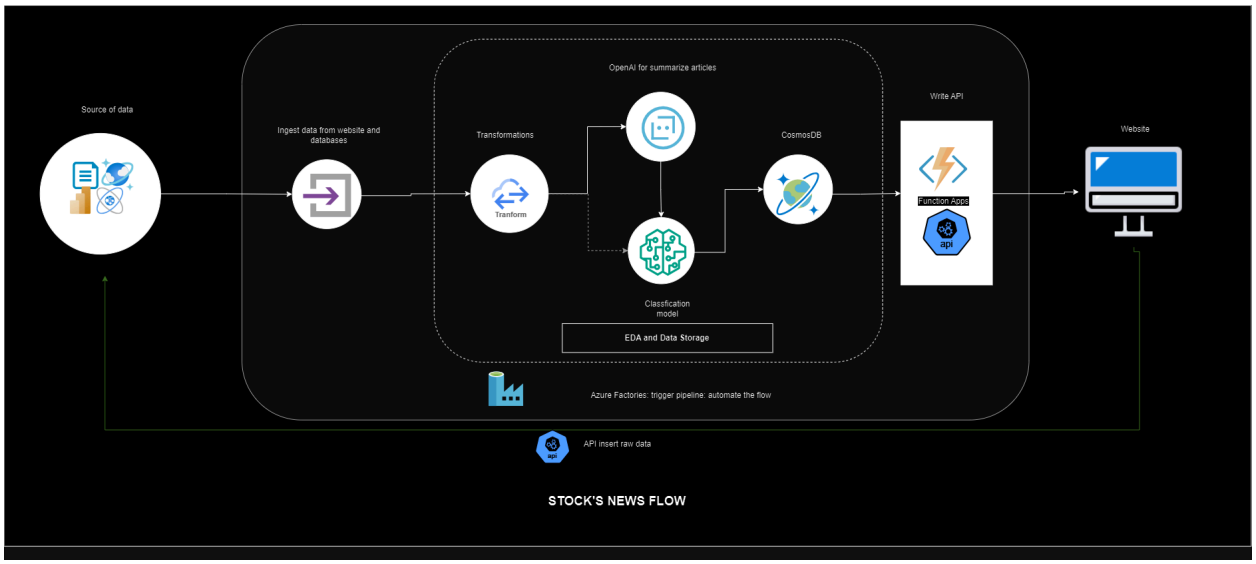
Giới thiệu

Team gồm 3 thành viên:
Hà Thanh Hương
Nguyễn Thị Minh Phượng
Nguyễn Ngọc Đức
Chủ đề: Đánh giá cảm xúc/phân loại tin tức chứng khoán

I: Workflow chung



1. Input data

- Nguồn từ các website khác (cafeF, 24h,...): Qua quá trình crawl data. Đây là phần dữ liệu để trainning và dữ liệu được lưu trữ lại để end-user có thể tìm kiếm. File .py: **\Code-crawl-and-clean-data\Crawling data with article manager.ipynb**
- Website của team: Qua chức năng tạo mới dữ liệu.
    - Sử dụng API insert: file .py: **\sentiment-analytics\insert-article-in-db\init.py**
    - **=> /code-model/SummarizeArticle2Model.ipynb**
    - **=> /code-model/phoBert_Inferrence.ipynb**

2. Xử lý data

- Sử dụng Azure Databricks để xử lý dữ liệu.
- File .py: **\Code-crawl-and-clean-data\article_managers.py**

3. Tóm tắt nội dung bài báo theo từng mã chứng khoán

- Viết trên Azure Databricks.

- Sử dụng OpenAI và các thành viên của team tóm tắt bài báo theo từng mã cổ phiếu và gán nhãn dữ liệu làm training set.
- Sử dung OpenAI để tóm tắt bài báo theo từng mã cổ phiếu: file .py: **/code-model/SummarizeArticle2Model.ipynb**

4. Gán nhãn dữ liệu: positive/negative/neutral:

- Viết trên Azure Databricks.
- Sử dụng OpenAI và các thành viên của team tóm tắt bài báo theo từng mã cổ phiếu và gán nhãn dữ liệu làm training set.
- Viết Sentiment analytics model (training): **/code-model/phoBert_Training.ipynb**
- Viết Sentiment analytics model (Inferrence): **/code-model/phoBert_Inferrence.ipynb**

5. Lưu trữ dữ liệu

- Sử dụng Azure CosmosDB để lưu trữ dữ liệu (hệ thống cloud)
- Dữ liệu dạng json
- Primary key: Mã chứng khoán, URL

