NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

CZ4034: Information Retrieval

Group 24 Assignment Report

Name
GUCON NAILAH GINYLLE PABILONIA
KAM CHIN VOON
LIANG ZHIHENG
LIM YUN HAN, DARREN
WANG ANYI

Table of Content

Table of Content 2
Introduction to TripAdvisor 3
Crawling 4
    Question 1 4
Indexing 12
    Question 2 12
    Question 3 20
Classification 24
    Question 4 24
    Question 5 35
Links 39
    Video Presentation YouTube Link 39
    Data Files Link 39
    Source Code Link 39

Introduction to TripAdvisor

TripAdvisor is a website that provides reviews and ratings of hotels, eateries, and tourist attractions from travellers all over the world. It is a popular platform for people to share their experiences and provide recommendations to others who are planning their trips.

TripAdvisor is useful because it allows users to access a large database of reviews and ratings, which can help them make informed decisions about where to stay, where to eat, and what to do during their travels. In 2021, TripAdvisor conducted a comprehensive survey* across 5 countries, including Singapore, to gain insights into the impact of reviews on consumers. The survey involved 9000 participants, with 3 out of 4 respondents acknowledging the extreme or very high importance of online reviews when making travel decisions. Interestingly, this held true across different types of reviews, with accommodations scoring an 82%, while eateries scored a 70%. These findings highlight the significant role of online reviews in influencing consumer decisions.

By creating a Singapore-specific search engine with hotel and eatery data crawled from TripAdvisor's website and utilising classification models such as sentiment analysis, users can benefit from it by easily finding the information they need to make informed decisions about their trip to Singapore. By incorporating sentiment analysis, the search engine can provide valuable insights from reviews and ratings, such as overall satisfaction levels, to help users make more informed choices about where to stay and dine during their stay in Singapore.

* Statistics from: https://www.tripadvisor.com/powerofreviews

Crawling

Question 1

<u>How you crawled the corpus (e.g., source, keywords, API, library) and stored it</u>

Overview:

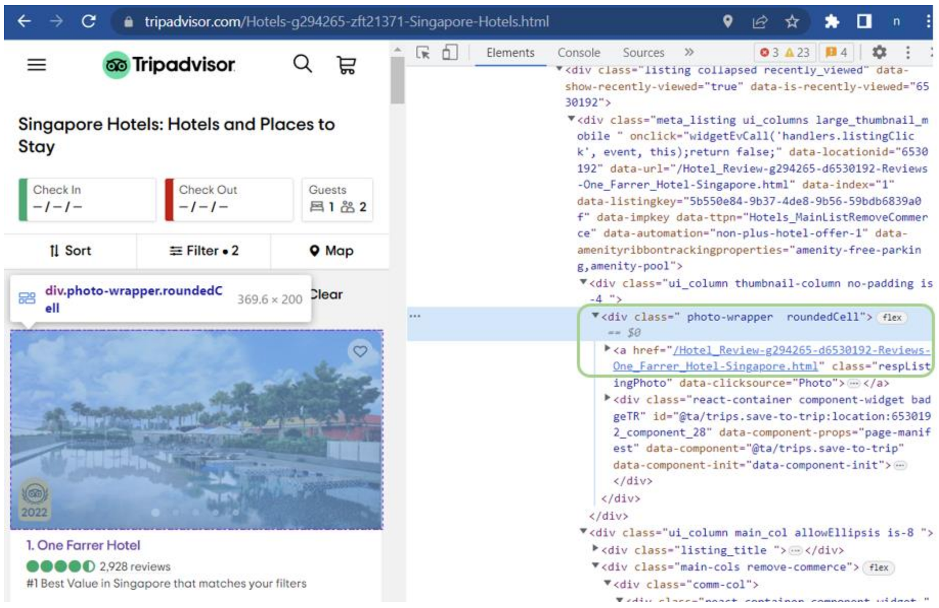
The corpus of Singapore hotels and eateries was obtained by crawling TripAdvisor's website. The web crawling process was automated using **Python Selenium**, which facilitated navigating the website programmatically and extracting data from it without the need for manual interaction. To obtain a comprehensive dataset, two levels of web crawling were carried out.

First-Level Crawling:

The first level involved the extraction of a list of URLs to Singapore hotels and eateries from the specific website's main pages (refer to Table 1). This was accomplished by utilising the inspect functionality to locate the div class containing the "a href" tag, as shown in Figure 1. The crawled links for hotels and eateries were then outputted inside their respective .csv files (refer to Figure 2).

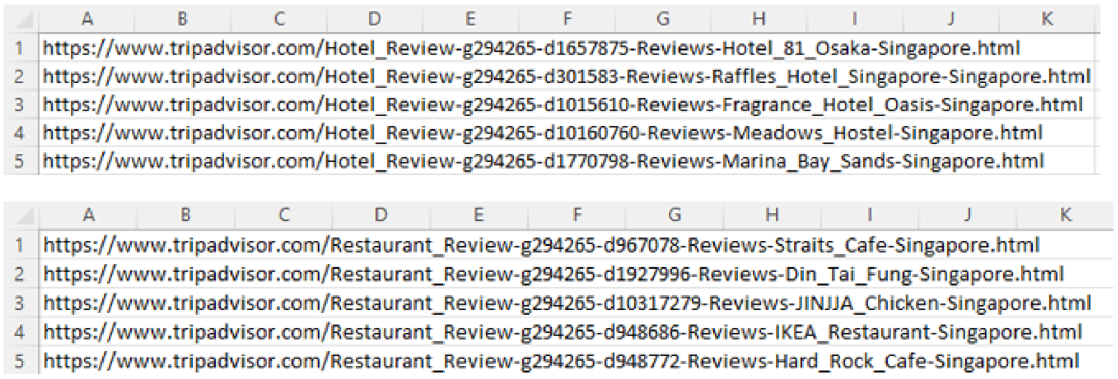
Table 1: Links used for web crawling task

Links used (TripAdvisor's Singapore Hotels and Eateries Pages)	
Hotels	https://www.tripadvisor.com/Hotels-g294265-zft21371-Singapore-Hotels.html
Eateries	https://www.tripadvisor.com/Restaurants-g294265-Singapore.html


tripadvisor.com/Hotels-g294265-zft21371-Singapore-Hotels.html
Tripadvisor
Singapore Hotels: Hotels and Places to Stay
Check In
-/-/-
Check Out
-/-/-
Guests
1 2
Sort
Filter • 2
Map
div.photo-wrapper.roundedCell
369.6 × 200
Clear
2022
1. One Farrer Hotel
2,928 reviews
#1 Best Value in Singapore that matches your filters
Elements
Console
Sources
3
23
4

```
<div class="listing collapsed recently_viewed" data-show-recently-viewed="true" data-is-recently-viewed="6530192">
  <div class="meta_listing ui_columns large_thumbnail_mobile " onclick="widgetEvCall('handlers.listingClick', event, this);return false;" data-locationid="6530192" data-url="/Hotel_Review-g294265-d6530192-Reviews-One_Farrer_Hotel-Singapore.html" data-index="1" data-listingkey="5b550e84-9b37-4de8-9b56-59bdb6839a0f" data-impkey data-ttpn="Hotels_MainListRemoveCommerce" data-automation="non-plus-hotel-offer-1" data-amenityribbontrackingproperties="amenity-free-parking,amenity-pool">
    <div class="ui_column thumbnail-column no-padding is-4 ">
      <div class=" photo-wrapper  roundedCell"> flex == $0
        <a href="/Hotel_Review-g294265-d6530192-Reviews-One_Farrer_Hotel-Singapore.html" class="respListingPhoto" data-clicksource="Photo">…</a>
        <div class="react-container component-widget badgeTR" id="@ta/trips.save-to-trip:location:6530192_component_28" data-component-props="page-manifest" data-component="@ta/trips.save-to-trip" data-component-init="data-component-init">…
        </div>
      </div>
    </div>
    <div class="ui_column main_col allowEllipsis is-8 ">
      <div class="listing_title ">…</div>
      <div class="main-cols remove-commerce"> flex
        <div class="comm-col">
```

Figure 1: Sample screenshot of usage of inspect functionality



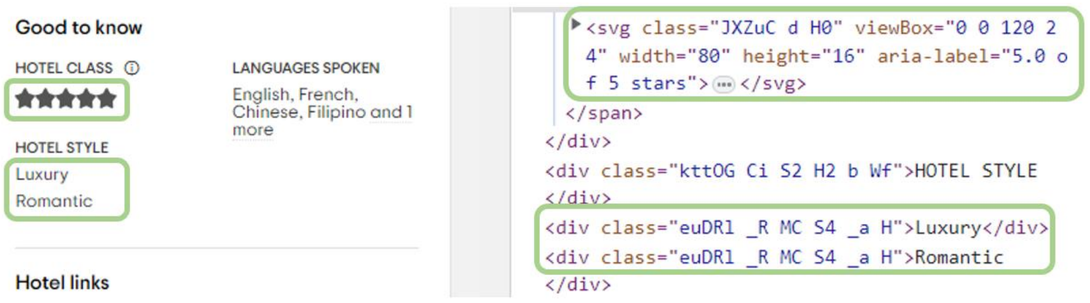
	A
1	https://www.tripadvisor.com/Hotel_Review-g294265-d1657875-Reviews-Hotel_81_Osaka-Singapore.html
2	https://www.tripadvisor.com/Hotel_Review-g294265-d301583-Reviews-Raffles_Hotel_Singapore-Singapore.html
3	https://www.tripadvisor.com/Hotel_Review-g294265-d1015610-Reviews-Fragrance_Hotel_Oasis-Singapore.html
4	https://www.tripadvisor.com/Hotel_Review-g294265-d10160760-Reviews-Meadows_Hostel-Singapore.html
5	https://www.tripadvisor.com/Hotel_Review-g294265-d1770798-Reviews-Marina_Bay_Sands-Singapore.html

	A
1	https://www.tripadvisor.com/Restaurant_Review-g294265-d967078-Reviews-Straits_Cafe-Singapore.html
2	https://www.tripadvisor.com/Restaurant_Review-g294265-d1927996-Reviews-Din_Tai_Fung-Singapore.html
3	https://www.tripadvisor.com/Restaurant_Review-g294265-d10317279-Reviews-JINJJA_Chicken-Singapore.html
4	https://www.tripadvisor.com/Restaurant_Review-g294265-d948686-Reviews-IKEA_Restaurant-Singapore.html
5	https://www.tripadvisor.com/Restaurant_Review-g294265-d948772-Reviews-Hard_Rock_Cafe-Singapore.html

Figure 2: Sample output of first level crawling

Second-Level Crawling:

The second level will involve crawling each individual hotel and eatery URL in the .csv file obtained during first-level crawling to extract targeted information and a number of reviews about each establishment. Samples of HTML elements used to identify the targeted information for both hotels and eateries can be seen in Figures 3 to 6, while a sample of the Python Selenium code used to crawl the targeted information, such as the reviews, using the identified HTML elements for hotels and eateries is shown in Figures 7 and 8 respectively.



