một tiêu chuẩn của trích chọn thông tin nhằm tìm ra những tri thức ẩn trong các cơ sở dữ liệu văn bản hay trên các kho dữ liệu trực tuyến. Hiện nay tồn tại một số phương pháp tính độ tương đồng giữa các câu, điển hình là các phương pháp dựa trên tính toán thống kê và các phương pháp dựa trên quan hệ ngữ nghĩa giữa tập các từ trong các câu đó.

1. Các thuật ngữ trong xử lý ngôn ngữ tự nhiên

Vai trò của [Xử lý ngôn ngữ tự nhiên - XLNNTN (Natural Language Processing-NLP)](https://ongxuanhong.wordpress.com/2015/08/13/xu-ly-ngon-ngu-tu-nhien-la-gi/) trong khai thác Big Data là không thể phủ nhận trong bối cảnh phát triển của doanh nghiệp hiện nay. Đối với ngôn ngữ tiếng Anh, ta đã được kế thừa nhiều tri thức cũng như nhiều công cụ có sẵn để áp dụng ngay vào thực tiễn. Tuy nhiên, đối với ngôn ngữ tiếng Việt, ta vẫn còn gặp nhiều khó khăn (nhân sự có chuyên môn còn hạn chế, ngữ liệu để huấn luyện chưa đủ lớn) bên cạnh những cơ hội rất lớn (thị trường Việt Nam chưa được khai thác) cho những ai đam mê lĩnh vực này.

Vì vậy, ở đây tôi xin lập ra danh sách các thuật ngữ thường gặp trong NLP để tiện tham khảo cũng như nhanh chóng tra cứu sơ để tiến hành nghiên cứu ngay các tài liệu khoa học.

**Natural Language Processing (NLP)** – Xử lý ngôn ngữ tự nhiên là lĩnh vực Khoa học máy tính kết hợp giữa Trí tuệ nhân tạo (Artificial Intelligence) và Ngôn ngữ học tính toán (Computational Linguistics) nhằm tập trung xử lý tương tác giữa con người và máy tính sao cho máy tính có thể hiểu hay bắt chước được ngôn ngữ của con người. Các ứng dụng thường thấy như hiện nay là Siri, Cortana và Google Now.

**Ambiguity** – nhập nhằng (ở nhiều cấp độ: lexical – từ vựng, morphological – hình vị, syntactic – cú pháp, semantic – ngữ nghĩa, domain – lĩnh vực). Ví dụ nhập nhằng từ “đậu” đại diện cho một hành động hay “đậu” đại diện cho một loài thực vật trong câu “Con ruồi đậu mâm xôi đậu”.

**Pre-processing –** tiền xử lý dữ liệu, xử lý sơ bộ văn bản: xóa bỏ những kí tự, những mã điều khiển, những vùng không cần thiết cho hệ thống gồm: tách đoạn/câu/từ (paragraph/sentence/word segmentation), làm sạch (cleaning), tích hợp (integreation), chuyển đổi (transformation), giảm số chiều (reduction).

1. Vai trò của Xử lý ngôn ngữ tự nhiên.

Xử lý ngôn ngữ tự nhiên có vai trò hết sức quan trọng trong ngành Khoa Học Máy Tính. Nó có vô vàn ứng dụng hữu ích trong cuộc sống cũng như nghiên cứu. Chúng ta có thể điểm qua một vài ứng dụng của xử lý ngôn ngữ tự nhiên như:

* *Nhận dạng chữ viết*: Có hai kiểu nhận dạng, thứ nhất là nhận dạng chữ in, ví dụ nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc (docx) của Microsoft Word chẳng hạn. Phức tạp hơn là nhận dạng chữ viết tay, có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng và thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).
* *Nhận dạng tiếng nói*: Nhận dạng tiếng nói rồi chuyển chúng thành văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.
* *Tổng hợp tiếng nói*: Từ một văn bản tự động tổng hợp thành tiếng nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang web, nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, tổng hợp tiếng nói là sự trợ giúp tốt cho người khiếm thị, nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa robot với người.
* *Dịch tự động* *(Machine translate)*: Như tên gọi đây là chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Một phần mềm điển hình về tiếng Việt của chương trình này là Evtrans của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web [vdict.com](http://vdict.com/) mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Tháng 10 năm 2008 có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và Google, một thời gian sau đó [Xalo.vn](http://xalo.vn/) cũng đưa ra dịch vụ tương tự.
* *Tìm kiếm thông tin (Information retrieval)*: Đặt câu hỏi và chương trình tự tìm ra nội dung phù hợp nhất. Thông tin ngày càng đầy lên theo cấp số nhân, đặc biệt với sự trợ giúp của Internet việc tiếp cận thông tin trở lên dễ dàng hơn bao giờ hết. Việc khó khăn lúc này là tìm đúng nhất thông tin mình cần giữa bề bộn tri thức và đặc biệt thông tin đó phải đáng tin cậy. Các máy tìm kiếm dựa trên giao diện web như Google hay Yahoo hiện nay chỉ phân tích nội dung rất đơn giản dựa trên tần suất của từ khoá và thứ hạng của trang và một số tiêu chí đánh giá khác để đưa ra kết luận, kết quả là rất nhiều tìm kiếm không nhận được câu trả lời phù hợp, thậm chí bị dẫn tới một liên kết không liên quan gì do thủ thuật đánh lừa của các trang web nhằm giới thiệu sản phẩm (có tên tiếng Anh là SEO viết tắt của từ Search Engine Optimization). Thực tế cho đến bây giờ chưa có máy tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người trừ trang [www.ask.com](http://www.ask.com/) được đánh giá là "hiểu" được những câu hỏi có cấu trúc ở dạng đơn giản nhất. Mới đây cộng đồng mạng đang xôn xao về trang Wolfram Alpha, được hứa hẹn là có khả năng hiểu ngôn ngữ tự nhiên của con người và đưa ra câu trả lời chính xác. Lĩnh vực này hứa hẹn tạo ra bước nhảy trong cách thức tiếp nhận tri thức của cả cộng đồng.
* *Tóm tắt văn bản*: Từ một văn bản dài tóm tắt thành một văn bản ngắn hơn theo mong muốn nhưng vẫn chứa những nội dung thiết yếu nhất.
* *Khai phá dữ liệu (Data mining) và phát hiện tri thức*: Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình học tập, khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển. Ở mức độ đơn giản khi kết hợp với máy tìm kiếm nó cho phép đặt câu hỏi để từ đó công cụ tự tìm ra câu trả lời dựa trên các thông tin trên web mặc cho việc trước đó có câu trả lời lưu trên web hay không (giống như trang Yahoo! hỏi và đáp, nơi chuyên đặt các câu hỏi để người khác trả lời), nói một cách nôm na là nó đã biết xử lý dữ liệu để trả lời câu hỏi của người sử dụng, thay vì máy móc đáp trả những gì chỉ có sẵn trong bộ nhớ. *(Nguồn: Wikipedia)*