6. API

- API search: Mục đích để lấy ra thông tin Mã chứng khoán (nội dung, thời gian tạo, nhãn dán) theo yêu cầu của end-users (khoảng thời gian, mã chứng khoán hoặc nhãn dán)
    - Chi tiết API search: file document: **/API specification.xlsx** và file .py: **/sentiment-analytics/get-stock-information/__init__.py**
- API insert: Mục đích để thêm mới bài báo
    - Chi tiết API insert: file document: **/API specification.xlsx** và file .py: **/sentiment-analytics/insert-article-in-db/__init__.py**

7. Website

- Front-end:
    - Bao gồm 2 chức năng search data và insert data.
    - Sử dụng ngôn ngữ HTML/CSS: để viết front-end.
    - File: **/Website_SA/index.html** và **/Website_SA/dashboard.html** và **/Website_SA/font-type.css**
- Back-end:
    - Mục đích để gọi API search và API insert data, kiểm tra required fields và logics kết quả.
    - Sử dụng ngôn ngữ javascript: để viết back-end.
    - File: **/Website_SA/xlsx.full.min.js** và **/Website_SA/jszip.js và /Website_SA/action**
- Folder ảnh: Chứa các ảnh cho màn hình. **/Website_SA/image/**
- Folder files: Chứa file template (trong phần insert data): **/Website_SA/file/**

II: Crawl data:

Nguồn dữ liệu

Kênh thông tin kinh tế - tài chính Việt Nam (cafef.vn): Kênh tin tức online phổ cập

Thu thập và quản lý dữ liệu

Sử dụng thư viện `BeautifulSoup` vì các trang báo có cấu trúc đơn giản và nội dung chủ yếu là dạng chữ. Nhược điểm của phương pháp này là sẽ không xử lý được những bài báo đặc biệt (như infographic hay feature story)

Dữ liệu là toàn bộ tin tức được đăng trên trang web từ năm 2016 tới nay, được quản lý nhờ 2 class:

1. `articles` là các bài báo, sử dụng thư viện `dataclass`, gồm các thông tin như đường dẫn, thời gian đăng, tiêu đề, nội dung… và những method hỗ trợ để đọc, ghi, thu thập nội dung từ đường dẫn, xử lý nội dung…
2. `article_managers` quản lý `list` các bài báo, gồm đọc, ghi, thu thập (từ sitemap chung -> từng submaps -> từng đường dẫn các bài báo), xử lý nội dung, lọc các chủ đề quan tâm (“Doanh nghiệp” và “Thị trường chứng khoán”)…

III: Model:

1. Summarize

- Sử dụng OpenAI API để tóm tắt các bài báo theo từng mã chứng khoán được đề cập trong bài và nội dung chính liên quan đến mã chứng khoán đó.
- Mô hình sử dụng: **gpt-3.5-turbo-instruct**
- Lý do lựa chọn model:
    - Mô hình này được thiết kế để xử lý các tác vụ dựa trên hướng dẫn, cho phép người dùng cung cấp một đoạn văn bản hướng dẫn cụ thể và yêu cầu mô hình thực hiện các tác vụ dựa trên hướng dẫn đó.
    - Tùy chỉnh thông số: mô hình này cho phép tinh chỉnh các thông số của mô hình như temperature, số lượng token tối đa, top-p probability, penalty frequency, penalty presence để điều chỉnh quá trình sinh văn bản và kết quả đầu ra.
    - Hiệu suất cải thiện: GPT-3.5-turbo-instruct được cải thiện so với các phiên bản trước đó, cho phép mô hình xử lý các tác vụ phức tạp và tạo ra kết quả chất lượng hơn.

    - Tiết kiệm thời gian: sử dụng mô hình GPT-3.5-turbo-instruct giúp tiết kiệm thời gian trong việc thực hiện các tác vụ liên quan đến xử lý ngôn ngữ tự nhiên và sinh văn bản.
- Giới hạn của mô hình:
    - Mô hình có giới hạn số lượng token tối đa của dữ liệu đầu vào và kết quả trả ra (tối đa 4097 tokens)

    - Mô hình trả ra có thể là một chuỗi ký tự xuất hiện trong bài báo nhưng không thực sự là một mã chứng khoán trong thực tế

2. Sentiment analyze

Với mục tiêu phân tích tình cảm, chúng em đánh giá mô hình phobert-base-v2 là thích hợp nhất hiện nay.