Good to know

HOTEL CLASS ★★★★★

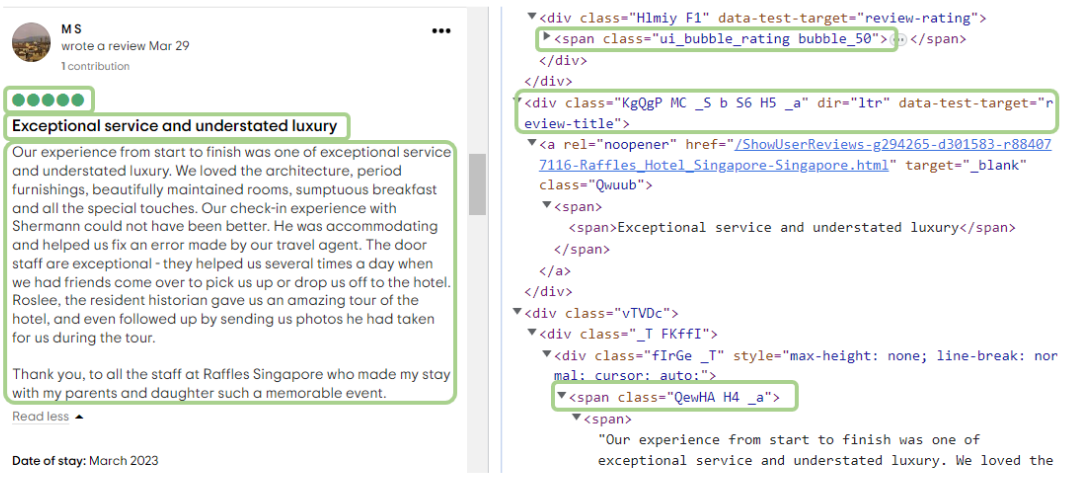
HOTEL STYLE
Luxury
Romantic

LANGUAGES SPOKEN
English, French, Chinese, Filipino and 1 more

Hotel links

```
    <svg class="JXZuC d H0" viewBox="0 0 120 2
    4" width="80" height="16" aria-label="5.0 o
    f 5 stars">⋯</svg>
  </span>
</div>
<div class="kttOG Ci S2 H2 b Wf">HOTEL STYLE
</div>
<div class="euDR1 _R MC S4 _a H">Luxury</div>
<div class="euDR1 _R MC S4 _a H">Romantic
</div>
```

Figure 3: Sample HTML Elements used to extract hotel information



M S
wrote a review Mar 29
1 contribution

Exceptional service and understated luxury

Our experience from start to finish was one of exceptional service and understated luxury. We loved the architecture, period furnishings, beautifully maintained rooms, sumptuous breakfast and all the special touches. Our check-in experience with Shermann could not have been better. He was accommodating and helped us fix an error made by our travel agent. The door staff are exceptional - they helped us several times a day when we had friends come over to pick us up or drop us off to the hotel. Roslee, the resident historian gave us an amazing tour of the hotel, and even followed up by sending us photos he had taken for us during the tour.

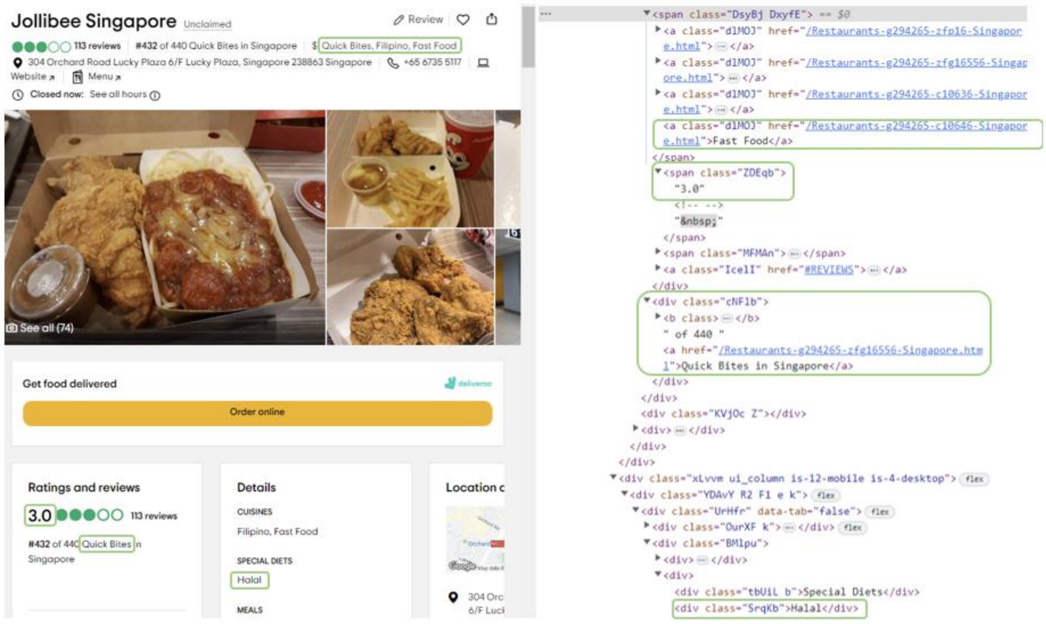
Thank you, to all the staff at Raffles Singapore who made my stay with my parents and daughter such a memorable event.

Read less

Date of stay: March 2023

```
▼<div class="Hlmiy F1" data-test-target="review-rating">
   ▶<span class="ui_bubble_rating bubble_50">⋯</span>
  </div>
 </div>
▼<div class="KgQgP MC _S b S6 H5 _a" dir="ltr" data-test-target="r
 eview-title">
  ▼<a rel="noopener" href="/ShowUserReviews-g294265-d301583-r88407
   7116-Raffles_Hotel_Singapore-Singapore.html" target="_blank"
   class="Qwuub">
    ▼<span>
       <span>Exceptional service and understated luxury</span>
     </span>
   </a>
 </div>
▼<div class="vTVDc">
  ▼<div class="_T FKffI">
    ▼<div class="fIrGe _T" style="max-height: none; line-break: nor
     mal; cursor: auto;">
      ▼<span class="QewHA H4 _a">
        ▼<span>
           "Our experience from start to finish was one of
           exceptional service and understated luxury. We loved the
```

Figure 4: Sample HTML Elements used to extract hotel reviews



Jollibee Singapore Unclaimed
Review
113 reviews | #432 of 440 Quick Bites in Singapore | $ Quick Bites, Filipino, Fast Food
304 Orchard Road Lucky Plaza 6/F Lucky Plaza, Singapore 238863 Singapore | +65 6735 5117 |
Website | Menu
Closed now: See all hours
See all (74)

Get food delivered
Order online

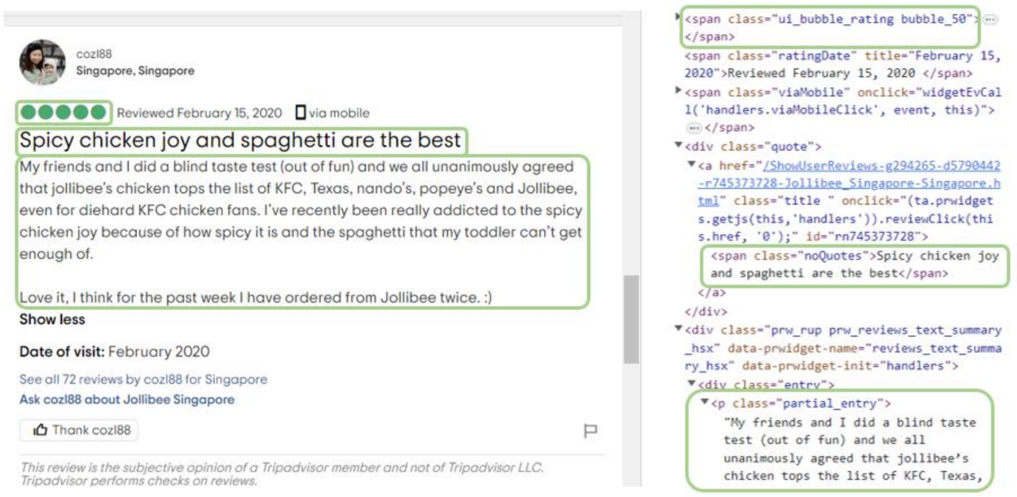
Ratings and reviews
3.0 113 reviews
#432 of 440 Quick Bites in Singapore

Details
CUISINES
Filipino, Fast Food
SPECIAL DIETS
Halal
MEALS

Location c
304 Orc
6/F Luck

```
▼<span class="DsyBj DxyfE"> == $0
  ▶<a class="dlMOJ" href="/Restaurants-g294265-zfp16-Singapor
   e.html">⋯</a>
  ▶<a class="dlMOJ" href="/Restaurants-g294265-zfg16556-Singap
   ore.html">⋯</a>
  ▶<a class="dlMOJ" href="/Restaurants-g294265-c10636-Singapor
   e.html">⋯</a>
   <a class="dlMOJ" href="/Restaurants-g294265-c10646-Singapor
   e.html">Fast Food</a>
 </span>
▼<span class="ZDEqb">
   "3.0"
   <!-- -->
   "&nbsp;"
  </span>
 ▶<span class="MFMAn">⋯</span>
 ▶<a class="IcelI" href="#REVIEWS">⋯</a>
 </div>
▼<div class="cNFlb">
  ▶<b class>⋯</b>
   " of 440 "
   <a href="/Restaurants-g294265-zfg16556-Singapore.htm
   l">Quick Bites in Singapore</a>
 </div>
</div>
<div class="KVjOc Z"></div>
▶<div>⋯</div>
</div>
</div>
▼<div class="xLvvm ui_column is-12-mobile is-4-desktop"> flex
 ▼<div class="YDAvY R2 F1 e k"> flex
  ▼<div class="UrHfr" data-tab="false"> flex
   ▶<div class="OurXF k">⋯</div> flex
   ▼<div class="BMlpu">
     ▶<div>⋯</div>
     ▼<div>
        <div class="tbUiL b">Special Diets</div>
        <div class="SrqKb">Halal</div>
```

Figure 5: Sample HTML Elements used to extract eatery information



cozl88
Singapore, Singapore

Reviewed February 15, 2020 via mobile

Spicy chicken joy and spaghetti are the best

My friends and I did a blind taste test (out of fun) and we all unanimously agreed that jollibee's chicken tops the list of KFC, Texas, nando's, popeye's and Jollibee, even for diehard KFC chicken fans. I've recently been really addicted to the spicy chicken joy because of how spicy it is and the spaghetti that my toddler can't get enough of.

