**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**Bùi Đức Lực**

**Nguyễn Hải Đăng**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG WEBSITE HỎI ĐÁP THÔNG TIN BÓNG ĐÁ**

**CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH**

**TP. HỒ CHÍ MINH, 2016**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**BÙI ĐỨC LỰC \_ 12520636**

**NGUYỄN HẢI ĐĂNG \_ 12520556**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG WEBSITE HỎI ĐÁP THÔNG TIN BÓNG ĐÁ**

**CỬ NHÂNNGÀNH KHOA HỌC MÁY TÍNH**

**GIẢNG VIÊN HƯỚNG DẪN**

**THS. NGUYỄN TRỌNG CHỈNH**

**TP. HỒ CHÍ MINH, 2016**

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …………………… ngày ………………….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

* 1. …………………………………………. – Chủ tịch.
  2. …………………………………………. – Thư ký.
  3. …………………………………………. – Ủy viên.
  4. …………………………………………. – Ủy viên.

**Lời cám ơn!**

Không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Được sự phân công của khoa Khoa Học Máy Tính trường Đại Học Công Nghệ Thông Tin, và sự đồng ý của Thầy giáo hướng dẫn Th.s Nguyễn Trọng Chỉnh nhóm đã thực hiện đề tài “Xây Dựng website hỏi đáp thông tin bóng đá”. Để hoàn thành tốt đề tài này, trước tiên, chúng em xin chân thành cảm ơn các thầy cô và ban giám hiệu trường ĐH Công Nghệ Thông Tin nói chung và các thầy cô trong khoa nói riêng đã tạo mọi điều kiện giúp đỡ chúng em trong quá trình học tập và nghiên cứu.

Kế đến, chúng em xin chân thành gửi lời cảm ơn đến thầy Nguyễn Trọng Chỉnh, là giáo viên hướng dẫn và cũng là người đã nhiệt tình chỉ bảo cho chúng em trong suốt quá trình hoàn thành đề tài.

Cuối cùng xin gửi lời cảm ơn đặc biệt đến gia đình luôn là một chỗ dựa vững chắc cũng như luôn tạo mọi điều kiện tốt nhất cho chúng em được học tập.

Tp Hồ Chí Minh, 10/06/2016

Nhóm sinh viên thực hiện

Bùi Đức Lực

Nguyễn Hải Đăng

MỤC LỤC

[Chương 1. TÊN CHƯƠNG 1 3](#_Toc367742496)

[1.1. Chủ đề cấp độ 2 3](#_Toc367742497)

[1.1.1. Chủ đề cấp độ 3 3](#_Toc367742498)

[1.1.2. Chủ đề cấp độ 3 3](#_Toc367742499)

[1.1.2.1. Chủ đề cấp độ 4 3](#_Toc367742500)

[Chương 2. TÊN CHƯƠNG 2 4](#_Toc367742501)

[2.1. Chủ đề cấp độ 2 4](#_Toc367742502)

[2.1.1. Chủ đề cấp độ 3 4](#_Toc367742503)

[2.1.1.1. Chủ đề cấp độ 4 4](#_Toc367742504)

[2.2. Chủ đề cấp độ 2 4](#_Toc367742505)

[2.2.1. Chủ đề cấp độ 3 4](#_Toc367742506)

[Chương 3. TÊN CHƯƠNG 3 5](#_Toc367742507)

[3.1. Chủ đề cấp độ 2 5](#_Toc367742508)

[3.1.1. Chủ đề cấp độ 3 5](#_Toc367742509)

[3.1.1.1. Chủ đề cấp độ 4 5](#_Toc367742510)

[3.2. Chủ đề cấp độ 2 5](#_Toc367742511)

DANH MỤC HÌNH VẼ

[Hình 1.1: Tên hình 1 3](#_Toc367742554)

DANH MỤC BẢNG

[Bảng 1.1: Tên bảng 1 3](#_Toc367742567)

[Bảng 2.1: Tên bảng 1 4](#_Toc367742568)

DANH MỤC TỪ VIẾT TẮT

TÓM TẮT KHÓA LUẬN

MỞ ĐẦU

Với nhu cầu trao đổi thông tin của con người ngày càng cao, thông tin tràn ngập trên mọi phương tiện truyền thông, đặc biệt là sự phát triển rộng rãi của mạng toàn cầu Internet, hằng ngày con người phải xử lý một lượng thông tin khổng lồ. Những thắc mắc của người dùng dưới dạng truy vấn sẽ được tìm kiếm và trả về một cách ngắn gọn, súc tích, chính xác nhất những gì mà họ mong muốn. Đó chính là mục tiêu của hệ thống hỏi-đáp tự động. Rất nhiều hệ thống hỏi đáp thông tin qua mạng ra đời nhằm đáp ứng nhu cầu này.

Qua báo cáo này, Nhóm sẽ trình bày những gì đã tìm hiểu được về một Hệ Thống Hỏi Đáp Tự Động. Từ đó xây dựng một Hệ Thống Hỏi Đáp Tự Động cụ thể.

# GIỚI THIỆU VẤN ĐỀ

## Giới thiệu bái toàn

Nguồn dữ liệu trên internet đang ngày một tăng lên không ngừng, cộng với nhu cầu trao đổi thông tin của con người ngày càng cao. Đặc biệt trong lĩnh vực bóng đá thì nguồn dữ liệu về thông tin bóng đá là rất lớn, và nhu cầu tìm kiếm thông tin về bóng đá của con người hàng ngày cũng nhiều không kém. Hiện nay các công cụ tìm kiếm như Bing hay Google sẽ giúp người dùng tìm kiếm được những thông tin cần thiết. Như đã biết, kết quả mà Bing Google hay một công cụ tìm kiếm nào khác hiện này trên internet, tất cả kết quả trả về sẽ là các đường link dẫn đến các trang chưa thông tin. Các trang thông tin hiện nay sẽ không tránh khỏi việc xuất hiện dữ liệu thừa không liên quan đến câu hỏi, đến thông tin người dùng mong muốn. Do đó buộc người dùng phải đọc và sàng lọc một cách kỹ càng các thông tin bên trong để tìm ra thông tin cần thiết. Nếu một trang web lớn chứ nhiều dữ liệu thì việc tìm kiếm thông tin cần thiết sẽ là rất khó khăn và sẽ mất khá nhiều thời gian, chưa kể đến việc thông tin chúng ta cần sẽ không có trong link đó và chúng ta phải chuyển sang link khác.

