FYP 21029

Food-Related Article Aggregation and Analytical System

Project Plan

**Supervisor:**
Dr. Tam, Anthony T.C.

**Industry Partner:**
Luk Advisor Limited

**Group Members:**
Chau Kin On (3035573848)
Cheung Man Chung (3035572026)
Tam Chin Hang (3035564457)
1. Project Background

Our industry partner, Luk Advisor Limited, is maintaining a price comparison site, Takeawayla.com, which focuses on takeaway food on different platforms in Hong Kong, such as Deliveroo and Foodpanda.

1.1. Current Situation

Currently, the company staff are finding and sourcing food-related articles by browsing on different websites and forums. The collected data is organized manually for identifying the latest trending food and restaurants. Then, the organized information will be used for writing new articles on a weekly basis and post them on Takeawayla.com for providing updated food suggestions to the visitors.

Fig 1. thumbnails of food articles from Takeawayla.com [link: https://takeawayla.com/latestnews]
1.2. Problem Identification

1.2.1. Collecting food articles manually
Numerous food articles are posted on different online platforms, such as forums and social media, everyday. In order to have most-updated information, regular article collection is required. However, reading a large amount of food articles can be quite time-consuming and apparently error-prone. If the content of the information is incomplete due to mistakes made during the data collection process, it may incur confusion and even misunderstanding, which will affect the company's image. Besides, when multiple staff members are at the collection stage at the same time, repeated articles may be read and collected, which takes unnecessary time and effort.

1.2.2. No statistical summary of the datasets
The collected articles are organized by placing the files in separate folders and renaming the filenames. In such a way, it is hard to extract simple metrics of the dataset manually, like the total number of articles collected and the word count of individual articles. Moreover, it is even more difficult to extract some customized metrics such as filtering articles by specifying food keywords and sorting the articles by the hit-rate.

1.2.3. Lack of food article analysis
To understand the recent trend of the foods and restaurants, the staff need to conduct certain analysis on the dataset. However, without proper analytical tools, it can be challenging for the staff to obtain an overview of the data, and some valuable information might be missed. It is also difficult for the staff to perform critical inferences on the analytical results.

1.2.4. Inconsistency among different authors
Every staff member has their own writing styles, habits and use of vocabulary, and they also have different perceptions while reading the same material. The inconsistency may be presented in the food articles written by different authors.
1.3. Our Solution

An aggregation and analytical system will be implemented in this project. The system will automatically listen to different forums and sites for food-related articles, and analyze whether these articles are popular or not. Similar to the belief of the Industry, we also appreciate the power and value of “Triple-A”, namely AI, Analytics and Automation.

The data aggregation part of the system can improve the efficiency in data collection, compared to manually browsing and sourcing data on the Internet.

The data analysis part of the system helps organize the data systematically and uses the NLP techniques to transform the data into machine-readable materials for providing statistical summaries and analyzing the trend on the aggregated article dataset.

A full-stack web development application will be implemented that the backend framework will configure with the analytical results, while the frontend is responsible for the data presentation and visualization.

If feasible, an article generator based on the NLP model will be made to create new food-related articles based on the aggregated article dataset, in order to provide information to visitors in a consistent writing style.
2. Objectives and Scope

2.1. Objectives

The main objective of this project is to create an aggregation and analytical system that is available to monitor and keep track of various food-related articles from distinct forums, food-platforms, websites, etc. and is capable of conducting analysis on the popularity of the articles. The system should also be competent to visualize the analytical results. An automated data aggregation model with a systematic database, a NLP model for data analysis and a full-stack web application will be included in the system.

2.2. Scope

The initial scope is set to be collecting data from food-related public sites, which have no membership restriction or subscription fee, and allow robot scraping. Here are some examples:

- FoodPanda: [https://www.foodpanda.hk/](https://www.foodpanda.hk/)
- Deliveroo: [https://deliveroo.hk/zh/](https://deliveroo.hk/zh/)
- OpenRice: [https://www.openrice.com/zh/hongkong](https://www.openrice.com/zh/hongkong)
- Uber Eats: [https://www.ubereats.com/hk](https://www.ubereats.com/hk)

This project mainly focuses on food articles in the Chinese language, especially Cantonese, as the target audience of the articles is Hong Kong citizens and the food introduced should be available in Hong Kong.
3. Project Methodology

In this section, the methodology for achieving the project’s objectives, which are mentioned previously, will be introduced.

The core technology used in this project is Natural Language Processing (NLP). The system relies on NLP techniques for data preprocessing, data cleaning and data processing. NLP is also crucial for the data analysis part.

According to our industry partner, in order to follow their development plan, Python and Javascript will be the major languages to be used in this project. Besides, the Scrum practice in software engineering will be followed, with the help of the project management web-application, Trello. To enhance team members’ collaboration, the version control tool, Github, will be used in this project. The system will be tested and hosted on cloud hosting service, AWS ECS.