Love it, I think for the past week I have ordered from Jollibee twice. :)

Show less

Date of visit: February 2020

See all 72 reviews by cozl88 for Singapore
Ask cozl88 about Jollibee Singapore

Thank cozl88

This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews.

```
▶<span class="ui_bubble_rating bubble_50">⋯
 </span>
 <span class="ratingDate" title="February 15,
 2020">Reviewed February 15, 2020 </span>
▶<span class="viaMobile" onclick="widgetEvCal
 l('handlers.viaMobileClick', event, this)">
 ⋯</span>
▼<div class="quote">
  ▼<a href="/ShowUserReviews-g294265-d5790442
   -r745373728-Jollibee_Singapore-Singapore.h
   tml" class="title " onclick="(ta.prwidget
   s.getjs(this,'handlers')).reviewClick(thi
   s.href, '0');" id="rn745373728">
     <span class="noQuotes">Spicy chicken joy
     and spaghetti are the best</span>
   </a>
 </div>
▼<div class="prw_rup prw_reviews_text_summary
 _hsx" data-prwidget-name="reviews_text_summa
 ry_hsx" data-prwidget-init="handlers">
  ▼<div class="entry">
    ▼<p class="partial_entry">
       "My friends and I did a blind taste
       test (out of fun) and we all
       unanimously agreed that jollibee's
       chicken tops the list of KFC, Texas,
```

Figure 6: Sample HTML Elements used to extract eatery reviews

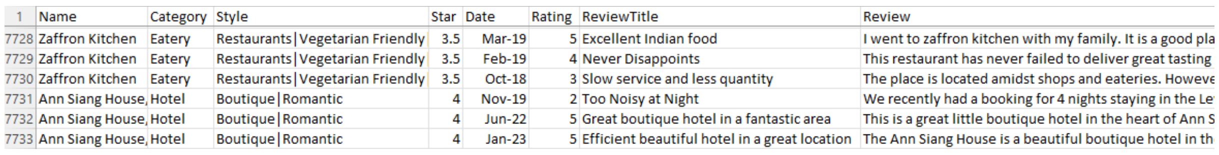
```
...
rating = int(container[j].find_element(By.XPATH, ".//span[contains(@class, 'ui_bubble_rating
bubble_')]").get_attribute("class").split("_")[3])/10
title = container[j].find_element(By.XPATH, ".//div[contains(@data-test-target, 'review-title')]").text
review = container[j].find_element(By.XPATH, ".//span[@class='QewHA H4 _a']").text.replace("\n", "  ")
...
```

Figure 7: Sample Python Selenium crawling code for hotels

```
...
rating = int(container[j].find_element(By.XPATH, ".//span[contains(@class, 'ui_bubble_rating
bubble_')]").get_attribute("class").split("_")[3])/10
title = container[j].find_element(By.XPATH, ".//span[@class='noQuotes']").text
review = container[j].find_element(By.XPATH, ".//p[@class='partial_entry']").text.replace("\n", " ")
...
```

Figure 8: Sample Python Selenium crawling code for eateries

The crawler extracted data from 210 hotels and 210 eateries, each having between 20 to 50 reviews. Both the extracted hotel and eatery information are outputted into a combined .csv file.



1	Name	Category	Style	Star	Date	Rating	ReviewTitle	Review
7728	Zaffron Kitchen	Eatery	Restaurants|Vegetarian Friendly	3.5	Mar-19	5	Excellent Indian food	I went to zaffron kitchen with my family. It is a good pla
7729	Zaffron Kitchen	Eatery	Restaurants|Vegetarian Friendly	3.5	Feb-19	4	Never Disappoints	This restaurant has never failed to deliver great tasting
7730	Zaffron Kitchen	Eatery	Restaurants|Vegetarian Friendly	3.5	Oct-18	3	Slow service and less quantity	The place is located amidst shops and eateries. Howeve
7731	Ann Siang House,	Hotel	Boutique|Romantic	4	Nov-19	2	Too Noisy at Night	We recently had a booking for 4 nights staying in the Le
7732	Ann Siang House,	Hotel	Boutique|Romantic	4	Jun-22	5	Great boutique hotel in a fantastic area	This is a great little boutique hotel in the heart of Ann S
7733	Ann Siang House,	Hotel	Boutique|Romantic	4	Jan-23	5	Efficient beautiful hotel in a great location	The Ann Siang House is a beautiful boutique hotel in th

Figure 9: Combined .csv file

What each column in the .csv file represents are as follows:

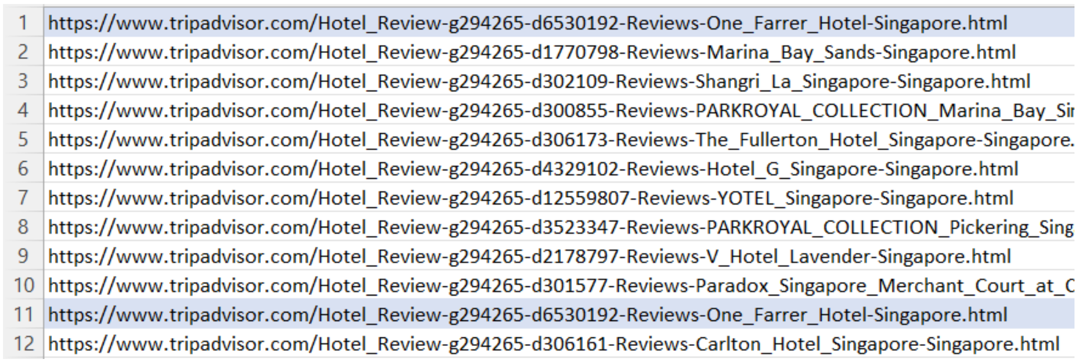
- Name: Name of Hotel or Eatery
- Category: Either “Hotel” or “Eatery”
- Style:
    - For Hotels: the style of hotel e.g. [‘Boutique’, ‘Romantic’]
    - For Eateries: a mix of type and special diets [‘Quick Bites’, ‘Halal’]
- Star:
    - For Hotels: Star Rating of Hotel
    - For Eateries: Average Star Rating of Reviews
- Date:
    - For Hotels: Date of Stay
    - For Eateries: Date of Visit
- Rating: Review rating of Hotel or Eatery
- ReviewTitle: Review title of Hotel or Eatery
- Review: Review text of Hotel or Eatery

This two-level crawling process will ensure that the dataset is complete and includes all the relevant information needed for indexing and classification.

For the full source code of all Python Selenium code and all HTML elements used to extract the targeted information, as well as the Python code used for the additional steps taken (preprocessing of crawled corpus) shown below, please view the “crawling” folder in the Source Code link found in the “Links” section of the report.

Issues Identified:

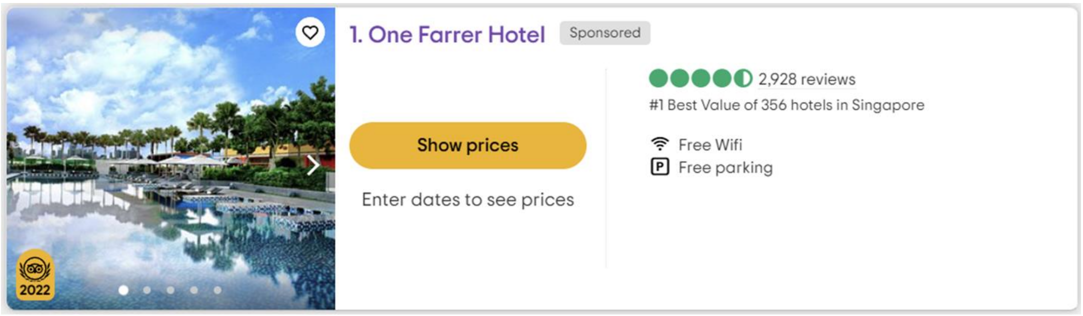
Through the process of crawling, repeated instances of hotel and/or eatery links were discovered (refer to Figure 10).



```
https://www.tripadvisor.com/Hotel_Review-g294265-d6530192-Reviews-One_Farrer_Hotel-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d1770798-Reviews-Marina_Bay_Sands-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d302109-Reviews-Shangri_La_Singapore-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d300855-Reviews-PARKROYAL_COLLECTION_Marina_Bay_Sir
https://www.tripadvisor.com/Hotel_Review-g294265-d306173-Reviews-The_Fullerton_Hotel_Singapore-Singapore.
https://www.tripadvisor.com/Hotel_Review-g294265-d4329102-Reviews-Hotel_G_Singapore-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d12559807-Reviews-YOTEL_Singapore-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d3523347-Reviews-PARKROYAL_COLLECTION_Pickering_Sing
https://www.tripadvisor.com/Hotel_Review-g294265-d2178797-Reviews-V_Hotel_Lavender-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d301577-Reviews-Paradox_Singapore_Merchant_Court_at_C
https://www.tripadvisor.com/Hotel_Review-g294265-d6530192-Reviews-One_Farrer_Hotel-Singapore.html
https://www.tripadvisor.com/Hotel_Review-g294265-d306161-Reviews-Carlton_Hotel_Singapore-Singapore.html
```

Figure 10: Sample of repeats of hotel links in .csv file

This is likely due to the occasional presence of "Sponsored" entries in the webpage, which sometimes appear as the first result after each pagination (refer to Figure 11).



