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 ?

# 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 ).

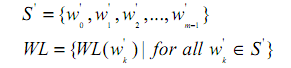
Đối với 1 từ Wk thì có nk hiển thị. Do đó ta nói từ Wk có một set các words location.

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.

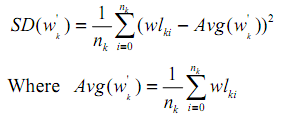


Và đây là dạng khác của biểu diển 1 đoạn văn bản.



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.



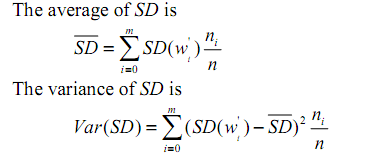
Từ công thức của một đoạn văn bản



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.

# 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. Ý tưởng chính của SVM: Là chuyển tập mẫu từ không gian biểu diễn Rn của chúng sang một không gian Rd có số chiều lớn hơn. Trong không gian Rd, tìm một siêu phẳng (hyperplane) tối ưu để phân hoạch tập mẫu này dựa trên phân lớp của chúng, cũng có nghĩa là tìm ra miền phân bố của từng lớp trong không gian Rn để từ đó xác định được phân lớp của 1 mẫu cần nhận dạng.

Ví dụ : để phân biệt hai loại trái cây táo và lề bằng SVM ta làm như sau :

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. Dữ liệu trong bước này là các cặp thuộc tính và nhãn tương ứng. 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 .

# Mô hình huấn luyện