Báo cáo tìm hiểu đề tài cuối kỳ

**Thành viên và phân công**

Danh sách các thành viên nhóm 5.

|  |  |  |
| --- | --- | --- |
| Tên | MSSV | Đóng góp (%) |
| Nguyễn Xuân Huy | 0712196 |  |
| Trần Văn Tiến | 0712446 |  |
| Diệp Thế Nghĩa | 0712301 |  |

# Giới thiệu về đề tài :

Đề tài của nhóm là tìm hiểu về phương pháp kiểm tra văn bản có ẩn dữ liệu hay không bằng Word Location . Các kỹ thuật ản dữ liệu được kiểm tra là NICETEXT, TEXTTO và Markov Chain. Từ phương pháp này , một machine sẽ được tạo ra và huấn luyện để phân biệt văn bản có nhúng bằng ba phương pháp nhúng ở trên không ?

# NICETEXT, TEXTO, Markov Chain

## NICETEXT

|  |  |  |  |
| --- | --- | --- | --- |
| **Style** | **Type** | **Bit** | **Word** |
| Footballer Fb\_Verb Fb\_Object | Footballer | 0 | Striker |
| 1 | Middlefier |
| Fb\_Verb | 0 | take |
| 1 | stop |
| Fb\_Object | 0 | a pass |
| 1 | a shot |

### TEXTTO

|  |  |  |
| --- | --- | --- |
| **TYPE** | **Index** | **Word** |
| THING | 21 | referre |
| ADVERB | 31 | wrongly |
| VERB | 30 | decide |
| ADJECTIVE | 18 | stupid |
| PLACE | 30 | stadium |

iluvu -> 18 21 30 31 30 -> The stupid referre decide wrongly in the stadium.

## Markov Chain

# Word Location:

1. Word location:

Trong phương pháp này, chúng ta xem các câu liền kề nhau như là 1 đoạn văn bản. Vậy chúng ta có thể xem một phần của văn bản hay toàn bộ văn bản như 1 phân đoạn. Trong thí nghiệm này. Tất cả các công việc ta sẽ xét trên các phân đoạn văn bản.

Một văn bản được chia thành nhiều phân đoạn. Trong mỗi phân đoạn ta sẽ có các từ, và thứ tự của nó.



Với Wi, 0<= i<= n-1, là thứ tự của từ trong phân đoạn.

Vị trí của từ trong văn bản được xác định



Dể dàng nhận thấy rằng, 0< WLi<1.

Thực ra , trong 1 phân đoạn văn bản. Chỉ có xác định một số hữu hạn m các từ ( Wk’ ) và mỗi từ thì có nk hiển thị ( lúc chổ này lúc chổ khác). Do đó ta có 1 phép tính.



Nghĩa là: Số từ có trong đoạn văn bản, sẽ bằng tổng tập hợp các từ có trong văn bản đó nhân với số lần xuất hiện của chính nó ( tôi hiểu vậy ).

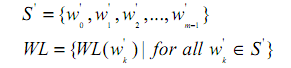
Đối với 1 từ Wk thì có nk hiển thị. Do đó ta nói từ Wk có một set các words location.

1 set vị trí của từ Wk  được biểu thị bằng LS ( Wk)



Đặt S’ là tập hợp các từ riêng biệt có trong Văn bản. WL là tập hợp địa chỉ của chúng.

Ta có công thức.

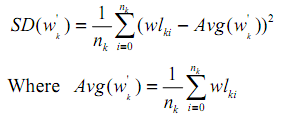


Và đây là dạng khác của biểu diển 1 đoạn văn bản.



1. Statistical Variables of Word Location ( thống kê biến đổi WL).

Trong văn bản tự nhiên. Sự lặp lại của một từ luôn thể hiện sự phân bổ không đồng đều. Một từ có thể xuất hiện nhiều ở một đoạn nào đó, nhưng ở chổ khác thì hiếm khi xuất hiện. Đặt SD ( chỉ số dàn trải : spread degree) để biểu thị điều đó.



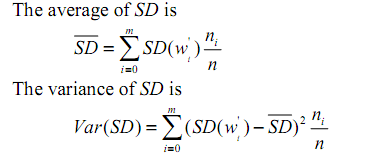
Từ công thức của một đoạn văn bản



Cho thấy,

Mỗi phân đoạn văn bản sẽ là tập hợp của một số từ ( dĩ nhiên là mỗi từ sẽ có 1 tập hợp vị trí của nó)

Chúng ta có thể do được sự phân bổ của từ trong một phân đoạn văn bản bằng cách kiểm tra sự phân bổ của chỉ số SD.



Giá trị SD trung bình và VarSD biểu thị đặc trưng phân bổ của 1 từ xác định. Do đó có thể dùng nó như 1 luật để phân loại văn bản thường và văn bản đã có ẩn dữ liệu.