CHƯƠNG 2 – GIẢI PHÁP XÂY DỰNG TẬP DỮ LIỆU CẶP CÂU SONG NGỮ

1. Giới thiệu.

Hiện nay có rất nhiều giải pháp xây dựng tập dữ liệu cặp câu song ngữ như: Xây dựng từ nguồn dữ liệu từ điển, báo điện tử, hoặc các kho dữ liệu dược xây dựng sẵn,.... Trong Đồ Án này sử dụng giải pháp xây dựng từ nguồn dữ liệu là các bộ Truyện song ngữ.

Trong nhiều năm trở lại đây, tầm quan trọng của kho dữ liệu song ngữ được đánh giá rất cao do đó việc xây dựng một kho dữ liệu song ngữ nhằm đáp ứng nhu cầu về thông tin, về học tập, dịch thuật,… là rất cần thiết. Hiện nay với lượng thông tin trên mạng toàn cầu đa phần là tiếng Anh, tại Việt Nam số lượng kho dữ liệu song ngữ Anh – Việt không nhiều và không được phổ biến rộng rãi, do đó trong đồ án này tôi đưa ra giải pháp để xây dựng kho dữ liệu song ngữ Anh – Việt phục vụ xử lý tiếng Việt nhằm đáp ứng nhu cầu sử dụng kho dự liệu cho giảng dạy, học tập tiếng Anh, dịch máy, xử lý ngôn ngữ tự nhiên,…

1. Mô hình tổng thể

Kiến trúc tổng thể của hệ thống bao gồm những thành phần sau:

- Bộ sưu tập dữ liệu: sưu tập các nguồn dữ liệu song ngữ Anh – Việt ban đầu từ ebook, văn bản song ngữ, các trang web song ngữ, từ điển,...

- Tiền xử lý dữ liệu: có thể nhập trực tiếp dữ liệu, xử lý thủ công hoặc hệ thống, chuẩn hóa dữ liệu trước khi đưa vào kho. Việc chuẩn hóa dữ liệu là việc chuyển đổi định dạng dữ liệu thành định dạng tương thích với mục đích cửa hệ thống.

- Khai thác dữ liệu: những ứng dụng của dữ liệu song ngữ sau khi xử lý.

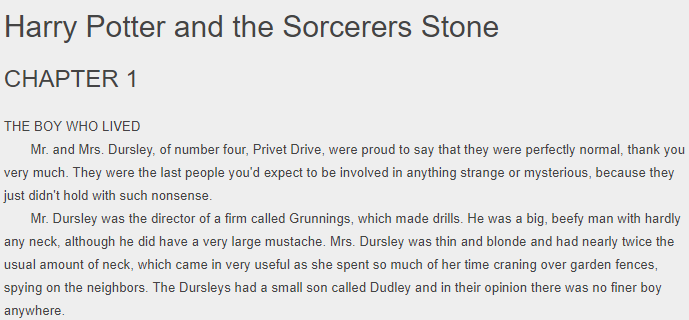
1. Xây dựng tập dữ liệu song ngữ

Để bảo đảm được hiệu quả khai thác, đúng mục tiêu nghiên cứu đã đặt ra, chúng ta cần áp dụng tuân theo một số tiêu chí trong khi xem xét lựa chọn dữ liệu đầu vào sau:

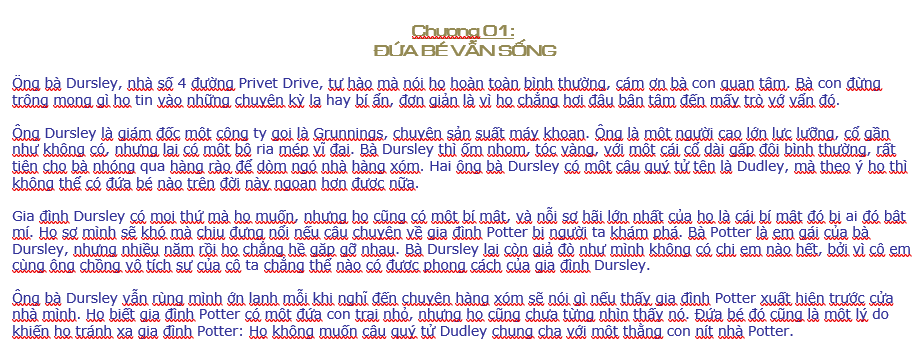
* Dữ liệu tiếng Anh, tiếng Việt phải là những câu chuẩn mực, có nghĩa, đúng ngữ pháp và được nhiều người chấp nhận, sử dụng.
* Văn bản song ngữ Anh – Việt phải do người viết, là bản dịch của nhau, không được dịch thoát ý, tóm lược hay giải thích/diễn giải.
* Các bộ Truyện song ngữ Anh – Việt được chọn làm dữ liệu đầu vào phải là những bộ Truyện dài, có thể có nhiều phần.
* Ưu tiên lựa chọn những dữ liệu song ngữ Anh – Việt nào mà đang tồn tại dưới dạng điện tử.