1. One Farrer Hotel Sponsored

2,928 reviews

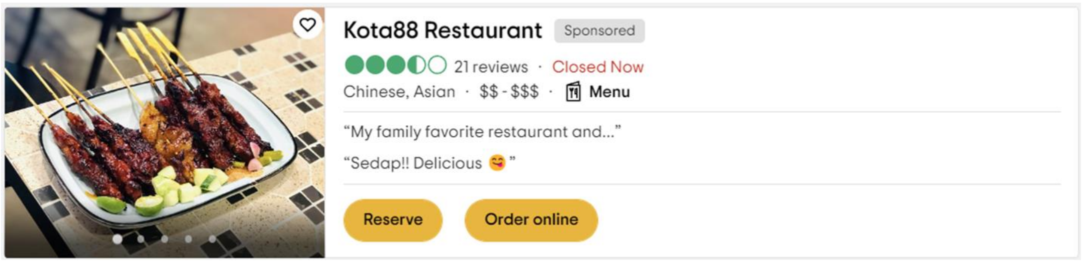
#1 Best Value of 356 hotels in Singapore

Show prices

Enter dates to see prices

Free Wifi

Free parking



Kota88 Restaurant Sponsored

21 reviews · Closed Now

Chinese, Asian · $$ - $$$ · Menu

"My family favorite restaurant and..."

"Sedap!! Delicious 😋 "

Reserve Order online

Figure 11: "Sponsored" hotel and eatery

Since the "Sponsored" entries have the same HTML tag as the other targeted links, checking and removal of duplicate links is done in the level two crawling code before the crawling of detailed information about the establishment begins.

```
...
df_eateries_withDups.drop_duplicates(subset=df.columns[0], inplace=True)
```

Figure 12: Main code statement that removes duplicate links from .csv file

Additional Steps Taken:
To ensure optimal performance of the indexed search engine and accuracy of classification analysis, a series of preprocessing steps were carried out to clean the crawled corpus of hotel and eatery data. The raw text data was found to contain noise that can impair the readability of reviews and hinder the effectiveness of any subsequent analysis.

For example, the effects of applying preprocessing steps such as conversion to lowercase, decoding of HTML entities, removal of non-ASCII characters, punctuation, and stopwords can be seen in Table 2.

Table 2: Links used for web crawling task

Before	After
DidnÃ¢â,¬â„¢t visit the restaurant for awhile. The standard in pork collar was hard. For the price I wasnÃ¢â,¬â„¢t expecting	didnt visit restaurant awhile standard t pork collar hard price wasnt expecting

A custom set of stopwords (refer to Figure 13) was used instead of the default set of stopwords in the nltk library as it removed words such as “not”, “wasn’t”, etc. which could be important for sentiment analysis as they can significantly change the polarity of a sentence (e.g. “the food was not good”, nltk would have removed the “not”, and the output would have become “food good”).

```
stop_words = set(['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves',
'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers',
'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs',
'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being',
'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the',
'and', 'if', 'or', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with',
'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after',
'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over',
'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where',
'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other',
'some', 'such', 'nor', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's',
'will', 'just', 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've',
'y', 'ma'])
```
Figure 13: Custom set of stopwords

This can result in more accurate insights and more effective classification models.

What kind of information users might like to retrieve from your crawled corpus (i.e., applications), with sample queries

The main objective of our project is to assist users in finding relevant information about a hotel or eatery such as its type and reviews.

Therefore, some recommended searches using keywords are:

- Name of hotel or eatery (e.g. “Concorde Hotel”, “JINJJA Chicken”)
- Type of hotel (e.g. “Budget”, “Romantic”)
- Type of eatery (e.g. “Halal”, “Quick Bites”)
- Hotel star rating (e.g. “Hotel 5 stars”)
- Average star rating of eatery (e.g. “Eatery 4 stars”)

The numbers of records, words, and types (i.e., unique words) in the corpus

Table 3: Corpus Information for Question 1 Part 3

	Before preprocessing	**After preprocessing**
Number of records	17669	17669
Number of words	1281072	715643
Number of types (i.e. unique words)	62475	26760

Given the data in Table 3, the crawled corpus fulfils the requirement of having at least 10,000 records and at least 100,000 words.

Indexing

Question 2

<u>Building a simple web interface for the search engine</u>

Overview:

Flask was used to create the web interface for the search engine in tandem with HTML and CSS for the frontend look of the interface, as well as Javascript for interactivity of the webpage.

To reduce clutter of information, the UI will be kept simple and clean.

UI Home page:

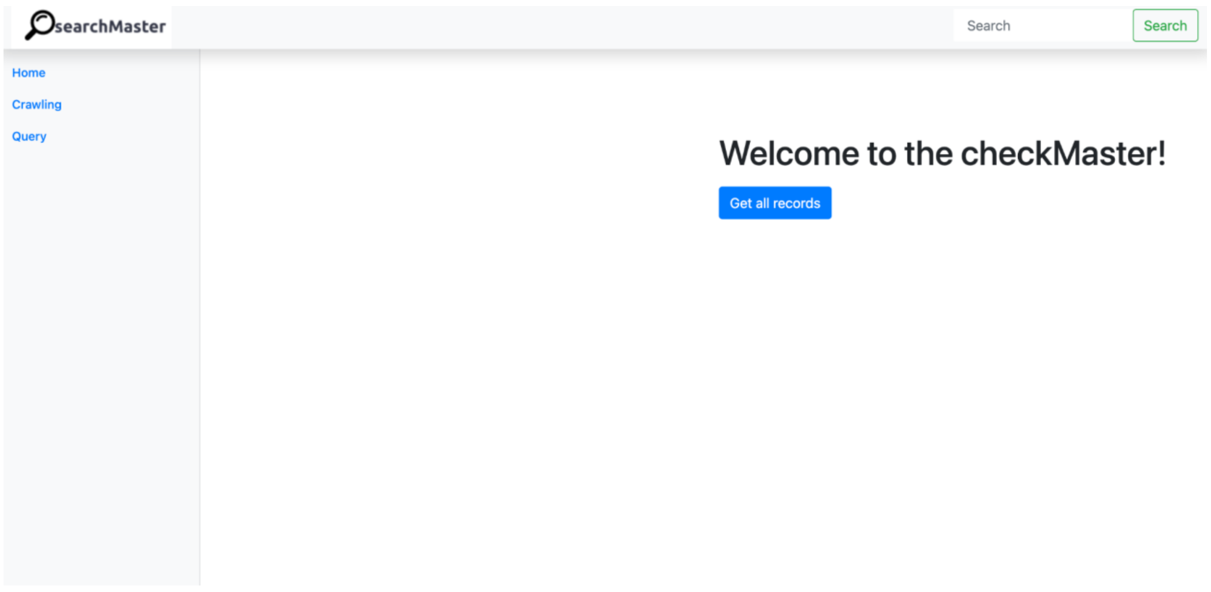
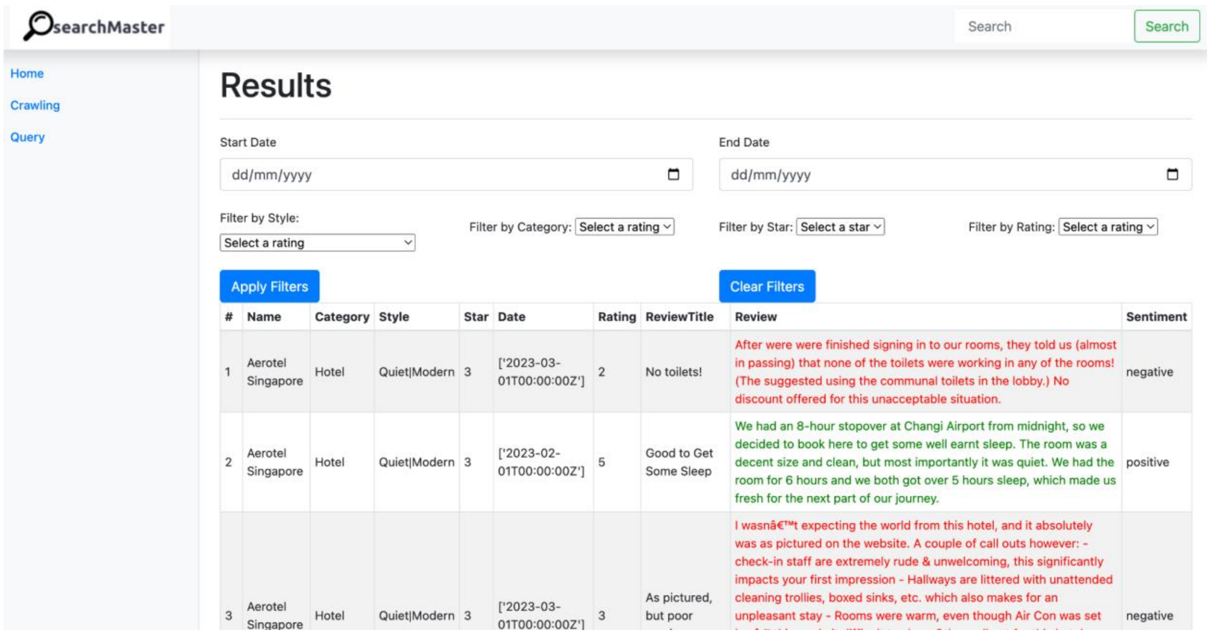


Figure 14: Home page of the web interface

Sidebar navigation allows for users' ease of access to different functions of the webpage. Simple blue button gets all records and allows the user to view all the records in the database.

