ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Báo cáo bài tập 1
Xử lý ngôn ngữ tự nhiên
Naive Bayes

GV hướng dẫn: TS. Nguyễn Thị Quý

Họ và tên	*MSSV*	*Mã lớp*
Trần Hoàng Bảo Ly	21521109	CS231.N21.KHTN
Lê Thanh Minh	21520063	CS231.N21.KHTN
Nguyễn Quốc Trường	21521604	CS231.N21.KHTN
Lê Thu Hà	21520800	CS231.N21.KHTN
Trần Xuân Minh	21520352	CS231.N21.KHTN

Hồ Chí Minh, tháng 4 năm 2023

Mục lục

1 Giới thiệu bài toán

1.1 Phân tích quan điểm (Sentiment Analysis)

Phân tích quan điểm là một ứng dụng của trí tuệ nhân tạo, nó sử dụng các thuật toán phức tạp để xử lý ngôn ngữ tự nhiên của con người (NLP) và xác định các đặc điểm cảm xúc tiêu cực/tích cực/trung lập, tại một thời điểm thông qua văn bản hoặc lời nói. Trong khuôn khổ môn học chúng ta chỉ làm việc với văn bản (text)

Hình 1: Mô tả về phân tích quan điểm (Nguồn: capitalhub)

Dựa trên phân tích ngôn ngữ tự nhiên dạng văn bản (text), để xác định xem liệu quan điểm của người viết về một vấn đề gì đó là gì.

1.2 Phân tích quan điểm dựa trên khía cạnh (aspect-based sentiment analysis - ABSA)

Phân tích quan điểm dựa trên khía cạnh (ABSA) là một kỹ thuật phân tích văn bản giúp phân loại dữ liệu theo khía cạnh và xác định quan điểm được gán cho từng khía cạnh đó. ABSA có thể được sử dụng để phân tích phản hồi của khách hàng bằng cách liên kết quan điểm cụ thể với các khía cạnh khác nhau của sản phẩm hoặc dịch vụ.

Ở đây, khía cạnh đề cập đến các thuộc tính hoặc thành phần của sản phẩm hoặc dịch vụ, ví dụ: giá cả, chất lượng, tốc độ sản phẩm/dịch vụ, ngoại hình/giao diện sản phẩm, ...

Như vậy ABSA có thể trích xuất:

- Quan điểm: ý kiến tích cực hoặc tiêu cực về một khía cạnh cụ thể.
- Các khía cạnh: danh mục, tính năng hoặc chủ đề đang được nói đến

1.3 Naive Bayes

Thuật toán Naive Bayes (như chúng ta đã biết trong chương 4) có thể được sử dụng để xử lý bài toán phân tích quan điểm, cụ thể ở đây là Multinomial Naive Bayes (MNB). Tuy nhiên, với bài toán phân tích quan điểm dựa trên khía cạnh mà bộ dữ liệu đặt ra, chúng ta phải làm sao?

2 Giải quyết

2.1 Giới thiệu

Trong bài toán về nhà hàng và khách sạn, các khía cạnh được quan tâm gồm: Thực thể (ENTITY), thuộc tính (ATTRIBUTE), và quan điểm (SENTIMENT) dựa trên khía cạnh đó.

2.2 Hướng giải quyết

Với bài toán phân tích quan điểm dựa trên khía cạnh, mỗi lời nhận xét có thể bao gồm nhiều khía cạnh. Giả sử mỗi một câu là một quan điểm về một khía cạnh nào đó. Vậy nên với mỗi một lời nhận xét chúng ta cố gắng tách thành các câu có nghĩa đầy đủ, mỗi một câu là một khía cạnh.

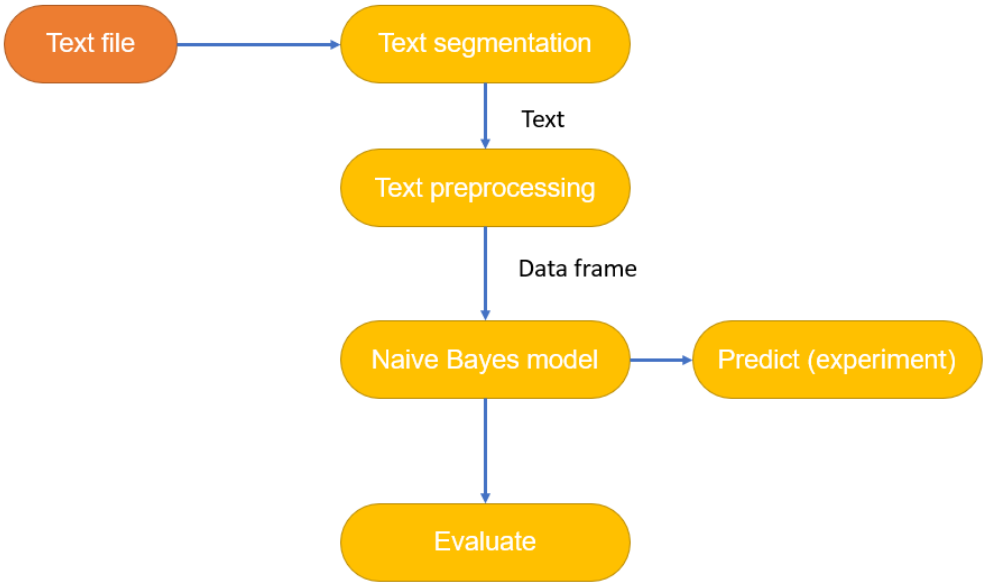
Mỗi một câu đi kèm với nó là ENTITY, ATTRIBUTE và SENTIMENT, việc của chúng ta là cố gắng dự đoán câu trên nói về cái gì, nói về vấn đề gì, và quan điểm của câu nói đó là gì.

Như vậy chúng ta sẽ cần 3 MNB model để giải quyết bài toán trên.

- Một model để xác định thực thể (Entity)
- Một model để xác định vấn đề (Attribute)
- Một model để xác định quan điểm (sentiment)

3 Quy trình thực hiện

3.1 Sơ đồ thực hiện



Hình 2: Quy trình thực hiện

3.2 Giải thích các bước

.

- Text_segmentation: Sau khi đọc file text, text gốc được tách thành các câu khác nhau, dựa vào các nhãn là các aspect cho trước.
- Text_preprocessing: Xử lý text đã được tách bằng cách loại bỏ các dấu câu, thực hiện ghép các từ đơn thành các từ ghép có nghĩa, chuẩn hóa. Sau đó sắp xếp thành dataframe với các aspect và sentiment tương ứng.
- Model: model được sử dụng là Multinomial Naive Bayes (từ thư viện sci-kit learn). Tiến hành fit model với các features được chọn theo phương pháp bag of words.
- Predict: Tiến hành thử nghiệm model với các text cụ thể, đánh giá kết quả sơ bộ, đặc biệt là với bài toán phân tích quan điểm dựa trên khía cạnh.
- Evaluate: đánh giá hiệu suất tổng thể của model trên cả tập train và tập test.

4 Đánh giá hiệu suất thực hiện

Phần đánh giá hiệu suất thực hiện đã được bao gồm trong notebook nhằm tăng tính thực tiễn.

Link đến github của assignment tại đây