|  |
| --- |
| **ĐẠI HỌC BÁCH KHOA HÀ NỘI**  **Trường Công nghệ Thông tin và Truyền thông**  **Description: logo_128**  **BÀI TẬP LỚN IT3940**  **Đề tài: Phân tích cảm xúc theo khía cạnh bình luận Lazada**  Mã lớp: 720287  GV hướng dẫn: **PGS.TS. Nguyễn Thị Kim Anh**  Sinh viên thực hiện:  **Lê Thành Trung 20194192**    **Hà Nội, 2022** |

**Mục Lục**

[Chương 1. Đặt vấn đề 3](#_Toc23410)

[Chương 2. Tổng quan hệ thống và công nghệ sử dụng 4](#_Toc17727)

[2.1 Data Pipeline 4](#_Toc2625)

2.2 Công nghệ sử dụng.................................................................................... 4

[2.2.1 Kafka 4](#_Toc7290)

[2.2.2 MongoDB 7](#_Toc200)

[Chương 3. Triển khai 7](#_Toc8308)

[3.1. Lấy dữ liệu bình luận từ Lazada 7](#_Toc10601)

[3.2. Triển khai Kafka 7](#_Toc21372)

[3.3. Xử lý dữ liệu 8](#_Toc129)

[3.4. Lưu dữ liệu vào mongoDB 9](#_Toc13624)

[Chương 4. Các hạn chế của giải pháp 10](#_Toc28305)

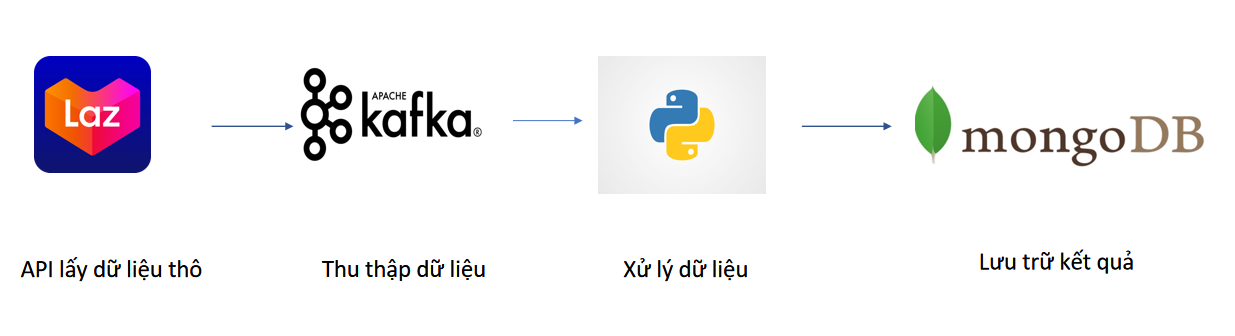
**Chương 1. Đặt vấn đề**

Trong thời đại 4.0 hiện nay, lượng dữ liệu được sản sinh với tốc độ chóng mặt từ các trang mạng xã hội, trang thương mại điện tử, các cảm biến GPS, ... Điều này khiến cho các công cụ truyền thống trở khó khăn trong việc lưu trữ, xử lý loại dữ liệu này. Vì vậy, công nghệ Big Data đã được phát triển để giải quyết vấn đề này.

Bài toán phân tích cảm xúc từ bình luận từ mạng xã hội, hoặc trang thương mại điện tử không phải đề tài mới. Tuy nhiên việc phân tích để biết được mỗi bình luận đang nói về những khía cạnh nào chưa có nhiều. Với cách phân tích các bình luận nói về một sản phẩm nào đó chúng ta có thể rút ra được độ tốt của các khía cạnh của sản phẩm. Do đó giúp cho người dùng có thể đưa ra quyết định mua hàng phù hợp hơn.

**Chương 2. Tổng quan hệ thống và công nghệ sử dụng**

**2.1 Data Pipeline**



Hệ thống được chia làm 4 pha:

- Sử dụng API của Lazada, thư viện requests python để lấy dữ liệu bình luận

- Public dữ liệu theo các danh mục sản phẩm lên Kafka

- Lấy dữ liệu từ Kafka, dùng NLP python để trích xuất cụm N-A và map vào bộ từ điển xây dựng sản để chuẩn hóa kết quả

- Lưu trữ kết quả vào mongoDB

**2.2 Công nghệ sử dụng**

**2.2.1 Kafka**

Diagram, schematic

Description automatically generated

Apache Kafka là một hệ thống phân tán message – distributed messaging system, Kafka được phát triển và duy trì bởi Apache, vì thế Kafka (message broker) có tên là Apache Kafka. Nền tảng này cho phép người dùng đọc, ghi, lưu trữ dữ liệu và xử lý các sự kiện (còn được gọi là records hoặc message) thông qua nhiều máy chủ khác nhau.

