CASE STUDY 03

Bài toán: "News Headlines Dataset For Sarcasm Detection - High quality dataset for the task of Sarcasm Detection" (Phát hiện tin bài châm biếm - Bộ dữ liệu chất lượng cao từ các trang tin chuyên nghiệp)

1. GIỚI THIỆU CHUNG

- "Phát hiện tin bài châm biếm" phát sinh từ bài toán gốc "Phát hiện tin giả mạo" - thường dùng cho mạng xã hội (vốn tin bài trên mạng xã hội viết với văn phong không chính thống)

- Đây là bài toán cũng đang khá hot, được nhiều người quan tâm (đặc biệt là Facebook đang gặp khó khăn trong vấn đề lọc tin)

- Thông tin về Dataset:

+ được crawl (thu thập) từ 2 trang web: 1) TheOnion toàn đăng những tin châm biếm và 2) HuffPost bao gồm những tin chính thống

+ Chỉ crawl Headline (Tiêu đề) của các bài viết.

+ Thống kê: dataset bao gồm:

Statistic/Dataset	Headlines
# Records	26,709
# Sarcastic records	11,725
# Non-sarcastic records	14,984

+ Nguồn dữ liệu:

@dataset{dataset,

author = {Misra, Rishabh},

year = {2018},

month = {06},

pages = {},

title = {News Headlines Dataset For Sarcasm Detection},

doi = {10.13140/RG.2.2.16182.40004}

}

2. YÊU CẦU

- Dự đoán 1 Headline x có phải là tin châm biếm hay không (bài toán Binary Classification)
- Dùng phương pháp Logistic Regression
- So sánh với kết quả của Naive Bayes, ...Liệu KNN, Decision Tree có phù hợp để áp dụng giải bài này không?Trả lời có phân tích diễn giải và kiểm nghiệm lại bằng thực nghiệm (nếu được)
- Viết báo cáo như các case study trước
- Lưu ý: Thực hiện 2 level :

+ Level 1: SV tự code lại các giải thuật

+ Level 2: Sử dụng thư viện có sẵn của Python/Matlab → cũng phải code nhưng gọi hàm trong thư viện

+ Level 3 (không khuyến khích): nếu không code được thì chạy tool của người ta để xem trả về kết quả thế nào.

3. HƯỚNG DẪN SƠ BỘ

Để làm được bài này thì phải xử lý được dữ liệu text.

Phase 0: Đăng ký tài khoản trên trang : https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection

và tự down dataset về.

Phase 1: Biểu diễn text thành vector (Tìm hiểu mô hình Vector Space Model trong lĩnh vực Information Retrieval)

Phase 2: Huấn luyện và dự đoán

- Thống nhất chung cả lớp để so sánh kết quả: 20.000 records đầu tiên để train và 6709 records còn lại để test.
- Tính accuracy, precision, recall, F1-score.

4. Notes:

- GV sẽ có 1 buổi hướng dẫn riêng (ngoài buổi học lý thuyết) về Information Retrieval