Nội dung: tìm hiểu về bài toán phân loại - classification với thuật toán naive bayes và SVM
Để cho đơn giản ta sẽ lựa chọn bài toán phân loại văn bản 2 lớp.
Đầu vào: 100 văn bản lấy từ báo theo 2 chủ đề là (mỗi chủ đề tầm 50 bài)

- Bài báo về khoa học, công nghệ
- Bài báo về chính trị, xã hội

Bạn có thể lấy báo tiếng Anh hoặc tiếng Việt

Trong tập đầu vào, mỗi chủ đề lấy ra 45 bài (tổng 90 bài) làm tập training và 10 bài còn lại (mỗi chủ đề 5 bài) làm tập test

Sau khi chạy 2 thuật toán naive bayes và SVM ở trên, test lại model thu được với 10 bài còn lại trong tập test và đánh giá sai số.
Các đơn vị đánh giá gồm

- True
- False
- True Negative
- False Positive

Lần 2 bạn tăng gấp 2 bộ test (lấy thêm tầm 50 bài/chủ đề) và đánh giá lại sai số của các thuật toán với các đơn vị đo như trên

Gợi ý:

- Nên chọn văn bản tiếng Anh để tách từ đơn giản (tách bằng dấu cách trống)
- Xây dựng tập từ điển (chứa các từ khác nhau trong bộ văn bản ban đầu)
- Biểu diễn lại văn bản ban đầu theo dạng vector (với số chiều vector chính là số từ trong từ điển, từ nào có xuất hiện trong văn bản thì có thể biểu diễn bằng giá trị nhị phân 1|0 hoặc là tần số của từ đó trong văn bản)
- Áp dụng code của thuật toán phân loại văn bản theo naive bayes và SVM vào để phân loại