Giống như những chương trình/ phần mềm message broker khác, nó được phát triển theo mô hình public/subscribe. Bên public dữ liệu được gọi là producer và bên nhận dữ liệu theo các topic sẽ được gọi là consumer.

Diagram

Description automatically generated

Kafka gồm các thành phần chính sau:

### Producer: Trước khi xử lý, nó sẽ thực hiện phân loại và lưu trữ các message dựa theo topic của chúng. Producer có nhiệm vụ publish message vào các topic thích hợp. Sau đó, khi dữ liệu được gửi đến partition của topic được lưu trữ tại Broker.

* Consumer: Các consumer sẽ được định danh theo từng group name, nhiều consumer có thể cùng nằm trong một topic.

### Partition: Partition là nơi lưu trữ dữ liệu cho các topic, mỗi topic có thể có một hoặc nhiều Partition khác nhau. Mỗi Partition sẽ được gán một ID và lưu trữ dữ liệu có định. Trong 1 Kafka cluster, một Partition có thể được sao chép thành nhiều bản và có một bản chính leader chịu trách nhiệm đọc ghi dữ liệu; các bản còn lại sẽ có tên là follower. Khi bản chính bị lỗi,sẽ có một bản follower lên thay thế thành leader

### Broker: Kafka cluster là một nhóm các server và mỗi nhóm server này sẽ được gọi là broker.

### Zookeeper: Zookeeper là một chức năng nhằm để quản lý và bố trí lại các broker.

  Một mô hình cấu trúc kafka đơn giản

Diagram

Description automatically generated

**2.2.2 MongoDB**

**MongoDB**là một database hướng tài liệu (document), một dạng NoSQL database. Vì thế, MongoDB sẽ tránh cấu trúc table-based của relational database để thích ứng với các tài liệu như JSON có một schema rất linh hoạt gọi là BSON. **[MongoDB](https://topdev.vn/viec-lam-it/mongodb-kt90)** sử dụng lưu trữ dữ liệu dưới dạng Document JSON nên mỗi một collection sẽ các các kích cỡ và cácMMMM document khác nhau. Các dữ liệu được lưu trữ trong document kiểu JSON nên truy vấn sẽ rất nhanh., cho phép dữ liệu lưu trữ không cần tuân theo một cấu trúc nhất định.

**MongoDB** đã trở thành một trong những NoSQL database nổi trội nhất bấy giờ, được dùng làm backend cho rất nhiều website như eBay, SourceForge và The New York Times.

**Chương 3. Triển khai**

**3.1. Lấy dữ liệu bình luận từ Lazada**

- API để lấy ID các sản phẩm của một danh mục với tham số là tên danh mục sản phẩm của Lazada

VD: dien-thoai-di-dong, may-tinh-bang, do-gia-dung-nha-bep,...

https://www.lazada.vn/{}/?ajax=true&page={}&spm=a2o42.searchlistcategory.cate\_5.9.46281e22mYNSDT

- API để lấy các bình luận của một sản phẩm với tham số là ID của sản phẩm được lấy từ bước trên

https://my.lazada.vn/pdp/review/getReviewList?itemId={}&pageSize=50&filter=0&sort=0&pageNo=1

**3.2. Triển khai Kafka**

- Khởi tạo server Kafka và tạo topic là danh mục sản phẩm

+ Khởi tạo zookeeper:

.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties

+ Khởi tạo kafka server:

.\bin\windows\kafka-server-start.bat .\config\server.properties

+ Tạo topic

.\bin\windows\kafka-topics.bat --delete --bootstrap-server localhost:9092 --topic dien-thoai-di-dong

- Sau đó tạo producer để gửi luồng dữ liệu vào topic tương ứng (danh mục sản phẩm)



