Đồ án 3 - Fake news detection

Tổng quan:

Tin giả đang là một vấn đề nhức nhối của xã hội, đặc biệt trong thời đại bùng nổ thông tin hiện nay. Đồ án này sẽ cho phép các bạn làm việc với dữ liệu văn bản tiếng Việt, xây dựng mô hình dự đoán tin giả và deploy mô hình này lên một trang web đơn giản.

Nhiệm vụ:

- Tiền xử lý văn bản tiếng Việt
- EDA
- Xây dựng mô hình máy học
- Deploy mô hình

Mục tiêu:

Làm quen quy trình làm việc với bài toán khoa học dữ liệu với dữ liệu văn bản thô, biết cách sử dụng một số thư viện phổ biến

Chi tiết

1. Nguồn dữ liệu: VNFD Dataset

- Tập dữ liệu 223 record bản tin tiếng Việt, gồm 2 nhãn: 1 (tin giả) và 0 (tin thật)
- Mô tả dữ liệu: Mô tả tập VNFD

2. Tiền xử lý văn bản tiếng Việt

- Các bước tiền xử lý văn bản cơ bản gồm: lowercase, loại stopwords, stemming, normalize tùy theo từng lĩnh vực, loại noise (HTML tag, các ký hiệu đặc biệt như @, #,...), dấu câu
- Nguồn tham khảo: Sơ lược về tiền xử lý văn bản
- Trong đồ án này, các bạn cần quan sát tập dữ liệu và tự lựa chọn, thử nghiệm các phương pháp tiền xử lý.

Lưu ý 1: với bước loại stopwords và tokenize, các bạn có thể tham khảo một số nguồn cho tiếng Việt:

- Stopword: Nguồn
- Tokenizer: VnCoreNLP

Lưu ý 2: nếu sử dụng stopwords và tokenizer (hay các tool normalize cho tiếng Việt khác), các bạn vui lòng trích dẫn link nguồn đã tham khảo. Nếu sử dụng 1 trong 2 nguồn gợi ý ở trên thì chỉ cần ghi tên tool đã sử dụng trong bài.

3. EDA (khám phá dữ liệu)

- Kiểm tra dữ liệu bị thiếu, sai kiểu dữ liệu (nếu có)
- Kiểm tra phân bố các class có chênh lệch không
- Các thông tin thống kê của văn bản (chiều dài trung bình mỗi record, v.v)

4. Mô hình hóa

- Thư viện gợi ý: scikit learn
- Có thể chọn các mô hình tuyến tính hoặc không tuyến tính
- Cần chọn ít nhất 2 mô hình máy học trở lên (nhưng không cần quá nhiều), huấn luyện trên tập dữ liệu đã xử lý và đánh giá mô hình
- Sau khi huấn luyện xong, cần lưu lại các mô hình để thực hiện inference trong bước 5

Lưu ý: Dữ liệu đầu vào của mô hình máy học cần thuộc kiểu numerical, nên cần có bước rút trích đặc trưng của văn bản trước khi đưa vào mô hình học

5. Deploy mô hình

- Thư viện gợi ý: Streamlit

Streamlit cho phép code giao diện một cách đơn giản hoàn toàn bằng Python để hỗ trợ deploy mô hình máy học lên web miễn phí, phục vụ việc demo. Cách sử dụng thư viện khá đơn giản nên các bạn có thể tham khảo ý tưởng giao diện tại Streamlit gallery.

- Khi chấm demo, mình muốn nhập trực tiếp một đoạn văn bản vào, chọn 1 trong các mô hình huấn luyện và mô hình cần trả ra kết quả tin giả hay tin thật, do đó các bạn cần load các mô hình đã lưu ở bước 4.

- Sau khi đã test xong phần web demo ở local, các bạn tham khảo Deploy Streamlit và làm theo hướng dẫn để deploy trang web lên Internet.

  Lưu ý 1: Các bạn cần có tài khoản Github để host trang web này.
  Lưu ý 2: Do tập dữ liệu sử dụng khá nhỏ nên kết quả dự đoán khi mình test demo không yêu cầu hoàn toàn chính xác.
  Lưu ý 3: Phần giao diện chỉ cần dễ nhìn, đảm bảo cho phép người dùng nhập văn bản, chọn loại mô hình và hiển thị kết quả là tin giả hay tin thật.

- Các bạn có thể tham khảo giao diện sau để hiểu cách mình sẽ test demo: Demo

Yêu cầu

1. Code

- Làm trực tiếp trên các file notebook .ipynb
- Code app streamlit .py

2. Báo cáo

Viết trực tiếp trong các ô markdown của file notebook đã code.

3. Bảng phân công công việc

Đầu notebook của mỗi nhóm cần có bảng danh sách tên và phân công công việc của các thành viên.
Các bạn lưu ý chia việc hợp lý, mỗi thành viên cần đảm nhận lượng công việc tương đương nhau.

4. Lưu ý

Bài làm giống nhau 0 điểm cả môn.
Ghi rõ nguồn tham khảo đầy đủ.
Cần trình bày phần báo cáo rõ ràng, dễ hiểu.

5. Nộp bài

Folder bài nộp cần có:

- File notebook của phần code (.ipynb)
- File app streamlit (.py)
- Bản pdf của các file notebook trên (export file .ipynb thành PDF)
- Đầu file notebook cần dẫn link trang web đã deploy

Nén folder thành một file, đặt tên theo cú pháp sau và nộp qua moodle:

<MSSV1>_<MSSV2>_<MSSV3>_<MSSV4>_<MSSV5>.zip

Thông tin liên hệ

TA: Nguyễn Ngọc Băng Tâm
Nếu có thắc mắc các bạn vui lòng liên hệ qua: bangtamnguyenn@gmail.com
với tiêu đề [NMKHDL] Project 03 - <vắn tắt chủ đề muốn hỏi>