Vậy vấn đề đặt ra ở đây là liệu có hệ thống nào, phần mềm nào sẽ giúp chúng ra việc đó, giúp chúng ta tìm kiếm thông tin chúng ta cần một cách chính xác mà lại không mất nhiều thời gian. Hệ thống giúp chúng ta làm điều này có tên là “Hệ thống hỏi đáp”. Trong đề tài này, nhóm sẽ xây dựng một ứng dụng dạng web, thực hiện việc trả lời những câu hỏi liên quan đến thông tin bóng đá do người dùng nhập vào

## Mục tiêu

Hệ thống được xây dựng với mục tiêu sau:

* Xây dựng giao diện trực quan, dễ dàng sử dụng.
* Phân tích được dạng câu hỏi do người dùng nhập vào dưới dạng ngôn ngữ tự nhiên.
* Rút trích được thông tin cần thiết từ nguồn dữ liệu khổng lồ trên internet dựa vào câu hỏi.
* Trả lời được các dạng câu hỏi một cách chính xác và nhanh chóng. Tiết kiệm thời gian tìm kiếm thông tin cho người dùng.

## Giải pháp cho hệ thống

## Phạm vi đề tài

Câu hỏi được giới hạn là kiểu câu đơn với nội dung câu hỏi được xác định:

* Thông tin về cầu thủ: năm sinh, tên, tuổi, chiều cao, số áo…
* Thông tin về trận đấu: thời gian diễn ra, thời gian kết thúc, kết quả của trận đấu…
* Thông tin về giải đấu: thời gian khai mạc, bế mạc, địa điểm,..

# CÁC VẤN ĐỀ LIÊN QUAN

## Giới thiệu về hệ thống hỏi đáp

### Tổng quan về hệ thống hỏi đáp tự động

Hệ thống hỏi-đáp tự động (Question Answering-QA) là một hệ thống được xây dựng để thực hiện việc tìm kiếm câu trả lời cho một câu hỏi của người dùng. Hệ thống hỏi-đáp tự động liên quan đến 3 lĩnh vực lớn là xử lý ngôn ngữ tự nhiên (Natural Language Processing), tìm kiếm thông tin (Information Retrieval) và rút trích thông tin (Information Extraction).

Hệ thống hỏi đáp tự động có thể coi như một lựa chọn thứ hai bên cạnh hệ thống trích chọn thông tin khi người dùng muốn tìm kiếm thông tin họ cần. Hệ thống trích chọn thông tin nhận đầu vào là các từ khóa và trả về tập các tài liệu liên quan (có chứa các từ khóa đó). Kết quả mà hệ thống trích chọn thông tin (máy tìm kiếm) trả lại cho người dùng là rất lớn, có thể lên đến hàng nghìn trang web mà phần nhiều không chứa thông tin người dùng mong muốn. Trong khi đó, hệ thống hỏi đáp nhận đầu vào là câu hỏi dưới dạng ngôn ngữ tự nhiên của người dùng, trả lại các đoạn văn bản ngắn (các snippet) chứa câu trả lời trực tiếp cho câu hỏi. Nghiên cứu về hệ thống hỏi đáp tự động hiện đang thu hút sự quan tâm của rất nhiều các nhà nghiên cứu từ các trường đại học, các viện nghiên cứu và cả các doanh nghiệp lớn trong ngành công nghệ thông tin, có ý nghĩa khoa học lẫn ý nghĩa thực tế. Rất nhiều các hội nghị thường niên về khai phá dữ liệu, trích chọn thông tin dành một chủ đề riêng cho các nghiên cứu về hệ thống hỏi đáp như TREC1 , CLEF2 … Bài toán xây dựng hệ thống hỏi đáp là một bài toán khó thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Chúng ta biết rằng ngôn ngữ tự nhiên vốn nhập nhằng, đa nghĩa, việc xác định được ngữ nghĩa của câu hỏi cũng như phát hiện ra câu trả lời là một thách thức không nhỏ. Không những vậy, giữa câu hỏi và câu trả lời còn tồn tại các quan hệ “ngầm” hay phụ thuộc vào ngữ cảnh.

Ví dụ câu hỏi “Ai là tác giả Nhật ký trong tù ?” Câu trả lời: “ Hồ Chí Minh viết Nhật ký trong tù trong thời gian Người bị chính quyền Tưởng Giới Thạch bắt giam và giải đi khắp các nhà giam ở Quảng Tây, Trung Quốc.”

“Nhật ký trong tù, dịch tiếng Hán- Ngục trung nhật ký là một tập thơ của Hồ Chí Minh,…”. Hay câu hỏi “Mozart sinh năm nào ?” và câu trả lời “….Mozart (1751 – 1791)…”. Để tìm được câu trả lời trên cho câu hỏi, hệ thống cần có cơ chế để biết được rằng “tác giả của một tập thơ là người viết tập thơ đó” hoặc cần học được các mẫu thường gặp của câu trả lời (các mẫu về ngày tháng năm sinh, về thời gian, địa chỉ….) tương ứng với từng loại câu hỏi. Các hệ thống Q&A trên thế giới hiện nay sử dụng rất nhiều các công cụ xử lý ngôn ngữ như: Bộ gán nhãn từ loại (POS Tagger), bộ nhận dạng tên thực thể (Named Entity Recognizer), bộ phân tích ngữ pháp (Parser)… và các tài nguyên ngôn ngữ như Wordnet, ontology để phân tích câu hỏi và trích xuất câu trả lời.

### Lịch sử phát triển của hệ thống hỏi đáp

Nghiên cứu về QA đã bắt đầu từ những năm 60 của thế kỷ trước. Tuy nhiên, những khó khăn trong việc xử lý ngôn ngữ tự nhiên đã giới hạn việc nghiên cứu QA vào việc nghiên cứu các hệ chuyên gia trả lời câu hỏi trong các lĩnh vực nhỏ hẹp. Vào những năm gần đây, nghiên cứu về hệ QA tổng quát đã được đẩy mạnh cùng với sự xuất hiện của tiểu ban QA (QA Track) ở hội nghị TREC ( Text Retrieval Conference), sự thảo luận về QA ở diễn đàn CLEF (Cross Language Evaluation Forum).

Thiết kế một hệ thống hỏi-đáp không phải là khái niệm mới. Một số hệ thống đầu tiên đã được ra đời từ những năm 1960. Một ví dụ của một hệ thống như vậy là hệ thống BASEBALL được phát triển năm 1961 do nhóm tác giả Green, Chomsky, và Laughery . Hệ thống này được thiết kế để cung cấp các thông tin về các số liệu thống kê của liên đoàn bóng chày Mỹ. Một hệ thống khác tương tự như BASEBALL đã được phát triển bởi Woods (1973) và được đặt tên LUNAR. LUNAR có thể trả lời các câu hỏi liên quan đến các mẫu đá trở về từ tàu thăm dò mặt trăng Apollo. Hệ thống này dịch các câu hỏi thành một hay nhiều truy vấn cơ sở dữ liệu. Hệ thống TEAM được phát triển bởi Grosz (1983) đã có những đặc điểm cơ bản như một chuỗi các biểu diễn ngữ nghĩa và một lược đồ phiên dịch làm cho nó tốt hơn 2 hệ thống trước. Điểm tương đồng giữa cả ba hệ thống là tất cả đều sử dụng cơ sở dữ liệu để lưu trữ cơ sở tri thức. Việc thiết kế của cơ sở dữ liệu này và các dữ liệu có cấu trúc được xây dựng một cách thủ công bởi tất cả các chuyên gia trong các lĩnh vực tương ứng.

Xử lý ngôn ngữ tự nhiên trong quá trình thực hiện đã nhận được rất nhiều sự quan tâm trong những năm 1970. Một trong những hệ thống như thế là Margie (Schank, Goldman, Riesbeck, & Rieger, 1975). Hệ thống này đã có thể xử lý các tài liệu mà sau này có thể trả lời các câu hỏi cơ bản. Điều này được thực hiện bằng cách phân tích và tổ chức tài liệu của thông tin trong một mô hình tương tự như bộ nhớ của con người. Điều này đã được cải thiện hơn bởi Lenhert, Dyer, Johnson, Yang, và Harley (1983) trong hệ thống Boris. Boris và Margie là các hệ thống gần giống nhất với hệ thống hỏi-đáp lĩnh vực rộng (open-domain) hoạt động ngày nay.

Hệ thống hỏi-đáp nhận được rất nhiều mối quan tâm của cộng đồng các nhà nghiên cứu. Điều này bắt nguồn từ việc một lượng tài liệu lớn có gắn chỉ mục sẵn có trên Internet, cùng với các thành tựu trong lĩnh vực rút trích thông tin và sự gia tăng nhu cầu thương mại cho các sản phẩm QA trên toàn cầu. Một lợi thế nữa, dễ dàng cho việc phát triển các hệ thống QA phức tạp hơn, là các nguồn tài nguyên như là WordNet (Fellbaum, 1998) và OpenCyc (Curtis, Matthews, & Baxter, 2005) . WordNet cung cấp sự truy cập tới các thông tin ngữ nghĩa và các mối liên hệ ngữ nghĩa giữa các cụm từ. OpenCyC cung cấp sự truy cập tới thông tin ontology và tri thức thông dụng. Cả 2 hệ thống này đều được cung cấp miễn phí. Hội thảo TREC (Text Retrieval and Evaluation Conference) về lĩnh vực QA năm 1999 đã đóng góp một phần to lớn cho sự phát triển của lĩnh vực QA với 20 thành viên tham gia. Sự kiện này đã thúc đẩy các nhà nghiên cứu chia sẻ kinh nghiệm và so sánh kết quả của họ sử dụng một độ đo chung.

Năm 2000, Carbonell và các đồng nghiệp trong bài báo The Vision Statement to Guide Research in Question Answering and Text Summarization đã đưa ra các tư tưởng chung cho việc nghiên cứu Q&A. Theo đó một hệ thống hỏi đáp được người dùng đánh giá là hữu ích nếu đáp ứng được các tiêu chuẩn:

* Tính hợp lý về thời gian (Timeliness): Câu trả lời phải được đưa ra trong thời gian ngắn, ngay cả khi có hàng ngàn người dùng cùng truy nhập hệ thống một lúc. Các nguồn dữ liệu mới cần phải được tích hợp vào hệ thống ngay khi chúng sẵn sàng để có thế cung cấp cho người dùng câu trả lời cho những câu hỏi về các sự kiện có tính thời sự.
* Tính chính xác: Tính chính xác của hệ thống hỏi đáp tự động là cực kì quan trọng bởi việc đưa ra câu trả lời sai còn tai hại hơn nhiều là không đưa ra câu trả lời. Nghiên cứu về Q&A cần tập trung vào việc đánh giá tính đúng đắn của câu trả lời đưa ra, bao gồm cả phương thức để phát hiện các trường hợp mà dữ liệu hiện thời không chứa câu trả lời cho câu hỏi. Các thông tin mâu thuẫn trong dữ liệu cũng cần được tìm ra và các thông tin này cần được xử lý theo một cách phù hợp, nhất quán. Để đạt được sự chính xác, hệ thống Q&A cần được tích hợp các nguồn tri thức (world knowledge ) và cơ chế “bắt chước” việc suy luận thông thường (việc bắt chước có thể hiểu như là một quá trình học).
* Tính khả dụng: Hệ thống Q&A cần đáp ứng được các yêu cầu cụ thể của một người dùng. Các ontology trên từng miền cụ thể và ontology trên miền mở cần được tích hợp trong hệ thống. Hệ thống Q&A cần có khả năng khai phá câu trả lời 5 từ bất kì dạng dữ liệu gì (văn bản, web, cơ sở dữ liệu, …) và đưa ra câu trả lời dưới định dạng mà người dùng mong muốn, cho phép người dùng miêu tả ngữ cảnh của câu hỏi và cung cấp các thông tin giải thích, trích dẫn nguồn cho câu trả lời.
* Tính hoàn chỉnh: Câu trả lời hoàn chỉnh cho câu hỏi của người dùng là điều mà các hệ thống Q&A hướng tới. Trong nhiều trường hợp (câu hỏi về danh sách, nguyên nhân, cách thức…), các phần của câu trả lời nằm rải rác trong một văn bản, thậm chí trong nhiều văn bản. Vì vậy cần phải hợp nhất các phần này dựa trên các thông tin liên kết để tạo ra câu trả lời hoàn chỉnh.
* Tính thích hợp của câu trả lời: Trong ngôn ngữ tự nhiên, câu hỏi đưa ra luôn gắn với ngữ cảnh nào đó và câu trả lời cũng nằm trong một ngữ cảnh nhất định. Câu trả lời mà hệ thống Q&A đưa ra phải phù hợp ngữ cảnh với câu hỏi. Một hệ thống Q&A có khả năng giao tiếp (interactive Q&A) là cần thiết trong nhiều trường hợp bởi chuỗi các câu hỏi liên quan đến một vấn đề sẽ giúp làm sáng tỏ thông tin mà người dùng đang hỏi. Việc đánh giá một hệ thống Q&A cần hướng người dùng bởi ý kiến người dùng là đánh giá tốt nhất cho tính thích hợp của câu trả lời. Các tiêu chuẩn trên được đặt ra với mong muốn xây dựng được một hệ thống Q&A hoàn chỉnh. Tuy nhiên, không phải hệ thống nào cũng có khả năng thông minh và hoàn thiện như thế. Các nghiên cứu về Q&A hiện nay đang tập trung vào xây dựng hệ thống hỏi đáp có tính chính xác cao và có khả năng sử dụng nguồn dữ liệu web khổng lồ trên Internet.

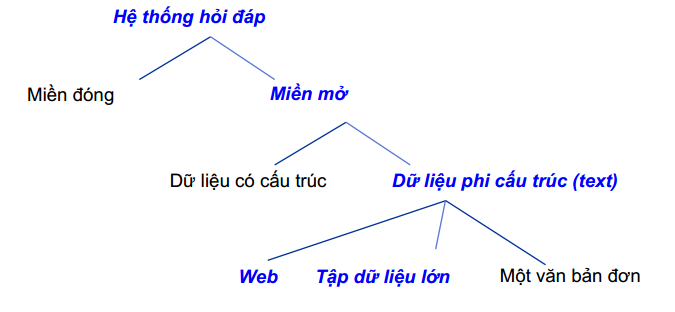
Rất nhiều phương pháp tiếp cận khác nhau được sử dụng trong các hệ thống hỏi-đáp gần đây như phương pháp thống kê, phương pháp dựa trên luật, và các phương pháp tổng hợp. Hầu hết các hệ thống hỏi-đáp ngày càng trở nên phức tạp, thường sử dụng rất nhiều các mô-đun khác nhau, chẳng hạn như tìm kiếm thông tin, phân tích cú pháp câu, phân tích loại câu hỏi, phân tích ngữ nghĩa, và thậm chí dùng phương pháp suy luận (Moldovan et al., 2002) để đánh giá, và xếp hạng câu trả lời. Với sự gia tăng tính phức tạp của hệ thống hỏi-đáp, sẽ rất khó khăn để xác định hiệu quả của hệ thống đến từ hiệu quả của giai đoạn nào.

### Phân loại hệ thống hỏi đáp

Có nhiều cách phân loại hệ thống hỏi đáp dựa trên các tiêu chí khác nhau như: Phân loại theo miền ứng dụng, theo khả năng trả lời câu hỏi, theo cách tiếp cận giải quyết bài toán…

#### Phân loại theo miền ứng dụng

Hệ thống hỏi đáp miền mở (open domain Question answering): Hệ thống trả lời bất kỳ câu hỏi nào được đưa vào. Khó khăn cho hệ thống miền mở đó chính là việc xây dựng các tri thức nên cho việc trả lời cũng như phân tích câu hỏi, các phương pháp hiện nay thường sử dụng một số các ontology khái quát hay các mạng tri thức như: wikipedia, bách khoa từ điển... Tuy nhiên, dữ liệu cho việc trích rút câu trả lời là phong phú, dễ thu thập.

Hệ thống hỏi đáp miền đóng (close domain Question answering): Hệ thống tập trung vào trả lời các câu hỏi liên quan đến một miền cụ thể (giáo dục, y tế, thể thao...). Xây dựng hệ thống hỏi đáp miền đóng được coi là bài toán dễ hơn so với xây dựng hệ thống hỏi đáp miền mở vì có thể sử dụng các tri thức miền (thường là ontology của miền cụ thể).

Các nghiên cứu hiện nay về Q&A đang tập trung vào xây dựng hệ thống hỏi đáp trên miền mở, sử dụng nguồn dữ liệu phi cấu trúc (kho văn bản lớn hay dữ liệu web) để tìm câu trả lời. Các nghiên cứu mới và cải tiến những phương pháp cũ để có thể áp dụng cho nguồn dữ liệu web vốn đa dạng, nhiều “nhiễu” và trùng lặp đang rất được quan tâm.

#### Phân loại theo khả năng trả lời câu hỏi

Hệ thống có khả năng trả lời các câu hỏi liên quan đến sự vật, hiện tượng,... dựa trên việc trích ra câu trả lời có sẵn trong tập tài liệu. Câu trả lời là các chuỗi ký tự trong một tài liệu. Kỹ thuật chính được sử dụng là xử lý chuỗi và từ khóa.

Hệ thống có cơ chế lập luận đơn giản: Trích xuất các câu trả lời có sẵn trong tập tài liệu sau đó sử dụng các suy luận để tìm mối liên kết giữa câu trả lời và câu hỏi. Hệ thống sử dụng các nguồn tri thức như ontology về từng miền cụ thể và ontology chung.

Hệ thống trả lời các câu hỏi yêu cầu khả năng tổng hợp: Các phần của câu trả lời được trích rút từ nhiều tài liệu sau đó được tổng hợp lại thành câu trả lời hoàn chỉnh. Câu hỏi thường là về danh sách, về cách thức, nguyên nhân...

Hệ thống có khả năng giao tiếp với người dùng: Trả lời chuỗi các câu hỏi của người dùng về cùng một vấn đề. Ví dụ các câu hỏi của người dùng như: “Giáo sư A sinh năm nào? Ở đâu? Ông ấy đang công tác ở đâu?”.

Hệ thống có khả năng lập luận tương tự: Có thể trả lời các câu hỏi có tính chất suy đoán, câu trả lời ẩn trong tập tài liệu. Hệ thống cần trích ra các luận chứng và sử dụng lập luận tương tự để tìm ra câu trả lời.