1. Chọn nguồn dữ liệu song ngữ.

Hiện nay rất nhiều bộ Truyện dài tập hay, nổi tiếng được đăng trên Internet với nhiều định dạng như: Pdf, Doc, mp3, mp4 , hình ảnh hoặc text.







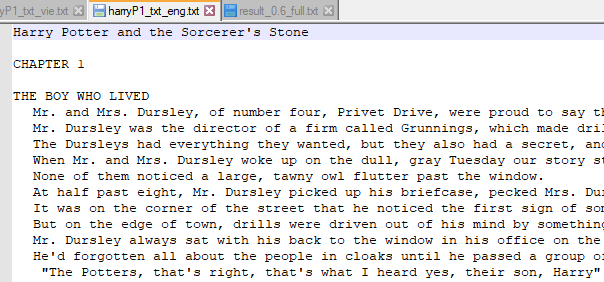
1. Xây dựng cấu trúc tập dữ liệu cặp câu song ngữ.

Về mặt tổ chức lưu trữ dự liệu tôi chọn việc lưu trư dưới dạng file text.

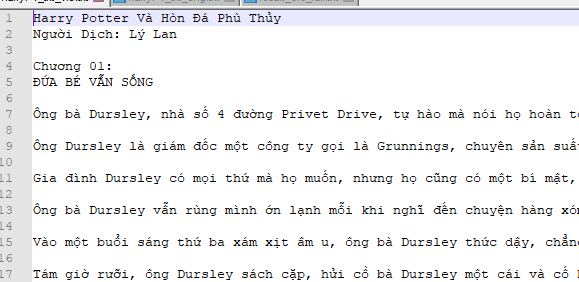
Nội dung của câu Truyện song ngữ được lấy về từ trang web sẽ được lưu thành 2 file: một file bản tiếng Anh, và một file bản tiếng Việt.

Kết quả tập dữ liệu cặp câu song ngữ được xử lý từ 2 file trên cũng được lưu trong một file text (“result.txt”).

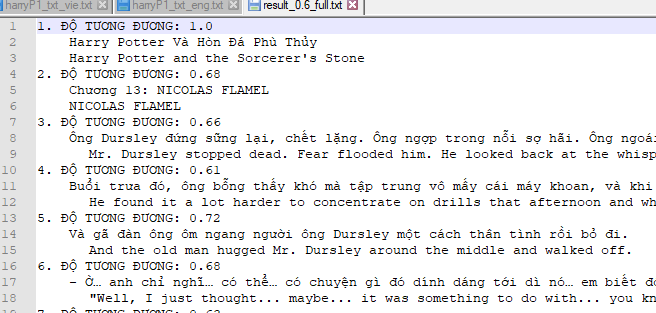
*File chứa nội dung Truyện bản tiếng Anh (tiếng Anh do người viết)*



*File chứa nội dung Truyện bản tiếng Việt (tiếng Việt do người viết)*



*File kết quả chứa các cặp câu song ngữ Anh – Việt*



CHƯƠNG 3 – PHÁT TRIỂN ỨNG DỤNG

Trong chương này trình bày các bước phát triển ứng dụng, một số kỹ thuật xử lý nguồn dữ liệu ban đầu thu thập được để xây dựng tập dữ liệu cặp câu song ngữ Anh – Việt.

1. **Giai đoạn thu thập dữ liệu**

Đầu tiên, để xây dựng được một tập dữ liệu cặp câu song ngữ đòi hỏi phải có lượng dữ liệu đầu vào lớn để có thể phân tích chính xác. Vì vậy, trong đồ án này, dữ liệu đầu vào là Truyện song ngữ là một bộ truyện dài tập, nhiều phần, là một bộ truyện nổi tiếng được viết bởi tác giả lừng danh và được dịch bởi người/đội dịch chuyên nghiệp.

Trong đồ án này bộ Truyện được chọn là bộ truyện Harry Potter tiếng Anh của tác giả *J. K. Rowling,* bản dịch tiếng Việt của dịch giả *Lý Lan*, được tài về từ website truyện song ngữ, nội dung được lưu vào file text với mỗi dòng trong file text là một đoạn của câu truyện, ta được dữ liệu đầu vào là: một file *harryP1\_txt\_eng.txt* chứa nội dung tiếng Anh do người viết, một file *harryP1\_txt\_vie.txt* chứa nội dung tiếng Việt do người viết.

1. **Giai đoạn xử lý dữ liệu đầu vào**

Ở giai đoạn này tôi dùng ngôn ngữ Python 3.6 (gọn nhẹ và hỗ trợ rất tốt trong việc xử lý ngôn ngữ tự nhiên) để tiến hành xử lý dữ liệu đầu vào, trên môi trường code là SublineText 3.

Quá trình xử lý dữ liệu đầu vào như sau:

* Mở file *harryP1\_txt\_vie.txt* chứa nội dung tiếng Việt.
* Đọc và lưu nội dung vào danh sách *list\_vie[],* mỗi phần tử trong list là một dòng trong file text.
* Tiếp theo xoá các phần không cần thiết như: ký hiệu xuống dòng.
* Sử dụng hàm *.split()* để tách đoạn và câu, lưu các đoạn và câu mới vào danh sách *list\_doan\_vie[]* để bắt đầu xử lý phân tích, ghép các cặp câu.
* Thực hiện các bước trên cho file *harryP1\_txt\_eng.txt* chứa nội dung tiếng Anh, cuối cùng ta được danh sách *list\_doan\_eng[].*

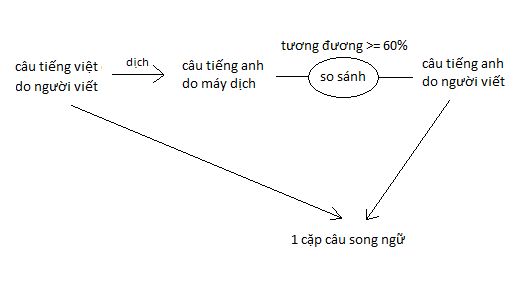
1. **Giai đoạn phân tích, ghép các cặp câu song ngữ**

Sau khi đã có 2 danh sách: *list\_doan\_vie[]* chứa các câu và đoạn tiếng Việt **(do người viết)** và *list\_doan\_eng[]* chứa các câu và đoạn tiếng Anh **(do người viết)**.

