**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

****

**Báo cáo tiểu luận chuyên ngành**

Đề tài: **Tìm hiểu các hướng tiếp cận dùng NLP (xử lý ngôn ngữ tự nhiên) và bài toán Sentiment Analysis (phân tích cảm xúc)**

GVHD: Trần Lê Minh Sang

SVTH: Nguyễn Đình Quyền 15110115

**Đặng Công Đạt 15110030**

***TP.HCM, ngày 15 tháng 9 năm 2018***

**I. Giới thiệu về xử lý ngôn ngữ tự nhiên  
1. Tổng quan**

Xử lý ngôn ngữ chính là xử lý thông tin khi đầu vào là “dữ liệu ngôn ngữ”, tức dữ liệu “văn bản” hay “tiếng nói”. Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) và tiếng nói đang dần trở nên kiểu dữ liệu chính con người có và lưu trữ dưới dạng điện tử.

Một bài viết hoặc một văn bản có thể có rất nhiều câu , và trên thực tế có đến hàng triệu văn bản hoặc hơn thế nữa . Trong tương lai , hầu như các kiến thức nhân loại sẽ được đưa vào máy tính bằng các chương trình nhận dạng chữ, thu nhập âm thanh, hoặc  
gõ thẳng vào máy tính . Vấn đề là làm sao chuyển đổi được khối dữ liệu văn bản khổng lồ này qua dạng khác để mỗi người có được thông tin và tri thức cần thiết từ chúng.

Xử lý ngôn ngữ tự nhiên đã được ứng dụng trong thực tế để giải quyết các  
bài toán như : nhận dạng chữ viết, nhận dạng tiếng nói, tổng hợp tiếng nói, dịch tự  
động, tìm kiếm thông tin, tóm tắt văn bản, khai phá dữ liệu và phát hiện tri thức.

**2. Cơ sở khoa học**

**2.1 Một số khái niệm cơ bản**

**2.1.1 Xử lý ngôn ngữ tự nhiên**

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh  
của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong  
trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó  
liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy  
và giao tiếp.[1]

**2.1.2 Trí tuệ nhân tạo**

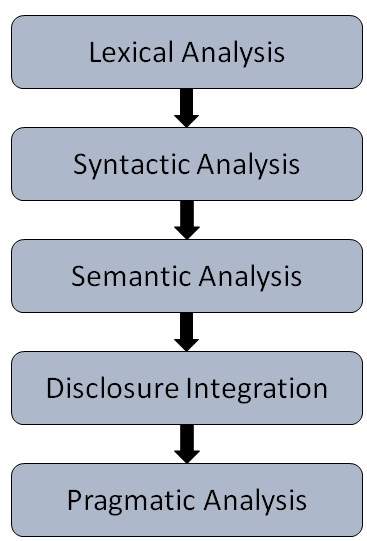
Trí tuệ nhân tạo hay trí thông minh nhân tạo (tiếng Anh: artificial intelligence hay machine intelligence, thường được viết tắt là AI) là trí tuệ được biểu diễn bởi bất cứ một hệ thống nhân tạo nào. Thuật ngữ này thường dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và ứng dụng của trí tuệ nhân tạo.

**2.2 Lợi ích của xử lý ngôn ngữ tự nhiên**

Có hàng triệu gigabyte mỗi ngày được tạo bởi blog, trang web xã hội và trang web .Có nhiều công ty thu thập tất cả các dữ liệu này để hiểu người dùng và niềm đam mê của họ và đưa những báo cáo này cho các công ty để điều chỉnh kế hoạch của họ.

Những dữ liệu này có thể cho thấy rằng người dân Brazil hài lòng với sản phẩm A có thể là một bộ phim hay bất cứ thứ gì ,trong khi người dân Hoa Kỳ hài lòng về sản phẩm B. Giống như những gì công cụ tìm kiếm làm, họ đưa ra kết quả phù hợp cho đúng người vào đúng thời điểm.

**2.3 Quy trình xử lý ngôn ngữ tự nhiên**

****

**Lexical Analysis** −  Nó liên quan đến việc xác định và phân tích cấu trúc của các từ. Lexicon của một ngôn ngữ có nghĩa là tập hợp các từ và cụm từ trong một ngôn ngữ. Phân tích Lexical phân chia toàn bộ đoạn txt thành các đoạn văn, câu và từ.

Ví dụ : Quá trình phân tích từ vựng cho câu lệnh gán position := initial + rate \*  
60 sẽ tách thành các token như sau:  
1. Danh biểu position  
2. Ký hiệu phép gán :=  
3. Danh biểu initial  
4. Ký hiệu phép cộng (+)  
5. Danh biểu rate  
6. Ký hiệu phép nhân (\*)  
7. Số 60  
Trong quá trình phân tích từ vựng các khoảng trắng (blank) sẽ bị bỏ qua.

**Syntactic Analysis (Parsing)** –Nó liên quan đến việc phân tích các từ trong câu cho ngữ pháp và sắp xếp các từ theo cách thể hiện mối quan hệ giữa các từ. Câu như “Trường học đi học” bị từ chối bởi bộ phân tích cú pháp tiếng Anh.

**Semantic Analysis** – Nó rút ra ý nghĩa chính xác hoặc ý nghĩa từ điển từ văn bản. Văn bản được kiểm tra ý nghĩa. Nó được thực hiện bằng cách ánh xạ các cấu trúc cú pháp và các đối tượng trong miền nhiệm vụ. Các phân tích ngữ nghĩa không quan tâm đến câu .

**Discourse Integration** – Ý nghĩa của bất kỳ câu nào phụ thuộc vào ý nghĩa của câu ngay trước nó. Ngoài ra, nó cũng mang lại ý nghĩa của câu ngay lập tức thành công. **Pragmatic Analysis** − Trong thời gian này, những gì được cho là được diễn giải lại theo ý nghĩa của nó. Nó liên quan đến việc phát sinh những khía cạnh của ngôn ngữ đòi hỏi kiến ​​thức thế giới thực.

**TÀI LIỆU THAM KHẢO**

[1]<https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn>

<https://likegeeks.com/nlp-tutorial-using-python-nltk/>

<https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm>