**Vietnamese Natural Language Processing**

**------------------------------**

**Mục lục**

1. **Giới thiệu**………………………………………………………….2
2. **Tiền xử lý**………………………………………………………….3
3. **Thuật toán tách từ**………………………………………………..3
4. **Đặc trưng của Tiếng Việt**…………………………………..3
5. **Tách từ là gì**?..........................................................................3
6. **Thuật toán tách từ**………………………………………….4
7. **Xử lý bằng Python**………………………………………….5
8. **Machine Learning trong NLP**…………………………………..7
9. **Ứng dụng**……………………………………………………7
10. **Định lý Bayes**……………………………………………….7
11. **Bài toán phân loại văn bản**………………………………...7
12. **Giới thiệu**

Xử lý ngôn ngữ tự nhiên là một nhánh của trí tuệ nhân tạo, là một kĩ thuật quan trọng nhằm giúp máy tính hiểu được ngôn ngữ của con người, qua đó hướng dẫn máy tính thực hiện và giúp đỡ con người trong những công việc có liên quan đến ngôn ngữ như: dịch thuật, phân tích dữ liệu văn bản, nhận dạng tiếng nói, tìm kiếm thông tin …

NLP (Natural Language Processing) là khái niệm để chỉ các kỹ thuật, phương pháp thao tác trên ngôn ngữ tự nhiên bằng máy tính. Trong NLP có 2 quan điểm cơ bản:

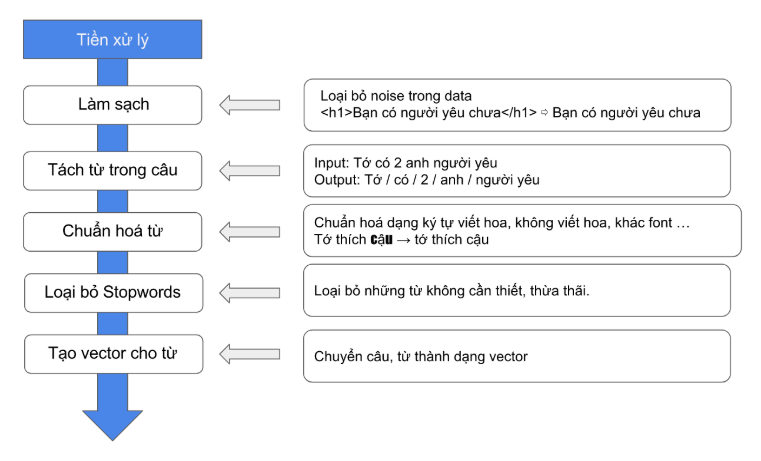
1. Xử lý các từ ngữ bằng máy tính.
2. Làm cho máy tính hiểu được các từ ngữ.

Ứng dụng của NLP:

1. Máy dịch
2. Chiết suất thông tin
3. Tóm tắt văn bản
4. Phân loại văn bản
5. Data mining, web mining

NLTK là một bộ công cụ dành riêng, hữu hiệu và mạnh mẽ nhất cho NLP và được tích hợp vào Python.

1. **Tiền xử lý**

****

1. **Thuật toán tách từ (Tokenizer)**
2. **Đặc trưng của tiếng Việt**

Khác với tiếng Anh, văn bản tiếng Việt đặt dấu cách giữa các âm tiết chứ không phải giữa các từ. Một từ có thể có một, hai hoặc nhiều âm tiết.

Tiếng Việt không sử dụng các **hình thái** để tạo ra các ý nghĩa của từ. Để tạo ra các sắc thái ý nghĩa khác nhau, tiếng Việt phụ thuộc vào **trật tự của từ**.

Vì vậy, việc phân tách từ trong tiếng Việt là một việc không thể thiếu.

1. **Tách từ là gì?**

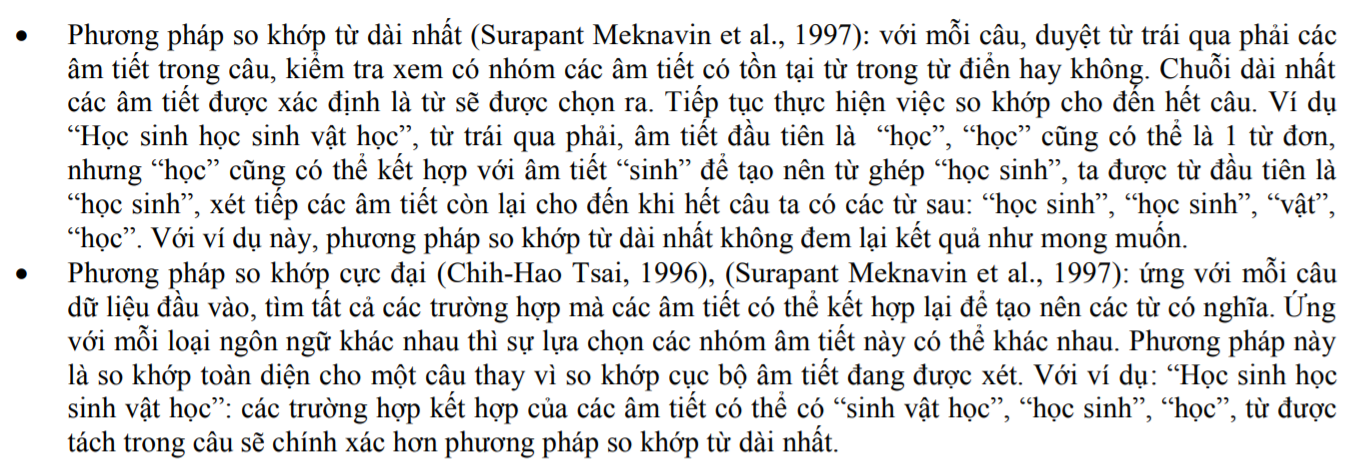
Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu.

