fit@hcmus

Đại học Quốc gia Thành phố Hồ Chí Minh
Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

Phân tích dữ liệu thông minh.

Đồ án thực hành

Natural Language Processing with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not

1. Giới thiệu

Đồ án này sẽ dựa trên cuộc thi Natural Language Processing with Disaster Tweets trên Kaggle

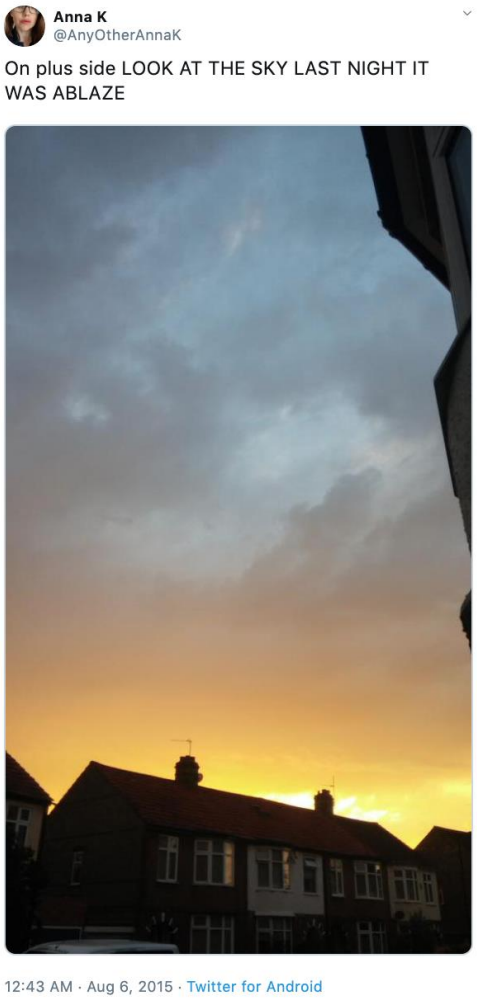
Mô tả:

Twitter đã trở thành một kênh liên lạc quan trọng trong trường hợp khẩn cấp.

Sự phổ biến của điện thoại thông minh cho phép mọi người thông báo các trường hợp khẩn cấp mà họ đang quan sát thấy trong thời gian thực. Do đó, nhiều tổ chức quan tâm đến việc xây dựng một chương trình tự động theo dõi Twitter nhằm phát hiện các tin khẩn cấp được người dùng đăng lên Twitter (chẳng hạn như các tổ chức cứu trợ thảm họa và hãng thông tấn báo chí).

Tuy nhiên, không phải lúc nào bài đăng của người dùng cũng có thể xác định rõ đó có thực sự thông báo về một thảm họa hay không.

Ví dụ, xem qua bài Tweet sau



Hình 1 - Ảnh chụp một Tweet [1]

Trong bài đăng này có từ "ABLAZE", nhưng không phải nói về thảm họa.

Trong cuộc thi này, các nhóm cần xây dựng một mô hình học máy dự đoán Tweet nào nói về thảm họa thực sự và Tweet nào không. Bạn sẽ có quyền truy cập vào tập dữ liệu gồm 10.000 tweet đã được phân loại thủ công.

2. Yêu cầu của đồ án

Các nhóm sẽ tham dự cuộc thi trên. Download bộ dự liệu, xây dựng mô hình dự đoán, submit lên Kaggle, xem kết quả.

a) Tổ chức thư mục cho đồ án: các nhóm có thể dùng một hay nhiều file notebooks nhưng phải được tách biệt rõ ràng cho từng giai đoạn: khám phá dữ liệu, tiền xử lý dữ

liệu, rút trích đặc trưng, xây dựng mô hình, đánh giá và phân tích kết quả.

b) Nên có nhiều hình vẽ biểu đồ trực quan để giải thích và trình bày trong quá trình làm.

c) Phải có giải thích rõ ràng cho mọi cell code trong file jupyter notebook. Tức là, mỗi cell code nên có một cell markdown kèm theo để giải thích.

3. Quy định nộp bài:

Mỗi nhóm cần nộp:

- File jupyter notebook. File này sẽ bao gồm source code và báo cáo. Nội dung cần trình bày bao gồm
    - Giới thiệu thành viên và phân công công việc
    - Giới thiệu chung về đồ án
    - Quá trình khám phá dữ liệu
    - Quá trình tiền xử lý dữ liệu
    - Quá trình tạo đặc trưng (feature engineering)
    - Xây dựng mô hình (Có thể dùng nhiều mô hình)
    - Thử nghiệm:
        - Training
        - Testing
        - Kết quả đạt được, nhận xét và phân tích
    - Kết luận và hướng phát triển. (Cần đưa kết quả trên leaderboard của Kaggle)
- Các mã nguồn python khác (nếu có)
- Một đoạn video ngắn (15-20p) trong đó nhóm trình bày đồ án của mình.

4. Thang điểm:

a. Kết quả trên Leaderboard (2đ) – Không tính Perfect score. (Dự kiến):

Top 10%: 2đ

Top 20%: 1.5đ

Top 30%: 1đ

Top 40%: 0.5 đ

b. Nội dung file notebook + code (5đ)

   Theo yêu cầu như phần 3. Các nhóm cần trình bày rõ ràng, chi tiết về mô hình mình đã dùng để đạt được kết quả như phần 4a.

c. Nội dung trình bày trong video (3đ)

5. Tham khảo

[1] https://www.kaggle.com/competitions/nlp-getting-started/overview