Ta lấy từng câu/đoạn trong *list\_doan\_vie[]* dịch sang tiếng Anh, ta được một câu/đoạn tiếng Anh **do máy dịch.**

Đem câu/đoạn tiếng Anh **do máy dịch** đó so sánh với từng trong *list\_doan\_eng[]* **(do người viết)**.

Nếu độ tương đương đạt từ 60% trở lên, ta ghi nhận lại câu tiếng Việt **do người viết**(gốc của câu tiếng Anh do máy dịch đang so sánh) và câu tiếng Anh **do người viết** đó là 1 cặp câu song ngữ.



Tiếp tục như vậy, cứ lấy từng câu/đoạn trong *list\_doan\_vie[]* dịch sang tiếng Anh, rồi so sánh với từng câu/đoạn trong *list\_doan\_eng[]* để tìm ra các cặp câu song ngữ tương đương với nhau, lưu kết quả ghi nhận được vào file *result\_0.6\_full.txt*.

1. **Lựa chọn độ tương đương phù hợp**

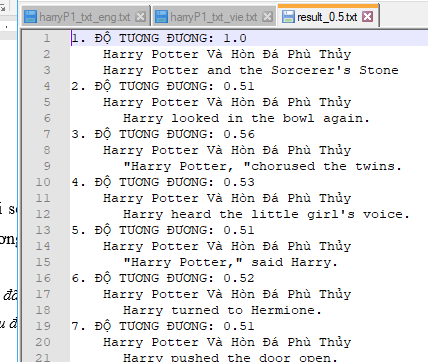
Độ tương đương ở đây, chính là mức độ giống nhau về mặt hình thức và ngữ nghĩa khi tiến hành so sánh một câu/đoạn tiếng Anh **do máy dịch** với một câu/đoạn tiếng Anh **do người viết**.

Nó cho biết liệu một câu/đoạn tiếng Anh được dịch bằng máy giống chính xác bao nhiêu phần trăm với câu/đoạn tiếng Anh do người viết.

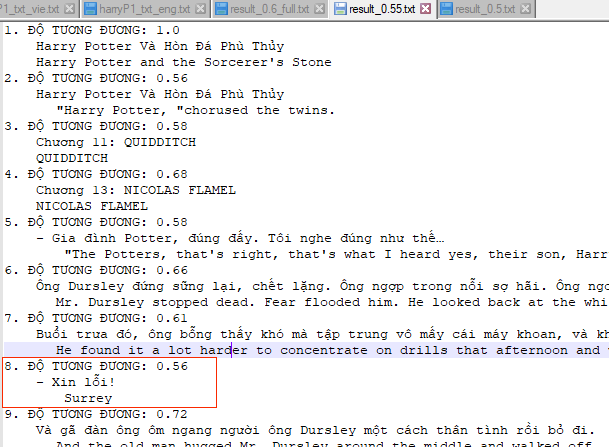
Trong Đồ án này tôi lựa chọn độ tương đương là từ 60% trở lên.

Vì sao lại chọn độ tương đương là từ 60% trở lên: trong quá trình thực hiện ứng dụng, tôi đã có chạy thử với các độ tương đương khác nhau (như >=50%, >=60%, >=65%, >=70%), sau đó theo quan sát thì:

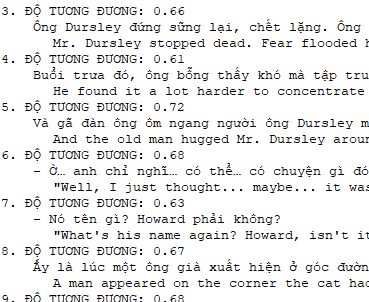
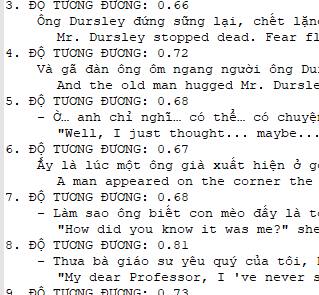
* *Với độ tương đương >=50%: số lượng câu/đoạn bị nhập nhằng là rất lớn.*



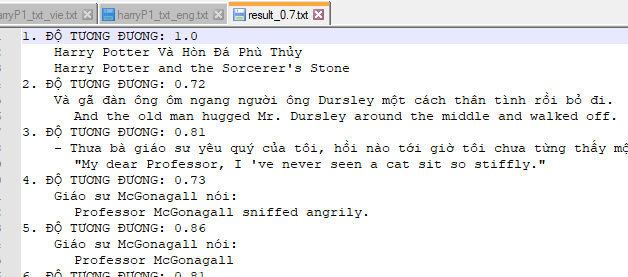
* *Với độ tương đương >=55%: số lượng câu/đoạn bị nhập nhằng giảm khá đáng kể nhưng vẫn còn một số câu sai nghĩa.*



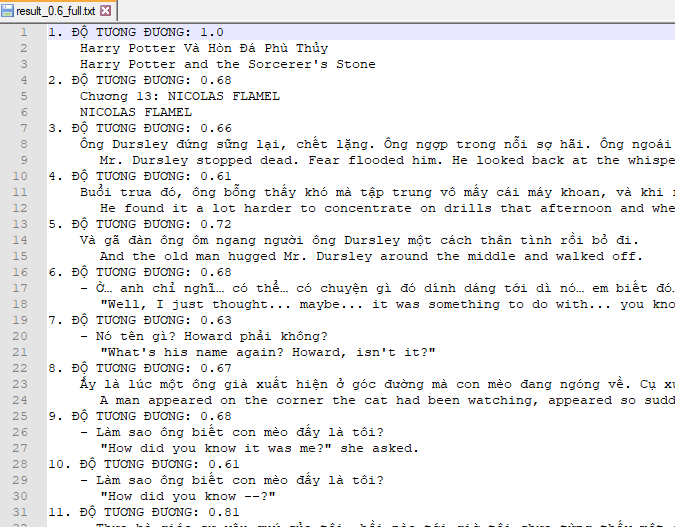
* *Với độ tương đương >=65%: một số câu tương nhau bị bỏ qua vì độ tương đương ở mức 61->64%.*