1. **Thuật toán tách từ**

* Hướng tiếp cận: dựa vào từ điển cố định (khi từ đơn không xuất hiện trong từ điển, coi như đây là một từ mới).

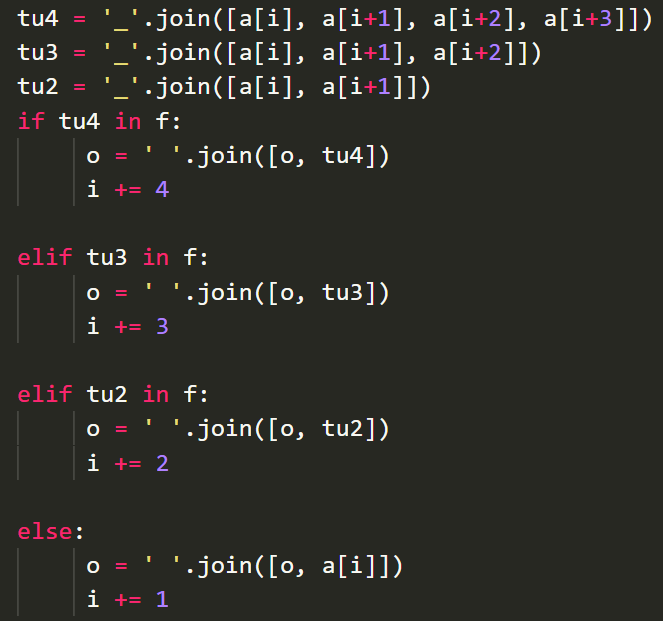
Sử dụng thuật toán “**so** **khớp từ dài nhất**” (Longest matching).

* Yêu cầu:
* Từ điển
* Chuỗi đầu vào đã tách các dấu câu và âm tiết
* Ý tưởng:
* Tách câu từ văn bản -> Tách dấu câu ra khỏi câu -> Tách từ
* Thuật toán tham lam
* Đi từ trái sang phải hoặc phải sang trái, lấy các từ dài nhất có thể, dừng lại khi duyệt hết.
* Tìm ra tất cả các phương án tách từ có thể được sinh ra từ từ điển và lựa chọn phương án tốt nhất.
* Nhược điểm: Học sinh học sinh vật học -> Học\_sinh học\_sinh vật học (cải thiện: thuật toán so khớp từ cực đại)

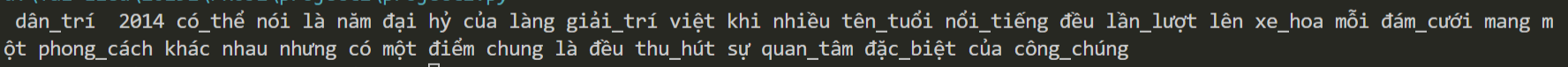


1. **Xử lý bằng Python**

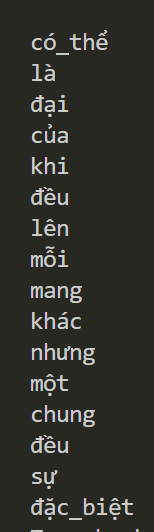
* Thuật toán:

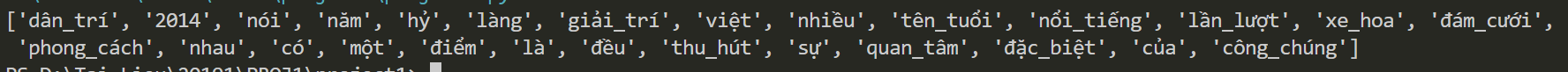


* Kết quả:



Sau khi loại bỏ **stopword**: sử dụng hàm **remove** để bỏ các stopword





1. **Machine Learning trong NLP**

**Phân loại văn bản bằng định lý Bayes**

1. **Ứng dụng**

* Hiểu được ý nghĩa, đánh giá, bình luận của người dùng từ mạng xã hội
* Phân loại email là spam hay không spam
* Tự động gắn thẻ cho những truy vấn, tìm kiếm của người dùng.
* Phân loại các bài báo điện tử.
* …

1. **Định lý Bayes**

* Là 1 phương pháp phân loại **có giám sát**, nội dung của văn bản đã được gán nhãn.
* Dễ hiểu và cài đặt
* Kết quả thu được tốt
* Là 1 phương pháp **rất quan trọng** trong NLP

1. **Bài toán phân loại văn bản**

* Vector đặc trưng **x** biểu diễn số lần xuất hiện các từ trong văn bản, **y** là các **nhãn** mà văn bản thuộc về.
* Tính giá trị **P(y|x)**, nếu giá trị này lớn hơn 1 giá trị **t** cho trước, ta kết luận nhãn của vector x là y.
* 10 thể loại: Chính trị xã hội, đời sống, khoa học, kinh doanh, pháp luật, sức khỏe, thế giới, thể thao, văn hóa, vi tính.
* Các bước thực hiện:

1. Chuẩn bị dữ liệu (Data Preparation): Phần B
2. Xử lý thuộc tính của dữ liệu (Feature Engineering)
3. Xây dựng mô hình (Build Model)
4. Tinh chỉnh mô hình và cải thiện hiệu năng (Improve Performance)