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

# 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