* *Với độ tương đương >=70%: số lượng câu/đoạn tương đương bị bỏ qua vô cùng lớn.*



* *Còn lại với độ tương đương >=60%, các tình trạng bị nhập nhằng, bỏ sót các câu/đoạn tương đương được giảm thiểu đáng kể, còn lại con số rất ít có thể chấp nhận được. Nên tôi quyết định chọn độ tương đương cho ứng dụng là >=60%.*

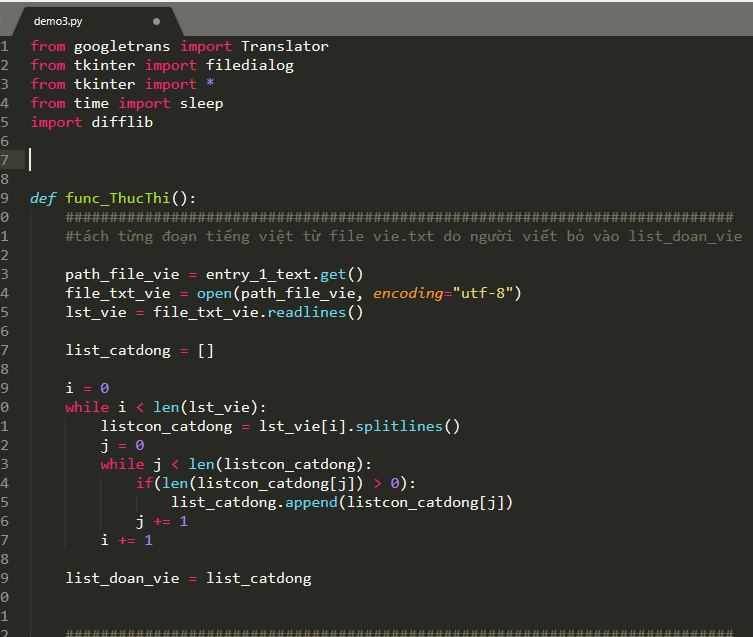


1. **Code và thiết kế giao diện**
2. Code

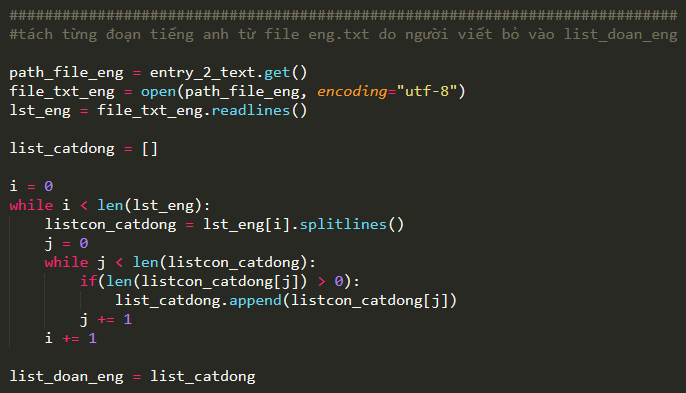
*Tại sao lại sử dụng Python trong xử lý ngôn ngữ tự nhiên?*

Python ra đời năm 1991, và là một ngôn ngữ thông dịch. Trải qua hơn 20 năm phát triển, Python là một trong những ngôn ngữ được sử dụng nhiều nhất trong dậy lập trình và nghiên cứu khoa học.  Rất nhiều trường đại học sử dụng Python để dậy về lập trình cho các sinh viên ngành Khoa Học Máy Tính. Rất nhiều công ty lớn sử dụng Python để xây dựng hệ thống như Google, Youtube, Instagram, Dropbox, Atlassian... Python là một ngữ sử dụng được cho nhiều mô hình lập trình, đơn giản khi học và sử dụng. Tôi sử dụng Python chưa lâu nhưng khi so sánh việc Code sử dụng Pythong thì nó ngắn hơn rất nhiều so với khi viết bằng PHP hoặc Java. Bạn có thể bay bổng tự do với Python hoặc cũng có thể bắt nó trở lên vững chắc và mạnh mẽ như Java. Theo những thông tin mà tôi được biết thì Python cũng là một ngôn ngữ rất phát triển trong lĩnh vực Data Science và Machine Learning. Python cũng cung cấp những hàm và thư viện xử lý ngôn ngữ tuyệt vời. Scikit-learn và Tensor-flow là 2 thư viện Machine Learning nổi tiếng được viêt bằng Python. Đứng ở góc độ người tiếp cận sau, cá nhân tôi thấy Python là một lựa chọn hợp lý khi làm Xử Lý Ngôn Ngữ Tự Nhiên.

