NATURAL LANGUAGE PROCESSING

Lecturer: Doctor Bui Thanh Hung
Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city
Email: hung.buithanhcs@gmail.com (buithanhhung@iuh.edu.vn)
Website: https://sites.google.com/site/hungthanhbui1980/

Bài 1:
Dựa trên bộ dữ liệu tự thu thập từ các trang báo mạng Tiếng Việt (5 lớp/ mỗi lớp 10 mẩu tin), hãy thực hiện các yêu cầu sau:

1. Tiền xử lý dữ liệu với Beautiful Soup, re,…
2. Tách từ (Tokenize) sử dụng thư viện pyvi hay underthesea
3. Trích xuất đặc trưng TF-IDF bằng thư viện sklearn
4. Đánh giá bộ dữ liệu với giải thuật KNN bằng phương pháp 5-Fold (k-fold)
5. Huấn luyện dữ liệu cho bài toán phân loại văn bản với tỷ lệ dữ liệu 8:2 (8 phần train, 2 phần test) sử dụng đặc trưng TF-IDF và 2 giải thuật Bayes, SVM.
6. Tính độ đo F1 score
7. Tính độ đo Accuracy
8. Tính độ đo Confusion Matrix
9. So sánh kết quả các độ đo 6,7,8 với 2 giải thuật học máy ở trên
10. Lưu model với giải thuật đạt kết quả tốt nhất
11. Xây dựng ứng dụng phân loại văn bản với đầu vào là 1 văn bản bất kỳ có thể tự ngõ hay từ 1 file, in kết quả ra màn hình

Bài 2:
Cho ví dụ sử dụng HashVectorizer như sau:

```
from sklearn.feature_extraction.text import HashingVectorizer
corpus = [
     ' Hôm_nay tôi đi_học',
     ' Hôm_nay tôi đi_học ở trường',
     ' Hôm_nay tôi nghỉ ở nhà',
     ' Hôm_nay tôi có đi_học không?',
]
vectorizer = HashingVectorizer(n_features=2**4)
X = vectorizer.fit_transform(corpus)
print(X.shape)
```

Hãy giải thích chi tiết các kết quả đạt được tương tự như các bước tính TF-IDF.

Bài 3:
Sử dụng HashVectorizer thay cho đặc trưng TF-IDF ở bài 1