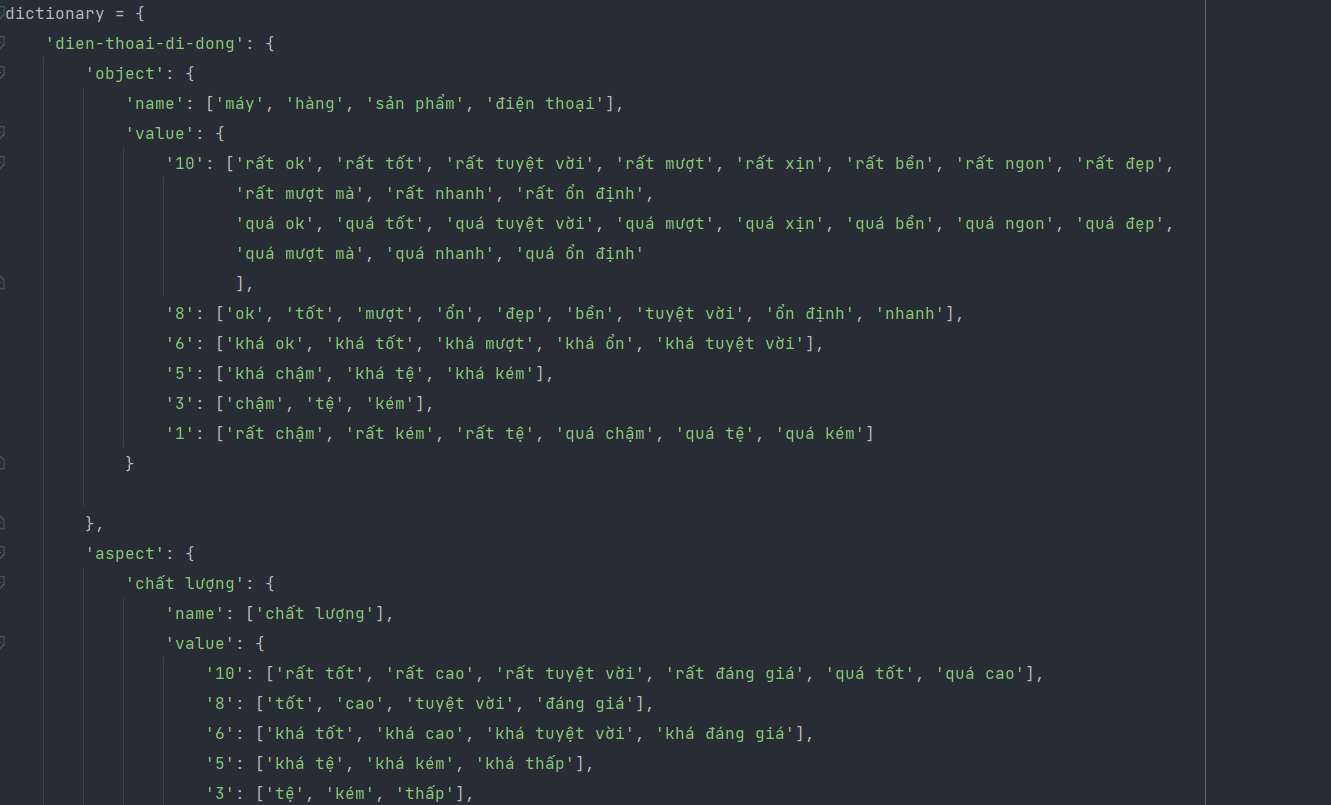
**3.3. Xử lý dữ liệu**

- tạo consumer để lấy dữ liệu từ kafka

- tiền xử lý dữ liệu văn bản tiếng Việt: xử lý viết tắt, teencode, gõ dấu tiếng Việt, ...