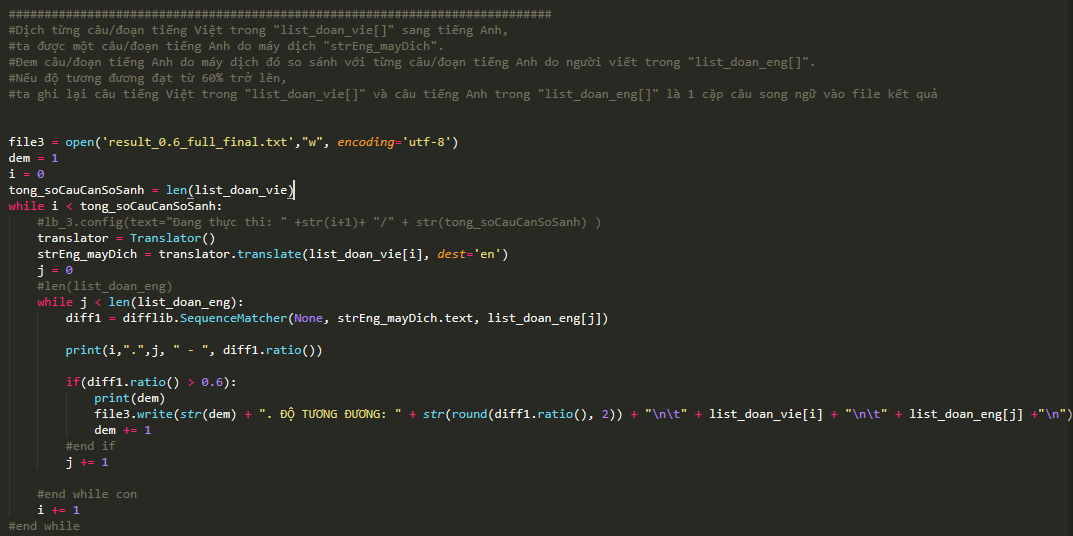
*Import các thư viện hỗ trợ cần thiết để xử lý ngôn ngữ tự nhiên và xử lý tách đoạn cho dữ liệu đầu vào bản tiếng Việt.*



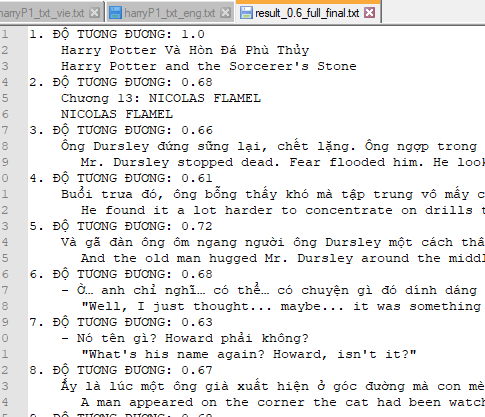
*Xử lý tách đoạn cho dữ liệu đầu vào bản tiếng Việt*



*Tiến hành dịch, so sánh và ghi file*



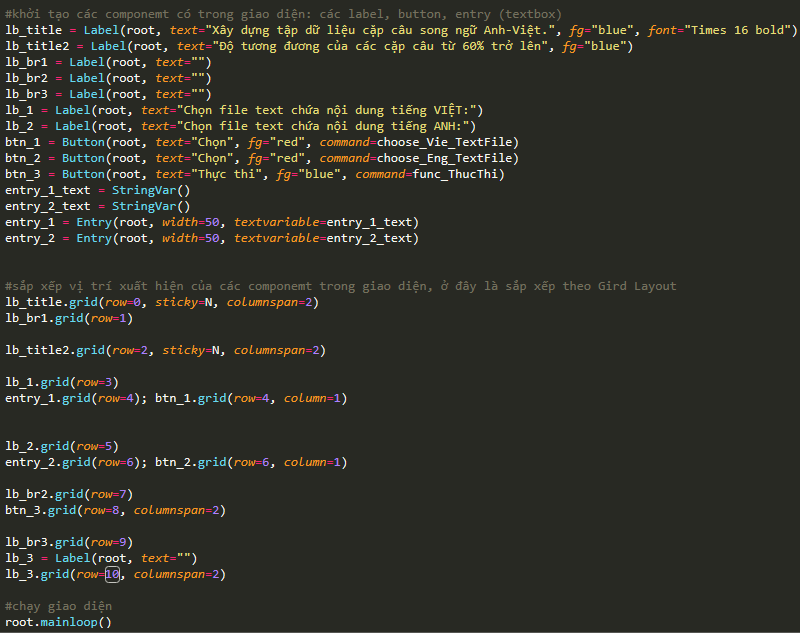
*Cấu trúc file kết quả (ảnh minh hoạ).*



1. Giao diện ứng dụng

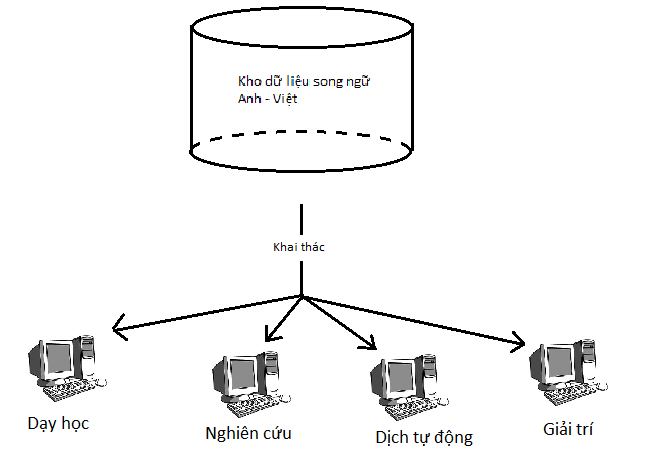
*Giới thiệu về Tkinter:* Tkinter là một gói trong Python có chứa module Tk hỗ trợ cho việc lập trình GUI. Tk ban đầu được viết cho ngôn ngữ **Tcl**. Sau đó Tkinter được viết ra để sử dụng Tk bằng trình thông dịch Tcl trên nền Python. Ngoài Tkinter ra còn có một số công cụ khác giúp tạo một ứng dụng GUI viết bằng Python như wxPython, PyQt, và PyGTK.

*Code xây dựng giao diện ứng dụng:*



CHƯƠNG 4 – KHAI THÁC TẬP DỮ LIỆU SONG NGỮ

Tập dữ liệu cặp câu song ngữ Anh – Việt chủ yếu được dùng để xây dựng hệ thống dịch tự động, trong nghiên cứu và rất hữu ích trong giáo dục. Tập dữ liệu song ngữ cặp câu Anh – Việt được khai thác trong việc học và giảng dạy ngoại ngữ, trong các trò chơi nhằm trau dồi vốn tiếng Anh. Ngoài ra, các kho dữ liệu song ngữ trên thế giới còn được sử dụng làm nguồn dữ liệu để sử dụng trong biên soạn phụ đề phim, trong việc xây dựng từ điển, hỗ trợ cho phiên dịch viên,…Với các nhà nghiên cứu, thì các kho dữ liệu song ngữ Anh – Việt được sử dụng trong việc tìm kiếm nghĩa các từ được dịch trong các câu với nhiều bối cảnh khác nhau.