1. `PhoBERT-base-v2` là mô hình thuộc class `Roberta`, là một dạng mô hình `BERT (Bidirectional Encoder Representations) transformer`, có năng lực thực hiện các thao tác xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) tốt nhất hiện nay. Các mô hình `transformer` có thể biểu diễn được ngữ cảnh của câu tốt hơn các dạng mô hình trước đây nhờ nhờ cơ chế chú ý (attention mechanism - giúp thể hiện mỗi từ còn chịu ảnh hưởng của những từ nào khác trong câu dưới dạng số mà mô hình có thể hiểu và học được). Các mô hình trước đây có thể kể đến như: CNN, RNN, SVM, hay đơn giản như Bag-of-Word.
2. `PhoBERT-base-v2` đã được huấn luyện trước với lượng lớn dữ liệu tiếng Việt (> 140 GB text), nhờ đó tiết kiệm được thời gian và chi phí huấn luyện từ đầu. Dù vậy, `PhoBERT` cần tiếp tục được điều chỉnh (fine-tune) thông qua 02 bước nữa: (1) để thực hiện được công việc phân tích tình cảm (là một thao tác phân loại trong NLP), và (2) để chuyên môn hóa để áp dụng với các dữ liệu liên quan tới tin tức tài chính/doanh nghiệp.

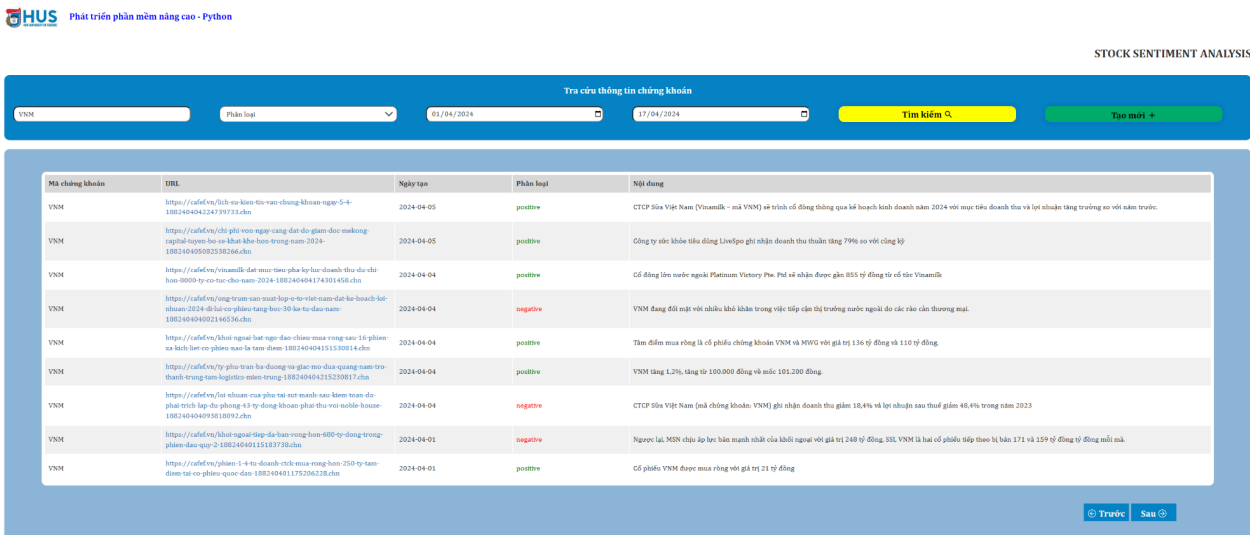
Quá trình huấn luyện `phobert`:

1. Bước 1: luyện phân tích tình cảm qua bộ dữ liệu ~30.000 review thương mại điện tử. Bước này chủ yếu sử dụng các công cụ trong hệ sinh thái của HuggingFace. Mô hình sau huấn luyện thuộc subclass `RobertaForSequenceClassification`, chuyên phân loại ngôn ngữ. Mô hình đạt độ chính xác trên dữ liệu kiểm tra ~80% (thấp hơn các kết quả cao nhất hiện nay ở ~85% nhưng cao hơn mô hình BoW thông thường ở ~70%)
2. Bước 2: chuyên môn hóa với dữ liệu ~700 dòng tóm tắt tin tức về tài chính/doanh nghiệp. Mô hình đạt độ chính xác trên dữ liệu kiểm tra ~85%.