UI for corpus search:



Results

Start Date: dd/mm/yyyy

End Date: dd/mm/yyyy

Filter by Style: Select a rating

Filter by Category: Select a rating

Filter by Star: Select a star

Filter by Rating: Select a rating

Apply Filters

Clear Filters

#	Name	Category	Style	Star	Date	Rating	ReviewTitle	Review	Sentiment
1	Aerotel Singapore	Hotel	Quiet|Modern	3	['2023-03-01T00:00:00Z']	2	No toilets!	After were were finished signing in to our rooms, they told us (almost in passing) that none of the toilets were working in any of the rooms! (The suggested using the communal toilets in the lobby.) No discount offered for this unacceptable situation.	negative
2	Aerotel Singapore	Hotel	Quiet|Modern	3	['2023-02-01T00:00:00Z']	5	Good to Get Some Sleep	We had an 8-hour stopover at Changi Airport from midnight, so we decided to book here to get some well earnt sleep. The room was a decent size and clean, but most importantly it was quiet. We had the room for 6 hours and we both got over 5 hours sleep, which made us fresh for the next part of our journey.	positive
3	Aerotel Singapore	Hotel	Quiet|Modern	3	['2023-03-01T00:00:00Z']	3	As pictured, but poor	I wasnâ€™t expecting the world from this hotel, and it absolutely was as pictured on the website. A couple of call outs however: - check-in staff are extremely rude & unwelcoming, this significantly impacts your first impression - Hallways are littered with unattended cleaning trollies, boxed sinks, etc. which also makes for an unpleasant stay - Rooms were warm, even though Air Con was set	negative

Figure 15: UI of corpus search

searchMaster allows users to browse through all the records in the database in a neat tabular manner which allows users to be able to digest the information easily. The filters are placed in the top half of the page allowing users to easily be able to filter and obtain the necessary information.

In addition, the search bar in the top right corner allows the user to search for reviews at any time on any page.

Reviews are also coloured with regards to their sentiment value which can be ‘positive’, ‘negative’ or ‘neutral’. Reviews with positive sentiments are coloured green while reviews which are negative or neutral are red in colour. This allows the user to clearly see the sentiments of any particular review without having to compare it with the sentiment column.

UI for crawling and incremental indexing of new data:

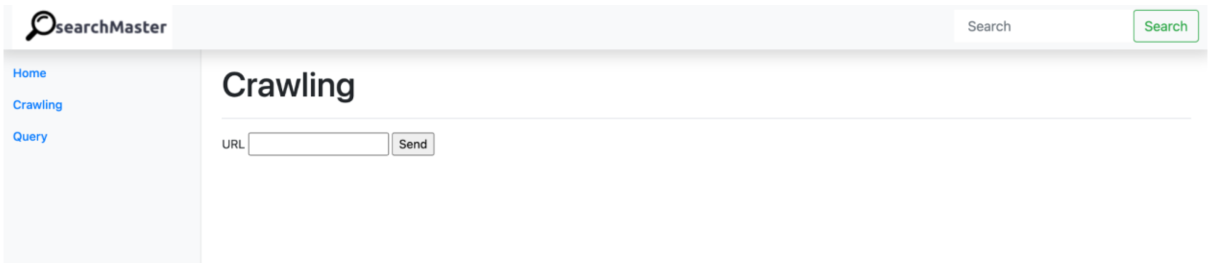
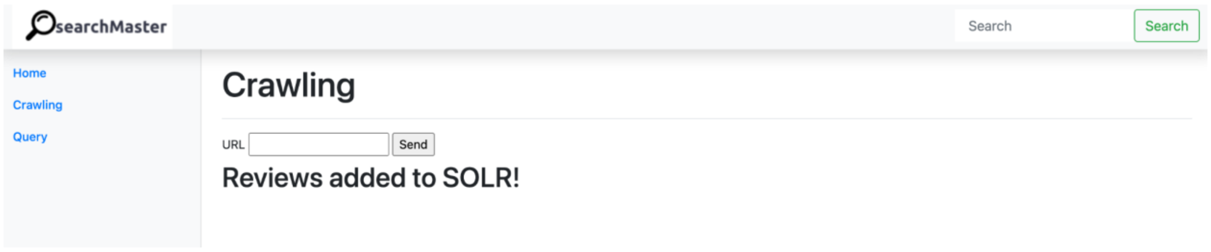


Figure 16: UI of crawling

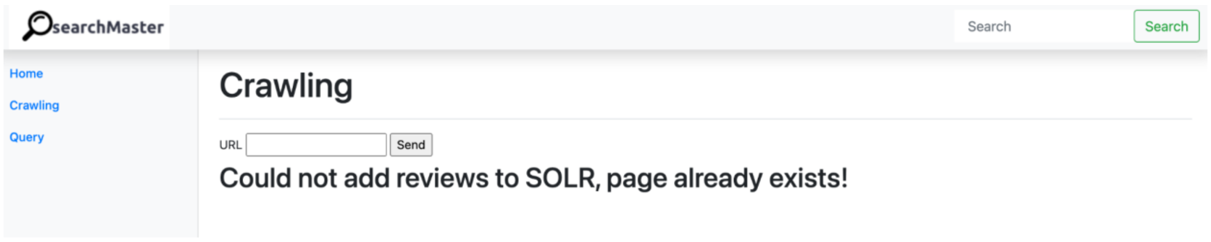
With only a simple input bar, the user is able to easily understand where to input the links for the application to start crawling.



Reviews added to SOLR!

Figure 17: Crawling output successful

After input of a link which has not been previously crawled, the response of the page will show a simple heading.



Could not add reviews to SOLR, page already exists!

Figure 18: Crawling output unsuccessful

After the input of a link which has already been crawled, the response of the page will show a simple heading to alert the user that the page already exists.

UI for location search:

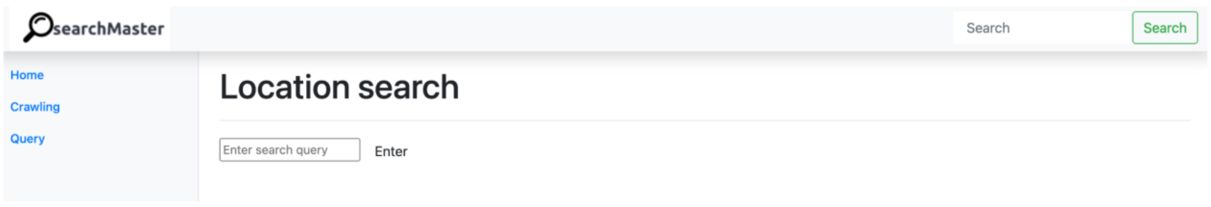
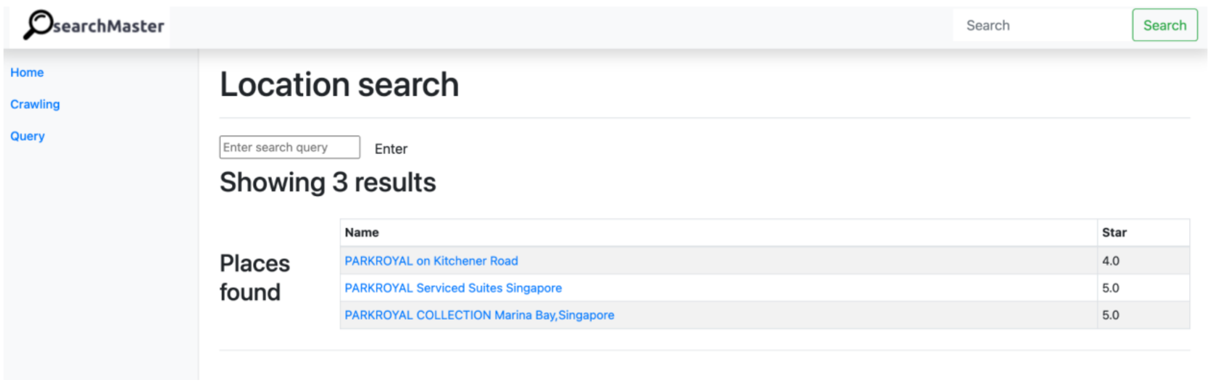


Figure 20: UI of location search

Location search is another feature of searchMaster. Location search enables the user with the ability to search for specific locations as well as offering search improvements such as geospatial search as well as recommendations such as more like this search.

UI for location page:



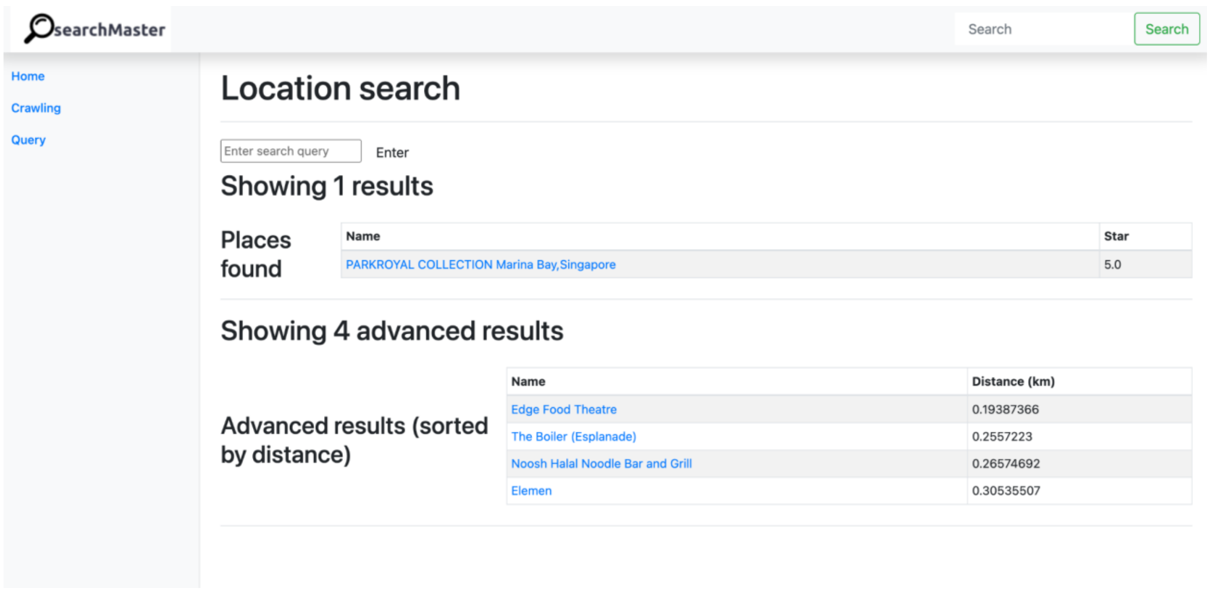
Showing 3 results

Places found

Name	Star
PARKROYAL on Kitchener Road	4.0
PARKROYAL Serviced Suites Singapore	5.0
PARKROYAL COLLECTION Marina Bay,Singapore	5.0

Figure 21: Query results for ‘parkroyal’

A simple search query yields a simple table which shows the places which match the query based on the Name index.