Dữ liệu nói chung và dữ liệu song ngữ Anh-Việt nói riêng sẽ giúp chúng ta rất nhiều trong vô vàn các ứng dụng khác nhau. Từ lĩnh vực Ngôn ngữ học so sánh đối chiếu, cho đến việc giảng dạy ngoại ngữ nói chung, môn biên dịch nói riêng, ngữ liệu song ngữ Anh-Việt sẽ giúp cho người học tự “nghiệm” ra các quy luật chuyển ngữ mà các cách tiếp cận truyền thống không thể bao quát hết được. Đặc biệt trong công tác dịch thuật chuyên nghiệp, kho ngữ liệu Anh – Việt này sẽ làm giảm đáng kể công sức dịch thuật thủ công. Nếu kho ngữ liệu này được tiếp tục cập nhật thì hiệu quả khai thác càng tăng gấp bội. Ngoài ra, cách thức này có thể áp dụng cho các cặp ngôn ngữ khác.

Qua phân tích các ứng dụng nói trên của song ngữ, ta thấy kho dữ liệu có thể khai thác phục vụ cho cả hai mục đích: về lý thuyết (giảng dạy dịch thuật) lẫn thực hành (công tác dịch thuật). Vì vậy, chúng ta cần tăng cường số lượng (thu thập thêm nhiều văn bản song ngữ), chủng loại (nhiều lĩnh vực khác nhau) và xử lý sâu (gán nhãn ngôn ngữ); đồng thời xây dựng các công cụ chuyển đổi tự động.

Ngoài ra còn có các công có thể hỗ trợ các thao tác khác của người dịch, như: giữ nguyên định dạng tập tin nguồn, tra cứu thuật ngữ, kiểm lỗi chính tả, kiểm tra tính nhất quán việc dịch các thuật ngữ, quản lý dự án dịch với nhiều cộng tác viên từ xa, … Tuy nhiên, các công cụ này hiện nay chủ yếu hoạt động hiệu quả khi phân tích tiếng Anh (và một số tiếng thông dụng ở Châu Âu), còn khi phân tích tiếng Việt sẽ có nhiều hạn chế do đặc thù ngôn ngữ tiếng Việt, vì vậy, cần có sự can thiệp (tiền xử lý) vào ngữ liệu trước khi sử dụng hoặc phải sử dụng phần mềm đặc thù cho tiếng Việt.

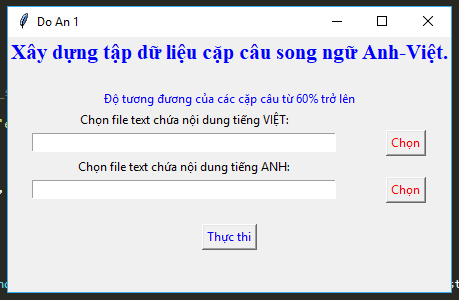
**CHƯƠNG 5 – KẾT QUẢ ĐẠT ĐƯỢC**

Trong quá trình thực hiện đồ án này tôi đã thu được nhiều kiến thức về xử lý ngôn ngữ tự nhiên , kho dữ liệu song ngữ và các vấn đề liên quan đến xử lý dữ liệu. Đồ án trình bày chi tiết các bước cơ bản để thực hiện xử lý dữ liệu ngôn ngữ tự nhiên, đưa ra các giải pháp, kỹ thuật xử lý và cập nhật kho dữ liệu song ngữ Anh – Việt. Tuy nhiên đồ án không tránh khỏi các hạn chế bao gồm: Nguồn dữ liệu song ngữ ở Việt Nam có chất lượng bản dịch không cao, đặc biệt đối với các trang web thường dịch ý, tóm lược nội dung của văn bản gốc do đó rất khó khăn cho việc lựa chọn nguồn dữ liệu cũng như canh đoạn, tách câu,…Việc cập nhật kho dữ liệu nói chung vẫn mang tính bán tự động, nhiều công đoạn thủ công. Chưa tìm hiểu kỹ khả năng ứng dụng của kho dữ liệu vào việc tự động mà chỉ dừng ở mức độ làm dữ liệu phục vụ việc tìm các cặp câu song ngữ tương đương nghĩa với nhau.

Qua quá trình làm đồ án, tôi xin đưa ra một số kiến nghị và hương phát triển của đồ án như sau: Do nhu cầu nghiên cứu và học tập tiếng Việt của sinh viên nước ngoài, cũng như nhu cầu học ngoại ngữ của sinh viên Việt Nam, tôi sẽ cố gắng tiếp tục bổ sung vào nguồn dữ liệu trên, không chỉ dừng lại ở 2 ngôn ngữ Anh – Việt mà còn có thể thêm nhiều ngôn ngữ khác như Pháp, Trung, Nhật, Hàn,…Cũng như tìm giải pháp tối ưu hơn để xây dựng tập dự liệu song ngữ hoàn thiện hơn.

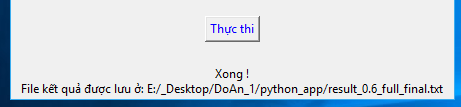
CHƯƠNG 6 – DEMO ỨNG DỤNG VÀ HƯỚNG DẪN SỬ DỤNG

* Khởi động ứng dụng.