Sau khi huấn luyện, mô hình có thể gán nhãn dữ liệu với class `pipeline` trong thư viện `transformer`

IV: Website

1. Search function



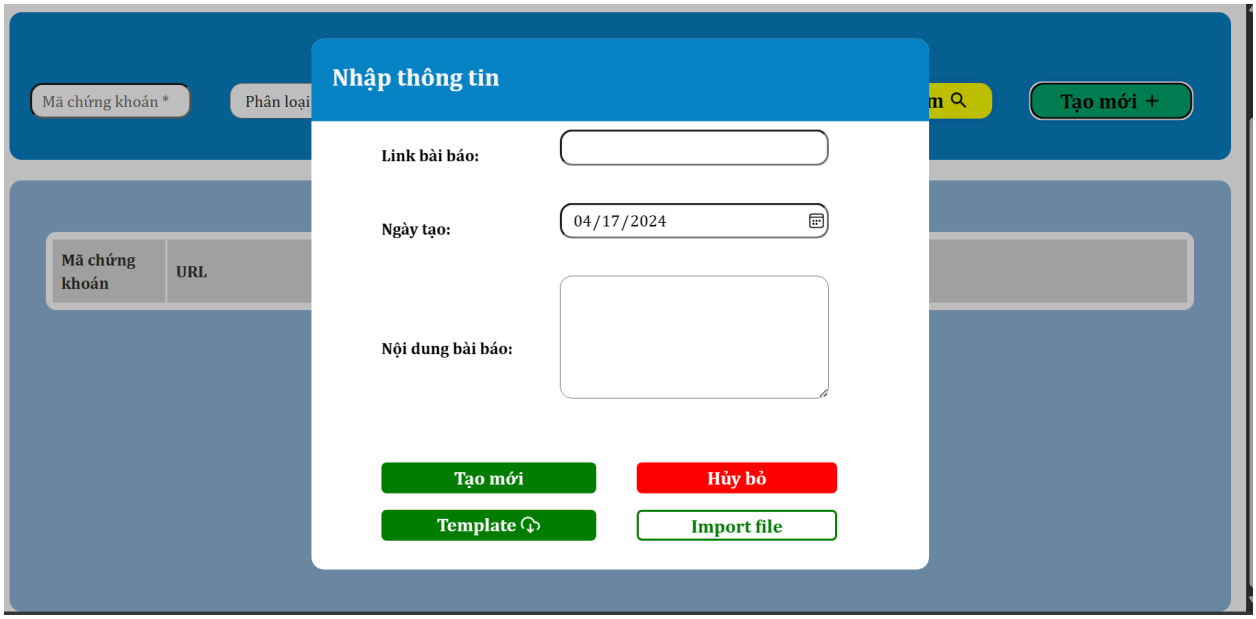
Mô tả chức năng tìm kiếm dữ liệu:

- Website cho phép tìm kiếm các bài báo liên quan đến một mã chứng khoán truyền vào
- Kết quả tìm kiếm trả ra bao gồm các thông tin:
    - Mã chứng khoán: Mã chứng khoán
    - URL: Link bài báo liên quan
    - Ngày tạo: Thời gian tạo của bài báo
    - Phân loại: Đánh giá nội dung bài báo: Positive/ Negative/ Neutral
    - Nội dung: Nội dung tóm tắt liên quan đến mã chứng khoán đang tìm kiếm

Cách sử dụng:

- Nhập các thông tin cần tra cứu, bao gồm:
    - Mã chứng khoán: thông tin bắt buộc
    - Phân loại: có 4 lựa chọn: Positive/ Negative/ Neutral hoặc Blank (tra cứu tất cả các phân loại)
    - Thời gian: Ngày tạo từ/ Ngày tạo đến
- Click vào button **Tìm kiếm**

2. Insert function



Mô tả chức năng thêm dữ liệu:

- Thêm 1 hoặc nhiều bài báo mới
- Thông tin bài báo bao gồm: Link bài báo, ngày tạo của bài báo, nội dung bài báo

Cách sử dụng:

2 cách:

- Cách 1: Thêm từng bài báo:
  - Click button “Tạo mới” trên màn hình chính
  - Điền thông tin link bài báo, ngày tạo của bài báo, nội dung bài báo vào ô tương ứng
  - Click vào tạo mới
- Cách 2: Thêm nhiều bài báo:
  - Click button “Tạo mới” trên màn hình chính
  - Tải “template” mẫu để điền theo yêu cầu (bao gồm thông tin link bài báo, ngày tạo của bài báo, nội dung bài báo vào ô tương)
  - Import file vừa điền theo template

V. Mục tiêu mở rộng

Những mục tiêu sau này có thể cải thiện:

1. Thêm nhiều nguồn thông tin hơn: các trang báo khác, các trang mạng xã hội, forum…
2. Tiếp tục huấn luyện chuyên môn hóa mô hình phân tích tình cảm với dữ liệu về tin tức tài chính/doanh nghiệp