Location search

Enter search query Enter

Showing 1 results

Places found

Name	Star
PARKROYAL COLLECTION Marina Bay,Singapore	5.0

Showing 4 advanced results

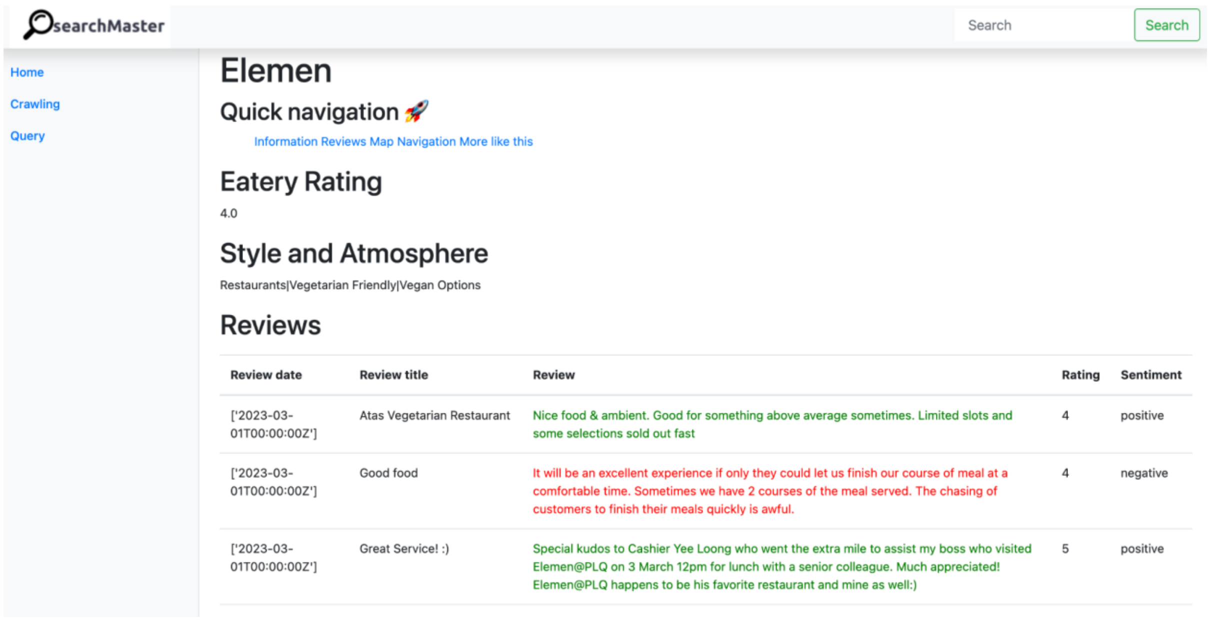
Advanced results (sorted by distance)

Name	Distance (km)
Edge Food Theatre	0.19387366
The Boiler (Esplanade)	0.2557223
Noosh Halal Noodle Bar and Grill	0.26574692
Elemen	0.30535507

Figure 22: Query results for 'top eateries near Parkroyal collection'

Search can also be tailored, searching using geospatial search yields results which are similarly presented in a tabular format. The reason for the similarity in the style of the output is so as to reduce information load on the user when they are presented with additional information.

UI of a location page instance:



Elemen

Quick navigation 🚀

Information Reviews Map Navigation More like this

Eatery Rating

4.0

Style and Atmosphere

Restaurants|Vegetarian Friendly|Vegan Options

Reviews

Review date	Review title	Review	Rating	Sentiment
['2023-03-01T00:00:00Z']	Atas Vegetarian Restaurant	Nice food & ambient. Good for something above average sometimes. Limited slots and some selections sold out fast	4	positive
['2023-03-01T00:00:00Z']	Good food	It will be an excellent experience if only they could let us finish our course of meal at a comfortable time. Sometimes we have 2 courses of the meal served. The chasing of customers to finish their meals quickly is awful.	4	negative
['2023-03-01T00:00:00Z']	Great Service! :)	Special kudos to Cashier Yee Loong who went the extra mile to assist my boss who visited Elemen@PLQ on 3 March 12pm for lunch with a senior colleague. Much appreciated! Elemen@PLQ happens to be his favorite restaurant and mine as well:)	5	positive

Figure 23: UI of a location instance

The hyperlink attached to every location name links to a single location page instance. This page provides additional information on top of the query indexing results such as geospatial graphs and recommendations to other places.

Queries, results and response time

Query 1: “JINJJA”, Qtime: 1ms; simple index search

http://localhost:8983/solr/reviews/select?fq=Name%3AJINJJA&indent=true&q.op=OR&q=*%3A*&useParams=

```
{
  "responseHeader":{
    "status":0,
    "QTime":1},
  "response":{"numFound":53,"start":0,"numFoundExact":true,"docs":[
      {
        "Name":"JINJJA Chicken",
        "spellCheck":["JINJJA Chicken"],
        "Category":"Eatery",
        "Style":"Restaurants|Fast Food|Halal",
        "distinctStyle":["Restaurants|Fast Food|Halal"],
        "Star":"4",
        "Date":["2019-10-01T00:00:00Z"],
        "Rating":"5",
        "ReviewTitle":"The fried chicken was amazing but didn't like kimbap. Visited far before coronavirus",
          "Review":"The fried chicken was delicious especially with the sauce. I'm craving it right now! The kimbap
wasn't nice but I don't think I like kimbap altogether. Visited way before the coronavirus",
        "Sentiment":"positive",
        "id":"3bd53922-a686-4a45-9b05-93f9f2bd1406",
        "_version_":1762344151331897348},
...
]}}
```

Query 2: Orchid~2, Qtime: 98ms; fuzzy search

http://localhost:8983/solr/reviews/select?fq=Name%3AOrchid~2&indent=true&q.op=OR&q=*%3A*&useParams=

```
{
  "responseHeader":{
    "status":0,
    "QTime":98},
  "response":{"numFound":738,"start":0,"numFoundExact":true,"docs":[
      {
        "Name":"Merci Marcel Orchard",
        "spellCheck":["Merci Marcel Orchard"],
        "Category":"Eatery",
        "Style":"Restaurants|Vegetarian Friendly|Vegan Options|Gluten Free Options",
        "distinctStyle":["Restaurants|Vegetarian Friendly|Vegan Options|Gluten Free Options"],
        "Star":"5",
        "Date":["2023-03-01T00:00:00Z"],
        "Rating":"5",
        "ReviewTitle":"Excellent â-ï¸â-ï¸â-ï¸â-ï¸â-ï¸",
```

        "Review":"I was very impressed with the quality of food and outstanding service from Abiel & Jasmine. Will surely be back soon!",
        "Sentiment":"positive",
        "id":"3a530623-1621-456d-b97b-1797b3789318",
        "_version_":1762344151557341185},
...]}}

3. “long wait time”~10, Qtime: 45ms; proximity search

http://localhost:8983/solr/reviews/select?indent=true&q.op=OR&q=Review%3A%E2%80%9Clong%20wait%20time%E2%80%9D~10&useParams=

{
  "responseHeader":{
    "status":0,
    "QTime":45},
  "response":{"numFound":700,"start":0,"numFoundExact":true,"docs":[
      {
        "Name":"Din Tai Fung",
        "spellCheck":["Din Tai Fung"],
        "Category":"Eatery",
        "Style":"Restaurants|Vegetarian Friendly|Vegan Options",
        "distinctStyle":["Restaurants|Vegetarian Friendly|Vegan Options"],
        "Star":"4",
        "Date":["2019-08-01T00:00:00Z"],
        "Rating":"4",
        "ReviewTitle":"Good Food",
        "Review":"Good Food, not long line waiting. Highly recommend itâ€™s super famous Xiao Long Bao. Very yummy. Itâ€™s great quality control, taste is exactly the same everywhere in the world.",
        "Sentiment":"positive",
        "id":"11269f27-d85b-4cdb-a410-dcf84fd63db5",
        "_version_":1762344150965944321},
...]}}

4. food^2 service, Qtime: 35ms; terms boosting

http://localhost:8983/solr/reviews/select?indent=true&q.op=OR&q=Review%3Afood%5E2%20service&useParams=

{
  "responseHeader":{
    "status":0,
    "QTime":35},
  "response":{"numFound":6164,"start":0,"numFoundExact":true,"docs":[
      {
        "Name":"Rosso Vino",
        "spellCheck":["Rosso Vino"],
        "Category":"Eatery",
        "Style":"Restaurants|Vegetarian Friendly|Vegan Options|Gluten Free Options",

"distinctStyle":["Restaurants|Vegetarian Friendly|Vegan Options|Gluten Free Options"],
"Star":"4",
"Date":["2023-02-01T00:00:00Z"],
"Rating":"5",
"ReviewTitle":"Superb river view Dinner",
"Review":"Fast reservation processing , quick serving of food . excellent Italian food and wine. Food is very fresh and delicious.",
"Sentiment":"positive",
"id":"36e1d6c2-5011-4181-b0a8-17575a1dd
"_version_":1762344151886594048},
...]}}

5. delic*, Qtime: 108ms; wildcard search

http://localhost:8983/solr/reviews/select?indent=true&q.op=OR&q=Review%3Adelic*&useParams=