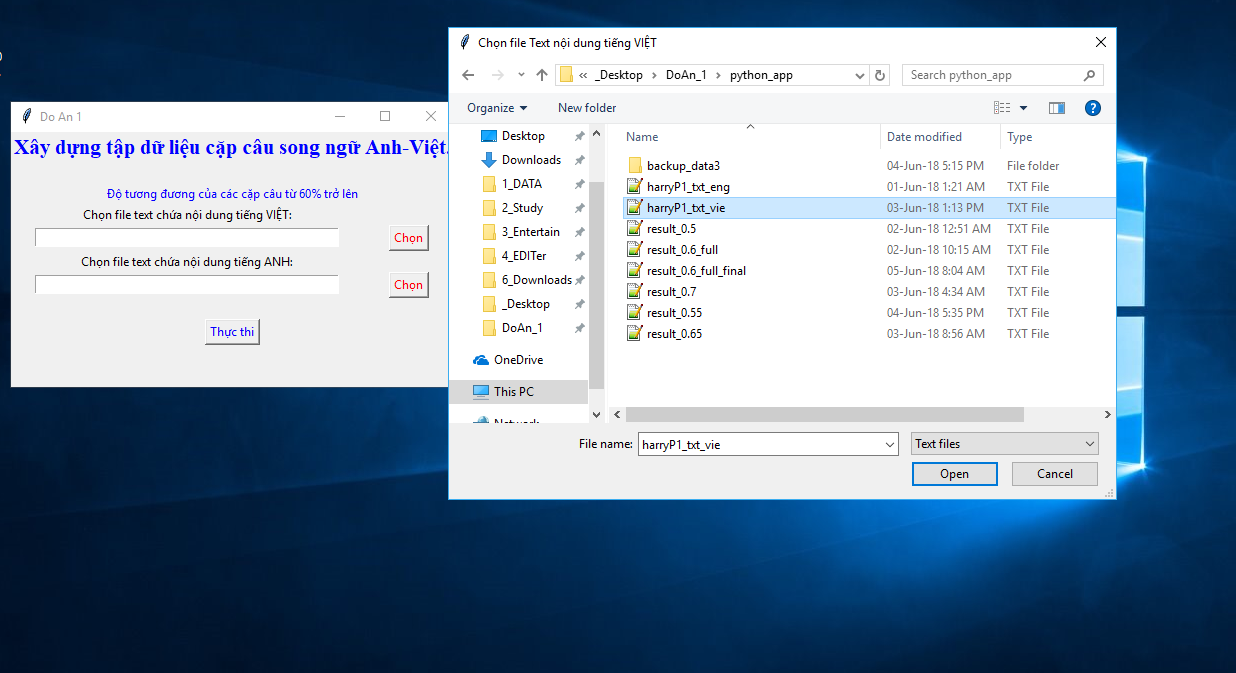


Giao diện ứng dụng hiện lên trong đó bao gồm:

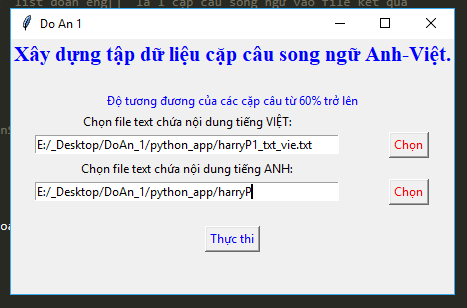
* Các Lable để hiển thị Title, tên của các ô nhập.
* Hai ô nhập file text dữ liệu đầu vào, có thể nhập trực tiếp đường dẫn đến file hoặc chọn file bằng cửa sổ Dialog.
* Hai nút “Chọn” file tương ứng cho hai ô nhập dữ liệu đầu vào.
* Cuối cùng là nút “Thực thi” để tiến hành quá trình xử lý hai file dữ liệu đầu vào.
* Ngoài ra khi Thực thi xong sẽ xuất hiện Lable thông báo hiển thị đường dẫn đến file kết quả.



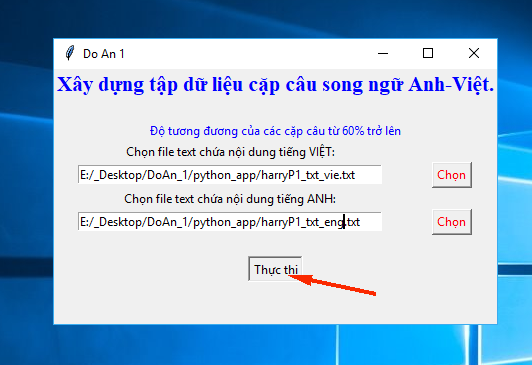
* Chọn file dữ liệu đầu vào.
* Nhấn nút Chọn bên phải ô nhập để chọn file tương ứng.
* Cửa sổ chọn file hiện ra để bạn chọn đúng file mình cần. Lưu ý, cửa sổ có tiêu đề nhắc bạn chọn đúng loại file, đúng bản ngôn ngữ của file.



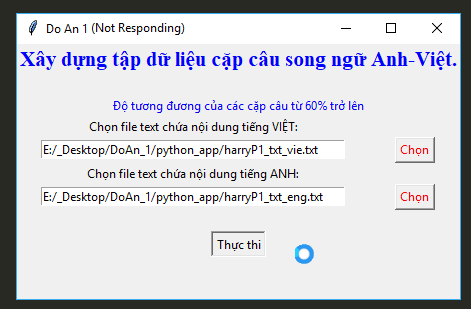
* Hoặc bạn có thể nhập trực tiếp đường dẫn đến file dữ liệu cần chọn.



* Thực thi chương trình.
  + Sau khi chọn file dữ liệu đầu vào xong, ta tiến hành chạy chương trình bằng cách nhấn vào nút “Thực thi”. Lưu ý, chương trình chỉ thực thi được khi có kết nối mạng.



* + Quá trình thực thi xử lý hai file dữ liệu đầu vào bắt đầu, lúc này bạn không thể thao tác gì trên ứng dụng.
  + Tốc độ xử lý nhanh hay chậm tuỳ thuộc vào số lượng câu/đoạn (độ dài dử liệu đầu vào) và cấu hình máy và tốc độ mạng của bạn.



* Thực thi xong.
  + Bên dưới nút “Thực thi” sẽ xuất hiện một đoạn thông báo kết quả, và đường dẫn đến file lưu kết quả thực thi.