<có nên cho ví dụ về 1 đoạn văn bản cụ thể hay không ? >

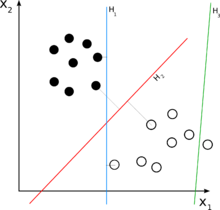
# Support Vector Machine

## Khái niệm

Support vector machines (SVMs) là phương pháp học giám sát gồm phân tích dữ liệu và nhận dạng mẫu.. Thuật toán SVM đầu tiên được Vladimir Vapnik giới thiệu và ông cùng Corinna Cortes đã đề xuất chuẩn SVM đang được sử dụng hiện nay.

SVM là phương pháp phân loại tuyến tính. Các mẫu sẽ được phân loại dựa trên các tính chất khác nhau của mẫu đó. Ví dụ một để phân loại cho các đối tượng thuộc một trong hai lớp khác nhau có n tính chất dùng để phân loại. Ta thể hiện n tính chất này thành trong không gian vector n chiều và tìm một siêu phẳng n-1 chiều để phân tách thành 2 lớp khác nhau.

Ví dụ có 2 lớp Táo và Lê, các tính chất dùng để phân biệt quả táo hay quả lê là chiều dài và cân nặng. Ta lấy thể hiện các tính chất này lên đồ thị 2 chiều thì được hình dưới. Với x1 thể hiện chiều dài ,x2 thể hiện cân nặng . Các điểm trắng thể hiện các quả lê trong khi các điểm đen là các quả táo.



Sau đó ta tìm một đường thẳng để tách 2 điểm táo và lê. Đường thẳng này được gọi là siêu phằng 1 chiều. Do có nhiều đường thẳng hợp phù hợp nên ta chon đường thẳng các xa cách xa các điểm trằng và đen nhất. Trong đồ thị trên thì đường phân loại là đường H2 màu đỏ .Khi đó nếu một quả có chiều cao là x1 và cân nặng là x2 nếu điểm (x1,x2) nằm bên trái đường thẳng H2 trên thì nó là quả táo , ngược lại nó là quả lê.

Trong trường hợp đối tượng được xét không có tính chất biến đổi tuyến tính. Ví dụ như các điểm táo và lê trong hính 1 nằm xen kẽ nhau không vẽ đường phân cách được ta phải dùng hàm kernel để mở rộng số chiêu không gian tính chất .Khi đó ta có thể khảo sát các tính chất chiều dài và cân nặng trong không gian 3 chiều hoặc 4 chiều.

Quá trình sử dụng SVM được chia làm 2 bước .

* Bước huấn luyện ta phải đưa dữ liệu để nó tìm để nó tìm hàm phân loại tốt nhất. Ví dụ ta đưa một số quả táo và lê với chiều dài và cân nặng của chúng cho SVM. SVM sẽ thể hiện đồ thị tính chất như ở trên và tìm đường thẳng để phân loại.
* Bước kiểm tra người ta chỉ đưa các thuộc tính và máy phải phân biệt nó thuộc nhãn nào. Tỉ lệ phân loại chính xác của máy là cơ sở để xem phương pháp SVM có thành công không . Trong ví dụ trên ta sẽ đưa một quả biết chiều dài và cân nặng để SVM phân loại xem nó là quả táo hay quả lê. Sau một số lần phân loại của SVM ta sẽ tính được xác suất phân loại thành công của SVM. Nếu xác suất này thấp ta phải kiểm tra và bổ sung thêm các tính chất được dùng để phân loại như màu sắc, độ bóng.

So với các phương pháp phân loại khác như cây quyết định và phân loại Bayes thì SVM có thể xứ lý trong đối tượng có nhiều tính chất hơn do sử dụng không gian n chiều và siêu phẳng. Đồng thời SVM cũng khử nhiễu tốt hơn

Phương pháp SVM có cài đặt sẵn trong MATLAB . Nên ta dễ dàng thực hiện việc huấn luyện và kiểm tra theo SVM.

## Vai trò SVM trong bài toán này

Trong bài toán trên ta cần phân biệt hai lớp là dữ liệu không nhúng (được gọi là dữ liệu tốt) và dữ liệu có nhúng watermark (gọi là dữ liệu xấu). Các tính chất dùng để phân loại là và. Do chỉ tính được các tính chất này trong văn bản mà chưa có kiến thức nào về mối quan hệ giữa các tính chất với dữ liệu tốt hay dữ liệu xấu nên nểu không dùng các phương pháp máy học như SVM thì ta. Nếu SVM phân loại thành công (độ chính xác khi kiểm tra đạt khoảng 85% trở lên) thì ta có thể khẳng định có thể phân loại dữ liệu tốt và dữ liệu xấu nhờ vào các tính chất này . Ngược lại các tính chất chưa đủ để kiểm tra xem văn bản có đã được nhúng watermark bên trong hay không .

# Mô hình huấn luyện