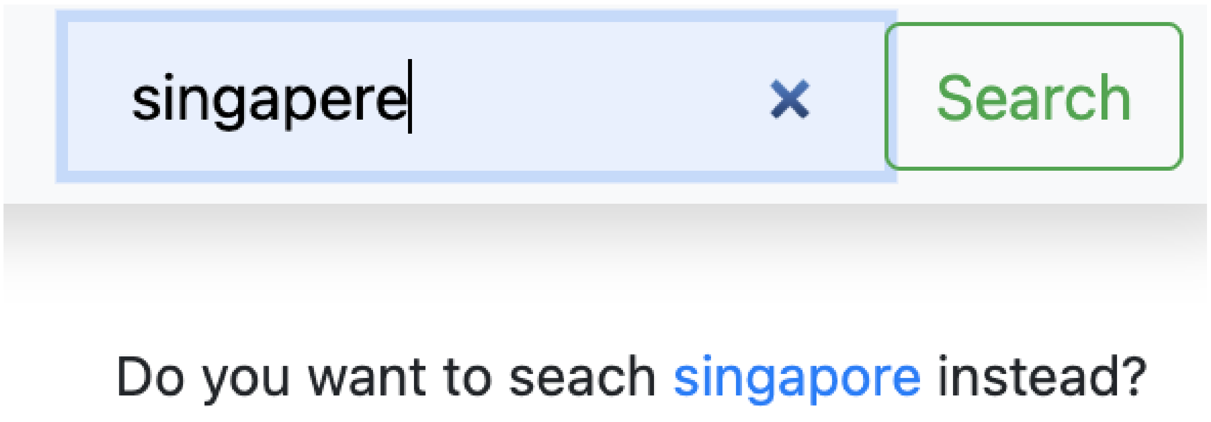
{
  "responseHeader":{
    "status":0,
    "QTime":108},
  "response":{"numFound":1021,"start":0,"numFoundExact":true,"docs":[
      {
        "Name":"10 Scotts",
        "spellCheck":["10 Scotts"],
        "Category":"Eatery",
        "Style":"Restaurants|Vegetarian Friendly|Gluten Free Options",
        "distinctStyle":["Restaurants|Vegetarian Friendly|Gluten Free Options"],
        "Star":"4",
        "Date":["2020-08-01T00:00:00Z"],
        "Rating":"5",
        "ReviewTitle":"High tea at 10 Scotts",
        "Review":"Excellent service by Shea and Nicolas - the food arrived promptly and was delicious. Would recommend the rosebuds tea",
        "Sentiment":"positive",
        "id":"14ca98ef-8c31-4307-b9f7-311147000b6c",
        "_version_":1762344150335750144},
...]}}

Question 3

In order to provide more accurate and relevant results to users, 4 innovations have been implemented in searchMaster for enhancing the indexing and ranking.

1. Spellcheck

Spellcheck is a feature that automatically provides spell suggestions to the user if the query may contain spelling errors or typos.
For instance, if the user wants to query "singapore", but he types it wrong and writes "singapere". Without spell check, the user will never find any records until he realises he misses spelling the word "singapore". With spell check, seachMaster will prompt the user that the query may contain miss spell words and give suggestions to the user. The user can click the suggested words for further query.



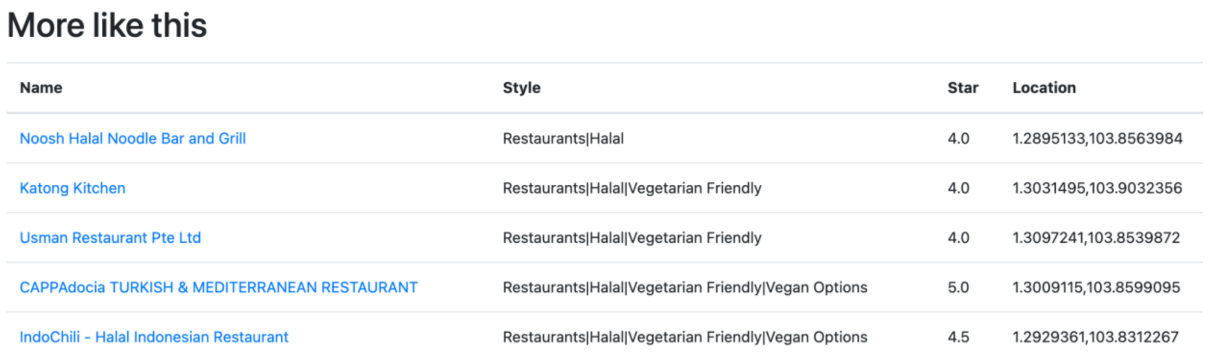
Do you want to seach singapore instead?

Figure 24: Spell Check for 'Singapore'

2. More like this

More like this is a search feature provided by Lucene, the search returns results that are similar to the results returned by the query. This query can be a field query or a function query. More like this is often used as a recommendation for users to get more results similar to the result they already searched for.

As an application for searching for hotels and eateries, searchMaster provides an all-round capability to provide additional recommendations to the user, based on the style of the hotel or eatery. One such example is when the user searches for a halal eatery “Noosh Halal Noodle Bar and Grill”, searchMaster will provide recommendations of other eateries which are also halal.


More like this

Name	Style	Star	Location
Noosh Halal Noodle Bar and Grill	Restaurants|Halal	4.0	1.2895133,103.8563984
Katong Kitchen	Restaurants|Halal|Vegetarian Friendly	4.0	1.3031495,103.9032356
Usman Restaurant Pte Ltd	Restaurants|Halal|Vegetarian Friendly	4.0	1.3097241,103.8539872
CAPPAdocia TURKISH & MEDITERRANEAN RESTAURANT	Restaurants|Halal|Vegetarian Friendly|Vegan Options	5.0	1.3009115,103.8599095
IndoChili - Halal Indonesian Restaurant	Restaurants|Halal|Vegetarian Friendly|Vegan Options	4.5	1.2929361,103.8312267

Figure 25: More like this recommendations

The more like this parser in Lucene works by tokenizing the search field, in this case, the style field and calculating the tf*idf for every token. In our case, since the style field contains the unique type of a particular eatery or hotel, the term frequency is at most 1, the document frequency is the number of eateries or hotels containing the unique style . The ranking of more like this is based on the ratings of the eateries or hotels similar to the current eatery or hotel.

More like this is used as a request handler instead of a search handler. To accommodate for this, a second core is created where the schema follows that of the first core ‘reviews’. In addition, the geo coordinates of every location are stored in this database. This reduces the overall size for storing additional information regarding a single place instance in the ‘reviews’ core.

URLs of query:
http://localhost:8983/solr/all_data/select?mlt=true&mlt.fl=Style&mlt.match.include=true&mlt.mindf=0&mlt.mintf=0&q=Name:%22Noosh%20Halal%20Noodle%20Bar%20and%20Grill%22

3. Multifaceted search

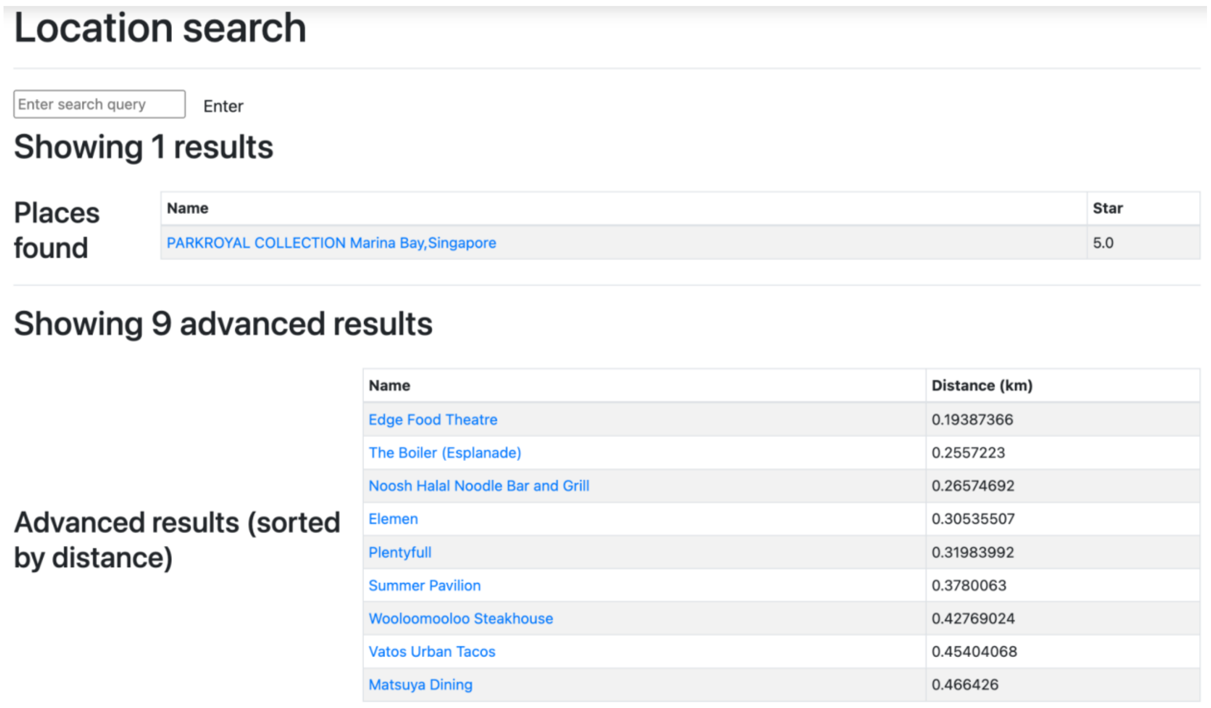
Multifaceted search allows users to refine their search results based on multiple aspects such as date range, style, category, star and rating. This feature is beneficial to the user because it makes the finding of specific records faster and easier, and brings a more customizable search experience to the user.

For example, when a user searches the hotel which contains “Parkroyal” in its name field, he will receive 195 records related to that query. User can filter the result based on the rating if he only wants to see the records which have a low rating. Without multifaceted search, the user has to read through all the results and find his desired information manually.

4. Geospatial search

In addition to multifaceted search, searchMaster is also able to cater to other aspects of the user’s needs. A use case is when the user would like to find eateries or hotels that are near their location. In a simple search, the user may not be able to find eateries or hotels which are close to them as a simple search does not have the capacity to infer the location, nor what the user would like to find. One such example is the search query ‘eateries near Parkroyal collection’. Parkroyal Collection is a hotel located in Marina Bay, Singapore, if the user inputs this query, the user would only be able to retrieve the hotel, this is however, not the information the user required, which are eateries near this hotel.

The solution is to incorporate geospatial search using SOLR. The coordinate locations of all places in the database are recorded in the database in geographical coordinates using google maps API, upon query runtime, SOLR will calculate the distance between 2 data entries using their respective locations and sort the results with respect to the distance. The results are returned to searchMaster and presented in such a way that is easily comprehensible to the user.