#### Phân loại theo hướng tiếp cận

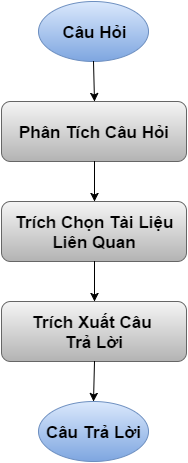
Hướng tiếp cận nông (shalow): Nhiều phương pháp sử dụng trong Q&A dùng các kĩ thuật dựa trên từ khóa để định vị các câu, đọan văn có khả năng chứa câu trả lời từ các văn bản được trích chọn về. Sau đó giữ lại các câu, đoạn văn có chứa chuỗi ký tự cùng loại với loại câu trả lời mong muốn (ví dụ các câu hỏi về tên người, địa danh, số lượng…).

Hướng tiếp cận sâu (deep): Trong những trường hợp khi mà hướng tiếp cận bề mặt không thể tìm ra câu trả lời, những quá trình xử lý về ngữ pháp, ngữ nghĩa và ngữ cảnh là cần thiết để trích xuất hoặc tạo ra câu trả lời. Các kĩ thuật thường dùng như nhận dạng thực thể (named-entity recognition), trích xuất mối quan hệ, loại bỏ nhập nhằng ngữ nghĩa,… Hệ thống thường sử dụng các nguồn tri thức như Wordnet, ontology để làm giàu thêm khả năng lập luận thông qua các định nghĩa và mối liên hệ ngữ nghĩa. Các hệ thống hỏi đáp dựa theo mô hình ngôn ngữ thống kê cũng đang ngày càng phổ biến.

### Các bước chung của hệ thống hỏi đáp

Một hệ thống hỏi đáp tự động thường gồm 3 bước chung sau:

* Bước 1: Phân tích câu hỏi: Bước phân tích câu hỏi tạo truy vấn cho bước trích chọn tài liệu liên quan và tìm ra những thông tin hữu ích cho bước trích xuất câu trả lời.
* Bước 2: Trích chọn tài liệu liên quan: Bước này sử dụng câu truy vấn được tạo ra ở bước phân tích câu hỏi để tìm các tài liệu liên quan đến câu hỏi.
* Bước3: Trích xuất câu trả lời: Bước này phân tích tập tài liệu trả về từ bước 2 và sử dụng các thông tin hữu ích do bước phân tích câu hỏi cung cấp để đưa ra câu trả lời chính xác nhất.



Các hệ thống hỏi đáp tự động hiện nay có kiến trúc rất đa dạng, tuy nhiên chúng đều bao gồm ba phần cơ bản như trên. Sự khác nhau chính giữa các hệ thống là ở quá trình xử lý trong từng bước, đặc biệt là ở cách tiếp cận trong việc xác định câu trả lời.

Cách tiếp cận theo trích chọn thông tin thuần túy (pure IR) là: chia nhỏ một tài liệu trong tập dữ liệu thành chuỗi các tài liệu con, trích chọn các tài liệu con có độ tương đồng lớn nhất với câu truy vấn (do bước phân tích câu hỏi tạo ra) và trả lại chúng cho người dùng.Thách thức lớn nhất ở đây là làm sao chia nhỏ được tài liệu thành các phần với kích cỡ tương ứng với kích cỡ của câu trả lời mà vẫn đủ lớn để có thể đánh chỉ mục được (nếu chia quá nhỏ thì số lượng tài liệu để đánh chỉ mục sẽ rất lớn, gây gánh nặng cho hệ thống trích chọn thông tin).

Cách tiếp cận theo xử lý ngôn ngữ tự nhiên (pure NLP) là: so khớp giữa biểu diễn ngữ pháp và (hoặc) biểu diễn ngữ nghĩa của câu hỏi với dạng biểu diễn ngữ pháp, ngữ nghĩa của các câu trong các tài liệu liên quan trả về. Khó khăn của cách tiếp cận này là hệ thống phải thực hiện việc phân tích ngữ pháp, ngữ nghĩa và so khớp đủ nhanh để đưa ra câu trả lời trong thời gian chấp nhận được, bởi số lượng các tài liệu cần xử lý là rất lớn trong khi các bước phân tích trên lại phức tạp và tốn nhiều thời gian.

Sự khác nhau trong cách trích xuất câu trả lời dẫn đến việc phân tích câu hỏi cũng trở nên đa dạng. Trong hướng tiếp cận theo trích xuất thông tin thuần túy, phân tích câu hỏi chỉ cần làm tốt việc tạo truy vấn, trong khi với hướng tiếp cận theo xử lý ngôn ngữ tự nhiên, câu hỏi cần được phân tích ngữ pháp, ngữ nghĩa một cách chính xác. Các hệ thống hiện nay thường là sự kết hợp giữa hai hướng tiếp cận, sử dụng hệ thống trích chọn thông tin để thu hẹp không gian tìm kiếm câu trả lời, đồng thời phân tích câu hỏi để tìm ra các thông tin về ngữ pháp, ngữ nghĩa nhằm tìm ra câu trả lời chính xác nhất. Kết quả của bước phân tích câu hỏi là đầu vào cho cả hai bước trích chọn tài liệu liên quan và trích xuất câu trả lời. Bước phân tích câu hỏi có ý nghĩa rất quan trọng, bởi nó ảnh hưởng đến hoạt động của các bước sau và do đó quyết định đến hiệu quả của toàn hệ thống. Chương 2 trình bày chi tiết các nội dung liên quan đến phân tích câu hỏi.

### Phương pháp đánh giá hệ thống hỏi đáp

Có hai phương pháp cơ bản dùng để đánh giá hệ thống hỏi đáp.

#### Trung bình thứ hạng đối xứng (MRR)

Chuẩn đánh giá này được đề xuất tại TREC-8. Chuẩn này được áp dụng cho các câu hỏi Factoid và non-Factoid như sau:

+ Cho một tập câu hỏi mẫu Q có NQ câu hỏi.

+ Tập tài liệu dùng để rút trích câu trả lời D.

+ Tập câu trả lời mẫu A={[ai, di]} là tập các cặp gồm câu trả lời ai và chỉ số di  của tài liệu trong D được dùng để rút trích câu trả lời đó.

