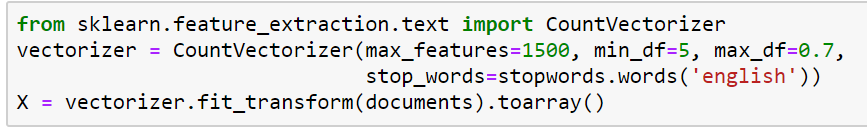
**LAB07 - RF IN NLP DATA**

1/ Thực hiện lại dự đoán nhận xét positive hay negative trên dữ liệu txt\_sentoken và làm thêm các chức năng sau :

* So sánh độ chính xác phân loại giữa Random Forest với SVM và Naïve Bayes
* Trong ví dụ mẫu ta vector hóa từng review trong documents thành một vector 1500 chiều. Bạn thử tăng giảm con số này từ 1000 cho đến 2000 và vẽ đồ thị cho biết độ chính xác tăng giảm ntn. Từ đó cho biết vector hóa document trong TH này nên dùng bao nhiêu chiều để đạt độ chính xác cao nhất



* Bạn thử viết hai reviews bằng tiếng Anh nhận xét về hai bộ phim nào đó (1 bộ phim bạn không thích và một bộ phim bạn không thích. Sau đó thử xem model được tạo ra có dự đoán chính xác nhận xét của bạn là negative hay positive không.

2/ Làm tương tự cho dữ liệu mới SMSSpamCollection về đánh giá một message là ham hay spam sau đó thực hiện các bước giống như câu 1

Gợi ý cách đọc dữ liệu: <https://archive.ics.uci.edu/ml/datasets/sms+spam+collection>

