

RNN, GRU

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/>

Ở bài tập 1 – Homework 1 có các yêu cầu sau:

Dựa trên bộ dữ liệu tự thu thập từ các trang báo mạng Tiếng Việt (5 lớp/ mỗi lớp 10 mẫu 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 gõ hay từ 1 file, in kết quả ra màn hình

Hãy cải tiến bài 1 (nội dung 3-11) theo yêu cầu sau:

- 1- Sử dụng đặc trưng TF-IDF với 2 mô hình học sâu RNN và GRU
- 2- Sử dụng Word embedding của Keras với 2 mô hình học sâu RNN và GRU