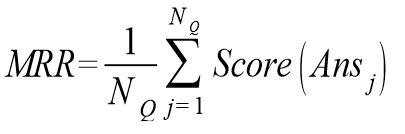
+ Hệ thống sẽ có 5 câu trả lời trong tập Ansq cho một câu hỏi q. Câu trả lời có thể là 50byte hoặc 250bytes kèm theo tài liệu dùng để trích câu trả lời.

Điểm số của kết quả trả lời Ansq là:

Score (Ansq) =

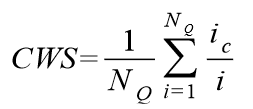
Với i<=5 là chỉ số của câu trả lời đún đầu tiên trong Ansq.

Nếu I > 5 thì Score(Ansq) = 0.

Trung bình thứ hạng đối xứng của hệ thống được tính là:

#### Điểm trọng số tin cậy (CWS)

Điểm trọng số tin cậy Confidence-Weighted Score (CWS) nhằm đánh giá khả năng hệ thống hiểu câu trả lời của nó như thế nào. Với số liệu được thiết lập như đánh giá trung bình thứ hạng đối xứng với rang buộc hệ thống chỉ được trả lời 1 câu trả lời. Điểm trọng số tin cậy được tính như sau:



## Giới thiệu về tìm kiếm thông tin

### Khái niệm search engine và cách sử dụng

#### Khái niệm search engine

Thuật ngữ Search Engine được dùng để nói đến 2 hệ thống tìm kiếm.

Do các chương trình máy tạo ra. (Crawaler-Based Search Engines và dạng thư mục do con người quản lý Human-Powered)

Hệ thống này sẽ tìm kiếm và lập chỉ mục website theo 2 cách khác nhau:

+ Crawler-Based Search Engnes: Các máy tìm kiếm thuộc loại này được sử dụng các chương trình được gọi là Robots hay Spiders (mình hiểu chung nó là những con bot hay con bọ. ). Các bot sẽ lần và tìm các website trên mạng và tự động phân tích và đưa website vào cơ sở dữ liệu của nó. Khi có một yêu cầu tìm kiếm thì các Search Engine đối chiếu với các từ khóa trong dữ liệu mà nó đã lập chỉ mục và trả lại các kết quả phù hợp tại trang kết quả. Các cỗ máy tìm kiếm có chế cập nhật nội dung của website định kỳ để có thể phát hiện sự thay đổi của trang web nếu có.

+ Human-Powered Directories: Các chỉ mục của website (ở đây mình gọi nó là sitemap) hoàn toàn phụ thuộc và sự quản lý của con người. Nếu chúng ta muốn các Search Engine tìm thấy website của mình thì chúng ta phải gửi bản đăng ký website đến các Search Engine. (VD: Công cụ Submit URL của Google, Thêm sitemap của website vào Search Console của Google) .

Các bạn có thể hiểu đơn giản là trong một website có những page nào đó ta muốn các con Bot không index nó thì, còn một số page khác ta lại muốn nó index.

#### Cách sử dụng

* Bộ thu thập thông tin (Crawler) :

Trước hết chúng ta cần hiểu  rằng các website trên thế giới được liên kết với nhau bằng hệ thống liết kết như mạng nhện vậy. Các SE ( Search Engine ) thả những con bọ tìm kiếm. Dựa trên những thuật toán có sẵn các con bọ tìm kiếm sẽ theo thuật toán và tìm đến các website.

* Bộ lập chỉ mục (Index)

Từ những website mà các con bọ tìm đến thì các con bot của SE sẽ tiếp tục đi theo các liên kết trong website hoặc ngoài website. ( Trong quá trình Crawler các con bot sẽ nhân lên thành nhiều những con bot khác và đi đến các đường dẫn và index các site khác ). Lập chỉ mục là dai giai  đoạn phân tích tài liệu (document) để xác định các chỉ mục biểu diễn các nội dung của bài viết.

* Kho dữ liệu Repository

Là một hệ thống lưu trữ có khác năng mở rộng, nó quản lý một tập lớn các trang web. Kho dữ liệu Repositoy thực hiện 2 chức năng chính. Một là cho phép Crawler lữu trữ các trang web. Hai là nó phải cung cấp API truy cập hiệu quả để bộ Indexer và Collection Analysis có thể sử dụng để lấy các trang từ kho dữ liệu.

Nếu bạn chưa hiểu Indexer và Collection Analysis là gì thì dưới đây sẽ là một chút thông tin về nó. Module Indexer và Collection Analysis có chức năng tạo ra nhiều loại chỉ mục khác nhau.

* Module Indexer tạo ra hai loại chỉ mục chính đó là:

Text Index: Chỉ mục nội dung.

Structure Index: Chỉ mục liên kết.

Dựa vào hai loại chỉ mục này Collection Analysis sẽ tạo ra nhiều loại chỉ mục hữu ích khác như:

Link index: Tạo chỉ mục liên kết các đoạn web.

Text Index: Phương pháp đánh chỉ mục dựa theo nội dung (text-based) là phương pháp quan trọng để định danh các trang web có liên quan đến yêu cầu tìm kiếm.

* Bộ tìm kiếm thông tin

Search Engine Search Engine là cụm từ dùng để mô tả toàn bộ hệ thống bao gồm thu thâp thông tin, bộ lập chỉ mục, bộ tìm kiếm thông tin. Search Engine sẽ tương tác với người dùng qua giao diện của website, nó sẽ có nhiệm vụ tiếp nhận và trả về những kết quả từ người dùng sau khi đối sánh và tham chiếu.  
- Bộ Query Engine

Bộ công cụ Query Engine có nhiệm vụ nhận và tìm kiếm các yêu cầu của người sử dụng. Bộ công cụ này sẽ dựa vào các chỉ mục tại kho lưu trữ dữ liệu Repository để trả về các truy vấn của người dùng trên trang kết quả. Nhưng để tra về những kết quả chính xác nhất với người dùng thì sẽ cần đến module xắp xếp.  
- Module:

sắp xếp Là một module có chức năng lọc các thông tin truy vấn từ người dùng từ rất nhiều các kết quả trong kho dữ liệu và trả về trang kết quả cũng như xắp xếp cho phù hợp nhất.