The whole project will be separated into several sprints. To generalize, the project is divided into 3 major sections:

3.1. Data Aggregation

For web-scraping tasks, Python packages Scrapy, BeautifulSoup and python-based Selenium will be used. Splash (Lua-based) will also be applied for scraping dynamic websites. PostgreSQL will be used as the database management system for storing and querying collected article datasets. The database will be accessed in the Data Processing part.

3.2. Data Processing

For the data pre-processing part, The Natural Language Toolkit (NLTK) package [2] and Python standard library will be used for primary text data cleaning and filtering functionalities. PyCantonese [3] and NLP Chinese Corpus [4] libraries will be used for obtaining common Cantonese and Mandarin corpus data. These tools will be used for advanced data cleaning on collected article texts. For the NLP model part, Cantoformer [5], Chinese BERT WWM [6] and Google BERT-Base [7] will be our pre-training BERT model candidates. The best BERT model among these candidates will be chosen as the primary BERT model to complete further text analysis including sentimental analysis, topic classification and summary generation. If possible, our chosen BERT model will be further used as an article generator to generate new human-like AI articles by using an aggregated article dataset.
3.3. Data Presentation & Visualization

*React* will be used as the Frontend framework for constructing the user interface and presenting the visualized analysis. It will also perform data fetching of the requested results and handle the corresponding user logic of the user actions. For the Backend, *Django* and *Django rest-framework* will be used to configure the analytical result, handling routing logic and communicating with the *React* Frontend.

The *Material-UI* [9] library will be applied for the UI styling while *D3.js* [10] and other graphic visualization libraries will be utilized for the data presentation and visualization. Word Cloud, Topic Classification and Article Summary will be used for demonstrating the Text-related analysis summary.

The process of data presentation and visualization is a significant part of the project, it will demonstrate our achievements in solving the project problems and show our effort on providing the solution to the industry partner. It is also imperative for meeting our project deliverable.

3.4. Division of Labour

<table>
<thead>
<tr>
<th>Section</th>
<th>Main member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project management</td>
<td>Tam</td>
</tr>
<tr>
<td>Project configuration</td>
<td>Cheung, Chau</td>
</tr>
<tr>
<td><strong>Front-end</strong></td>
<td></td>
</tr>
<tr>
<td>UI design &amp; UX (User Logic)</td>
<td>Chau, Tam</td>
</tr>
<tr>
<td>Data fetching</td>
<td>Chau</td>
</tr>
<tr>
<td>Data presentation and visualization</td>
<td>Chueng, Tam, Chau</td>
</tr>
<tr>
<td><strong>Back-end</strong></td>
<td></td>
</tr>
<tr>
<td>Database</td>
<td>Tam</td>
</tr>
<tr>
<td>Data crawling</td>
<td>Chau, Cheung</td>
</tr>
<tr>
<td>Data processing</td>
<td>Chau, Tam, Cheung</td>
</tr>
</tbody>
</table>
## 4. Project Schedule

<table>
<thead>
<tr>
<th>Date/Period</th>
<th>Tasks and Deliverables</th>
<th>Milestones</th>
</tr>
</thead>
</table>
| 3 Oct 2021        | **Deliverable of Phase 1 (inception)**  
  ● Project Plan *(this document)*  
  ● Project Webpage |            |
| Oct 2021          | Experimental research:  
  Research/Study on required tools & models  
  **Completing project configuration** | M1         |
| 30 Nov 2021       | **Completing automatic data crawling backend functionality** | M2         |
| 31 Dec 2021       | **Completing data cleaning and processing functionality (phase 1)***                  | M3         |
| 1-9 Jan 2022      | Preparation of First Presentation                                                      |            |
| 10-14 Jan 2022    | **First Presentation**                                                                  |            |
| 23 Jan 2022       | **Deliverable of Phase 2 (elaboration)**  
  ● Interim Report |            |
| 20 Feb 2022       | **Completing data processing functionality (phase 2)***                               | M4         |
|                   | **Completing data fetching, UI & UX (Experimental) on Frontend**                      | M5         |
| 31 Mar 2022       | **Completing data presentation and visualization functionality**                     | M6         |
| 18 Apr 2022       | **Deliverable of Phase 3 (construction)**  
  ● Preparation of Final Presentation  
  ● Final Report |            |
| 19-22 Apr 2022    | **Final Presentation**                                                                 |            |
| 4 May 2022        | **Project Exhibition**  
  ● 3-mins video  
  ● Poster |            |
| 31 May 2022       | **Project Completion**                                                                 |            |

*Data processing is splitted into 2 phases for implementation due to its complexities*
5. References