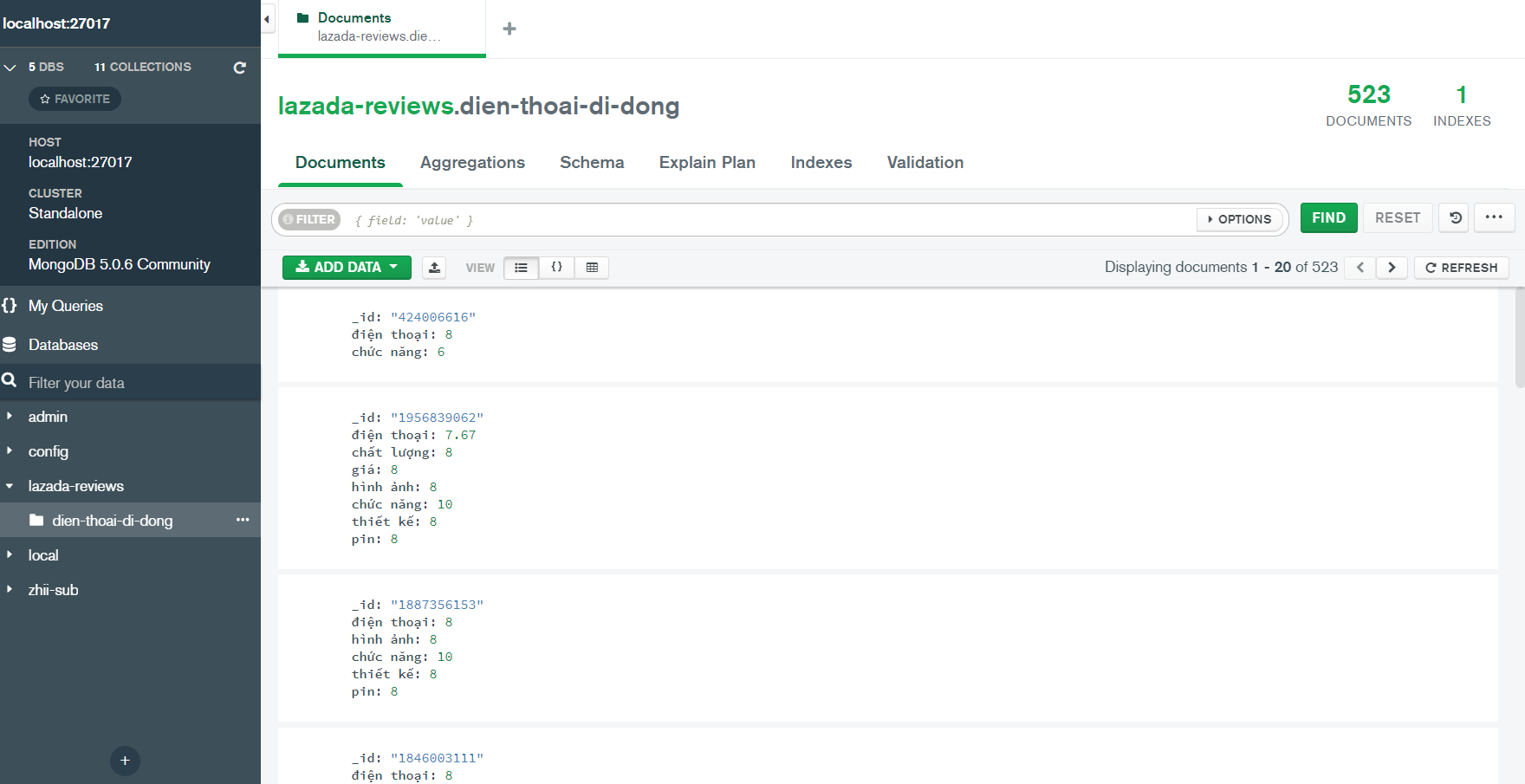
- dùng thư viện underthesea để trích xuất N-A dưa vào pos\_tag

- chuẩn hóa kết quả bằng cách so khớp với từ điển được xây dựng sẵn



**3.4. Lưu dữ liệu vào mongoDB**

- dữ liệu được lưu vào mongoDB



**Chương 4. Các hạn chế của giải pháp**

- Dữ liệu bình luận được lấy từ Lazada từ nhiều đối tượng người nên văn phong, cấu trúc bình luận không được nhất quán. Đa phần bình luận kiểu văn nói nên việc xác định cấu trúc để áp dụng tập luật dựa vào dependency\_parsing mang lại kết quả không cao. Giải pháp hiện tại là sử dụng pos\_tag lấy danh từ và tính từ theo sau. Mặc dù đem lại kết quả tốt hơn nhưng không phải cách giải quyết tốt

=> Cần có một cơ chế lọc bình luận tốt hơn

- Do các người dùng khác nhau không sử dụng một từ duy duy nhất để chỉ một khía cạnh (ví dụ: thiết kế, vẻ ngoài, bề ngoài,... ) nên cần có giải pháp để đưa về cùng một từ chỉ khía cạnh đó.

Giải pháp hiện tại: xây dựng bộ từ điển để quy đổi danh sách các từ chỉ cùng một khía cạnh về một từ duy nhất. Tuy nhiên cách làm thủ công này không khả thi nếu muốn tổng quát, mở rộng

=> Cần phương pháp xác định các từ chỉ cùng một khía cạnh

- Các tính từ, các từ chỉ sắc thái của tính từ tạo ra các mức độ nhận xét khác nhau cho từng khía cạnh.

Giải pháp hiện tại: xây dựng bộ từ điển, gán các giá trị rời rạc (điểm từ 1 đến 10) cho các tính từ. Cách làm này mang tính chủ quan chỉ dựa trên kinh nghiệm và hoàn toàn thủ công nên cũng không phù hợp cho trường hợp tổng quát

=> Cần phương pháp xác định mức độ sắc thái cho mỗi tính từ đối với các khía cạnh.