### Bing và cách sử dụng Bing Search API

#### Giới thiệu về Bing

Bing (trước đây là Live Search, Windows Live Search và MSN Search) là [bộ máy tìm kiếm web](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) (được quảng cáo là một bộ máy "ra quyết định"), đại diện cho công nghệ tìm kiếm hiện nay của Microsoft. Được Giám đốc Điều hành của Microsoft [Steve Ballmer](https://vi.wikipedia.org/wiki/Steve_Ballmer) tiết lộ vào ngày [28 tháng 5](https://vi.wikipedia.org/wiki/28_th%C3%A1ng_5) năm [2009](https://vi.wikipedia.org/wiki/2009) tại hội nghị All Things D tại San Diego, Bing là một sự thay thế cho Live Search; bộ máy tìm kiếm này được đưa lên trực tuyến hoàn toàn vào ngày [3 tháng 6](https://vi.wikipedia.org/wiki/3_th%C3%A1ng_6) năm [2009](https://vi.wikipedia.org/wiki/2009).

**Tính năng giao diện**

* Hình nền về các nơi trên thế giới thay đổi hàng ngày có các thông tin mà bạn có thể xem bằng cách rê chuột lên hình ảnh.
* Phân đề mục nội dung của kết quả (tách các phần riêng cho hình nền, bản đồ, thời tiết, trang hâm mộ, v.v.)
* Khung duyệt trang bên trái. Bao gồm điều hướng và, trên các trang kết quả, sẽ liên quan đến các tìm kiếm và tìm kiếm trước
* Xem thử mở rộng ở bên phải với danh sách các url tương ứng hoặc quan trọng trong một bài
* Liên kết con. Trên một số kết quả nhất định, trang kết quả tìm kiếm cũng hiển thị các liên kết đề mục bên trong một trang (Wikipedia)
* Mở rộng xem đối với thông tin từ bên thứ ba có thể xem được từ Bing. Cách này hoạt động được với trang Wikipedia

**Tính năng phương tiện**

* Xem thử thu nhỏ video, khi rê chuột lên biểu tượng thu nhỏ video, đoạn video sẽ tự động chơi
* Tìm kiếm hình ảnh từ trang kết quả hình ảnh liên tục cuộn có các thiết lập thay đổi được như kích thước, trình bày, màu sắc, kiểu và người[[8]](https://vi.wikipedia.org/wiki/Bing#cite_note-8).
* Tìm kiếm video với thiết lập thay đổi được độ dài, kích thước màn hình, độ phân giải và nguồn.

**Thông tin tìm kiếm cải tiến**

* Tỷ số thể thao và thống kê về Đội bóng và Cầu thủ
* Liệt kê khác sạn trong thành phố
* Liệt kê các hãng kinh doanh
* Liệt kê về người
* Bộ sưu tập
* Trích dẫn tài chính
* Thông tin xe cộ
* Thông tin giao thông hiện tại
* Tìm kiếm địa phương hóa cho nhà hàng và dịch vụ
* Các bình luận về nhà hàng
* Xếp hạng người nổi tiếng (xRank)
* Tin tức về người nổi tiếng
* Các bộ phim đang chiếu trong khu vực
* Phép tính (2 \* pi \* 24)
* Câu trả lời thức thời (What is the capitol of Germany ?)
* So trùng đúng nhất (cùng với các trang tương tự)
* Thông tin giá vé máy bay và tình trạng chuyến bay
* Mua hàng và Bing Cashback
* Thông tin sức khỏe
* Dò tình trạng gói hàng

#### Bing search API

Bing Search API là công cụ hỗ trợ các lập trình viên có thể lấy được các kết quả từ Bing trong các ứng dụng hoặc website, kết quả này dưới dạng JSON hoặc XML.

Hiện nay Bing Search API đã được chuyển qua Window Azure Marketplace.

Window Azure Marketplace cung cấp những chức năng:

- Cho phép truy cập các kết quả trang web, hình ảnh, tin tức và video cũng như những câu gợi ý tìm kiếm hay kiểm tra chính tả.

- Kết quả cải thiện và chính xác hơn so với API 2.0

- Hỗ trợ kiểu dữ liệu JSON, XML và Odata

- Có cơ hội kiếm tiền từ các ứng dụng của Window Azure Marketplace

- Có thể truy cập hàng trăm dữ liệu khác để phục vụ cho ứng dụng.

**Cách sử dụng:**

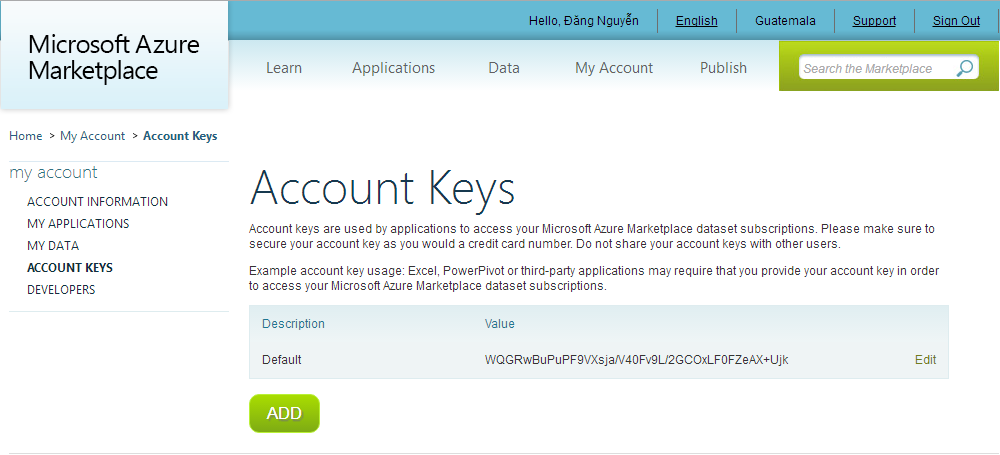
- Lấy Account key:

+ Truy cập vào <http://www.bing.com/developers/>

+ Click vào Get Started trong mục Search API

+ Bing đưa ra 1 vài lựa chọn tùy theo nhu cầu, ở đây chúng ta sẽ chọn 5000 query trong 1 tháng. Sau đó click vào Sign Up

+ Tạo 1 Window Azure Marketplace Account

+ Đăng kí Bing Search API

- Sau đây là Java Code:

+ Khai báo Account Key



+ Khai báo UrlPatern:



Ở đây query chính là câu mà người dùng nhập vào để tìm kiếm

$format= chính là kiểu dữ liệu được trả về, ở đây là JSON

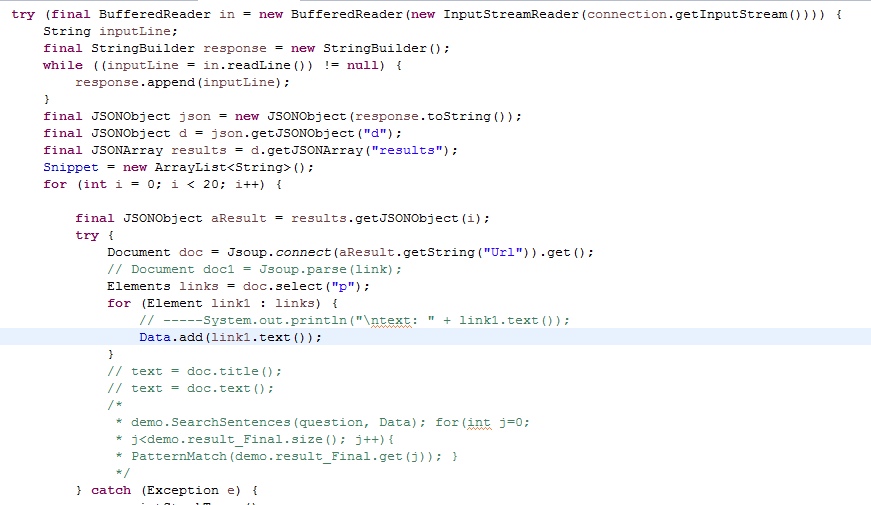
+ Encode Account Key:



+ Kết nối với URL:



+ Chúng ta sẽ nhận được kết quả trả ra là một JSON Object, trong đó có chưa chuỗi các *url, snippet, title* là kết quả trả ra từ Bing



+Kết quả với câu truy vấn “Ai là vua phá lưới wc 2014”:

Bảng 2.1: Tên bảng 1

# PHƯƠNG PHÁP TRẢ LỜI CÂU HỎI

## Xác định dạng câu hỏi

Việc hệ thống tìm đúng được tài liệu liên quan đến câu hỏi, hay xác định được nội dung câu hỏi sẽ hỏi về vấn đề gì và sau đó là xác định regex để tìm câu trả lời phụ thuộc khá nhiều vào quá trình xác định dạng câu hỏi. Do đó quá trình xác định dạng câu hỏi quyết định một phần không nhỏ đến hiệu năng của hệ thống.

Quá trình xác định dạng câu hỏi nhận input là câu hỏi dưới dạng ngôn ngữ tự nhiên do người dùng nhập vào. Dựa vào tập câu hỏi khảo sát sẽ xây dựng biểu thức chính quy cho mỗi dạng câu hỏi.

Để xác định được dạng câu hỏi, ta sử dụng các keyword trong câu hỏi. Việc xác định keyword trong câu hỏi được thực hiện thông qua việc thu thập bộ câu hỏi liên quan, từ đó đưa ra tập các keyword chung để xác định dạng câu hỏi phù hợp.

Ví dụ: Muốn xác định đâu là dạng câu hỏi về tên người ta xác định như sau:

* Tập các keyword thu thập có liên quan đến dạng câu hỏi về tên người: Ai, người nào, cầu thủ nào, huấn luyện viên nào, là ai, là người nào, là cầu thủ nào, là huấn luyện viên nào,…
* Xác định regex tương ứng cho câu hỏi: **“(Ai|ai|Người nào|người nào|Cầu thủ|cầu thủ)(.\*)") || question.matches("(.\*) ai (.\*)”**
* Câu hỏi nhập vào dưới dạng String, do đó sẽ dùng hàm **“matches”** của String để xác định: “**question.matches("(Ai|ai|Người nào|người nào|Cầu thủ|cầu thủ)(.\*)") || question.matches("(.\*) ai (.\*)")**”
* Kết quả trả về sẽ là một số, con số này sẽ đại diện cho loại câu hỏi, mỗi loại câu hỏi sẽ có mốt số riêng biệt. Con số này sẽ được dùng để xác định biểu thức chính quy thích hợp cho loại câu hỏi đó.

### Chủ đề cấp độ 3

#### Chủ đề cấp độ 4

## Phân tích câu hỏi

Khi một câu hỏi được người dùng nhập vào, câu hỏi sẽ được phân tích trước khi đưa vào hàm tìm kiếm.

Sẽ có hai dạng để phân tích câu hỏi. Một dạng dùng để tìm kiếm tài liệu từ Bing và Google và một dạng dùng để lucene tìm kiếm đoạn tài liệu thích hợp.

### Dạng tìm kiếm trên Bing và Google

Để thực hiện tìm kiếm tài liệu thông qua Bing hay Google, trước hết ta phải đưa dạng câu hỏi về định dạng tìm kiếm. Định dạng tìm của của Bing và Google cũng giống với câu hỏi dưới dạng ngôn ngữ tự nhiên, nhưng thay vào đó khoảng cách giữa các từ là khoảng trắng ta sẽ thay thế bằng dấu “+”

Ví dụ: khi nhập vào câu hỏi “Ai là vua phá lưới world cup 2014”. Qua quá trình phân tích câu hỏi trên sẽ trở thành:

“Ai+là+vua+phá+lưới+world+cup+2014”

Quá trình phân tích này được thực hiện bằng cách dùng hàm “replace”.

“**string.replace(' ', '+');**”

### Dạng áp dụng cho lucene

Quá trình phân tích câu hỏi gồm các bước: phân đoạn từ (Tokenizer), gán nhãn từ loại (Vntagger), loại bỏ từ không cần thiết trong câu hỏi (StopWord).

Trong quá trình phân tích một câu hỏi, nhóm có sử dụng hai công cụ được phát triển bởi tác giả Lê Hồng Phương là tokenizer và vntagger. Sau khi đã phân đoạn và gán nhãn từ loại, sử dụng biểu thức chính quy để loại bỏ stopword.

### Phân đoạn từ

#### Giới thiệu về công cụ Tokenizer

#### Quá trình phân đoạn

TÀI LIỆU THAM KHẢO

Theo chuẩn IEEE