Location search

Enter search query Enter

Showing 1 results

Places found

Name	Star
PARKROYAL COLLECTION Marina Bay,Singapore	5.0

Showing 9 advanced results

Advanced results (sorted by distance)

Name	Distance (km)
Edge Food Theatre	0.19387366
The Boiler (Esplanade)	0.2557223
Noosh Halal Noodle Bar and Grill	0.26574692
Elemen	0.30535507
Plentyfull	0.31983992
Summer Pavilion	0.3780063
Wooloomooloo Steakhouse	0.42769024
Vatos Urban Tacos	0.45404068
Matsuya Dining	0.466426

Figure 26: Geospatial search

URLs of query:
http://localhost:8983/solr/all_data/select?q=*:*&fq={!geofilt}&sfield=location&pt=1.2917922,103.8571184&d=10&sort=geodist()%20asc&fl=location,Name

<u>Crawling to Indexing</u>

In order to showcase sentiment analysis in the web application, in the process of crawling a link, the sentiment of every review was also extracted and pushed into the SOLR database, in addition to the coordinates of the locations which are extracted using Google API.

To prepopulate the SOLR database, csv files were generated which stored information on top of those that were crawled from the Tripadvisor website, such as the sentiment as well as the latitude and longitude of each location. The coordinates of the locations were placed in a different csv file (all_data.csv) while the sentiment was placed together with the crawled reviews (reviews.csv).

Subsequent calls to crawl new links add data to the SOLR database as well as the 2 csv files for future prepopulation of SOLR database to be kept locally.

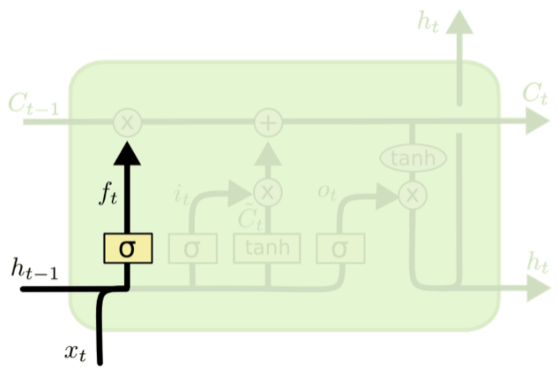
Classification

Question 4

Review of Existing Techniques

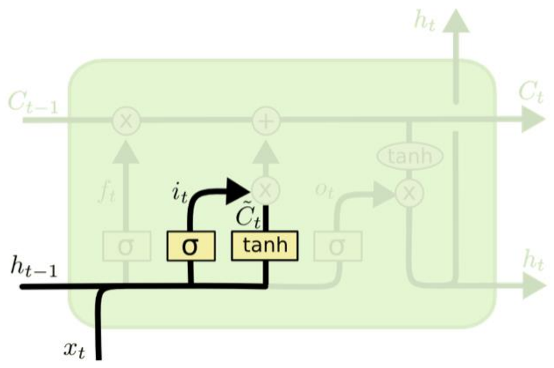
Long Short Term Memory (LSTM):

The introduction of LSTM units aims to solve the problem of vanishing gradients in RNNs, which happens when the magnitude of gradients decreases exponentially when propagated backwards in the network, causing weights in the early layers to be unable to update. RNNs therefore fail to capture long term dependencies in a sentence. LSTM units, on the other hand, can maintain information in memory across the layers using the cell state that runs across the units. A set of gates is used to control when information enters the memory, when it is output, and when it is forgotten.



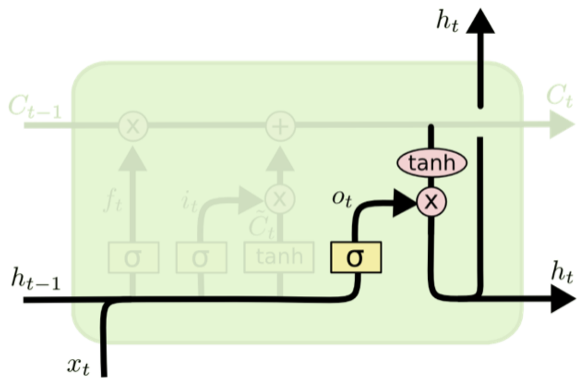
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Simply put, the forget gate controls how much previous information should be retained in the cell state C_t by looking at the previous hidden state h_{t-1} and the current input x_t , and this value is multiplied to the cell state.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

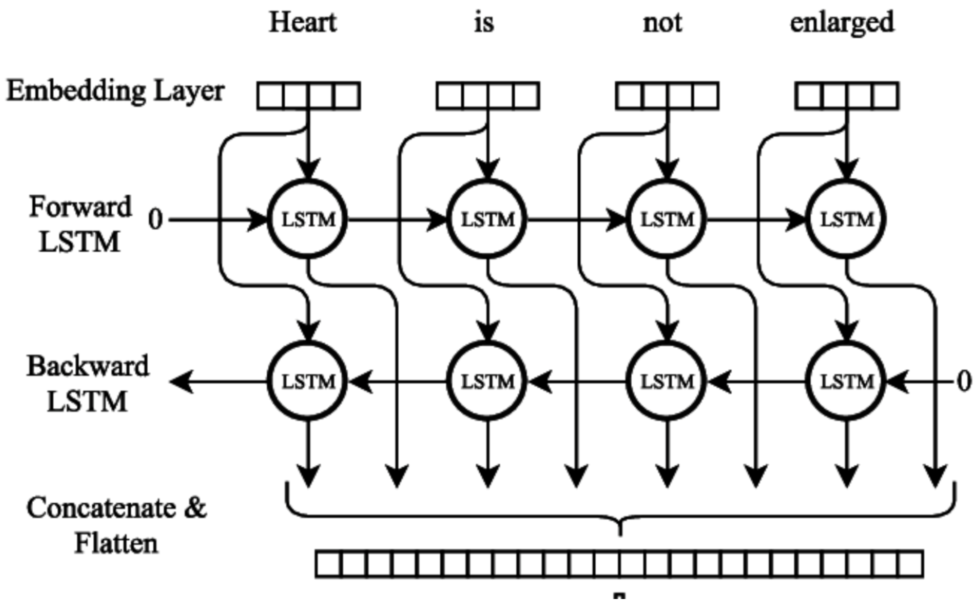
The input gate then decides what values to update (i_t) and add ($\hat{C}_t$) to cell state. These 2 values are combined and appended to the cell state.



$$o_t = \sigma\left(W_o\,[h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Finally, the output layer decides what value to output to the next LSTM unit and subsequent layer based on the current cell state, the previous hidden state and the current input.

Bidirectional LSTM:



The Bidirectional Long short term memory (Bi LSTM) essentially processes information from the front of the input to the back, and from the back to the front simultaneously. This helps the network to have access to more context information (both from the start of the sentence to current word and end of sentence to current word), and it is hence better able to capture long term dependencies.

Text preprocessing (stopwords removal, oversampling, tokenization, social media text processing etc.)

1. Since datasets are typically imbalanced, consisting of mostly positive reviews and lesser negative / neutral reviews, we utilised a module called "SMOTE"

which stands for Synthetic Minority Over-sampling Technique. This module comes from the imblearn library or imbalance learn library. The module "SMOTE" generates new samples by interpolation of the existing data; however, the data it generates will differ from the existing data. By using "SMOTE", we bring up the sample count of the minority classes to be the same amount as the majority class data to better train the model.

2. We adopted a lexicon based approach for microtext normalisation on the data, to handle acronyms and emoticons occurring in textual data from online sources like ours. Specifically, we imported a dataframe from the Abbreviations and Slangs for text preprocessing to convert common out-of-vocabulary (OOV) terms (internet abbreviations and textual emoticons) to in-vocabulary (IV) words. An example of our lexicon:

OOV form	**IV form**
g8	great
idk	I do not know
:-)	smile
:)	smile
:-(	frown

   Even though punctuations are typically removed from textual data during the preprocessing stage, some of these, such as emoticons, are important for sentiment analysis. We therefore convert them to in-vocabulary tokens before removal of other punctuations.
3. Short forms, such as "he's" or "I'm", were normalised into "he is" or "I am" respectively. This ensures that phrases are represented in a consistent format (in their expanded forms). Following which, unnecessary punctuations were then removed and words were also tokenized using the nltk tokenizer. This converts them into a more structured format for the machine learning task.
4. Stop words are frequently appearing words that typically carry little to no information. These words should be removed so that our models can focus on more informative words for sentiment analysis. The removal could also reduce the dimension of our input data to reduce computation costs. Although stop words may be important in providing contextual information in some cases, such as "flight ***to*** Singapore", we believe that this information is unlikely to be

useful for the task of sentiment analysis, and they incur more costs than benefits. We removed custom stopwords in the English language similar to that described in question 1, since using the NLTK stopwords will result in words like 'not' being removed, which will change the sentiment of sentences.

5. Words were then lemmatized using the NLTK WordNetLemmatizer to convert them into their meaningful base forms. We chose lemmatization instead of stemming since we will later use the gloVe embedding in the deep learning model and stemming may result in words not present in the English corpus or result in words of completely different meanings (eg. "cater" → "cat"), which in some cases would also affect the polarity of words.

Performance Metrics

To evaluate the models used for classification we used a few different metrics to better understand how our model is working. The metrics used are:

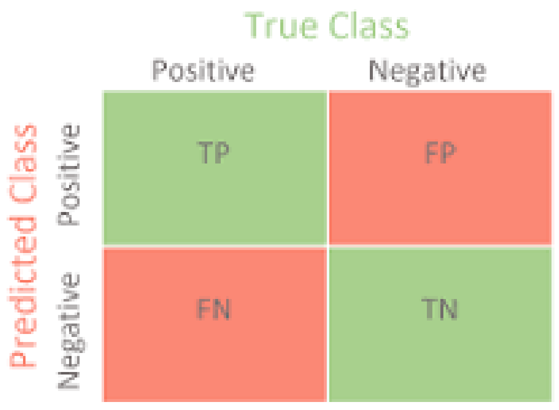
- Records classified per second
- Confusion matrix
- Precision
- Recall
- F1-score
- Accuracy

Records classified per second

This metric measures the time taken for our model to make a prediction on a record after it has been properly processed. Since we are running our models on large amounts of data at a time, it is important that this should be as low as possible.

Confusion Matrix

A confusion matrix is a summary of the prediction results. The number of correct and incorrect predictions are summarised with count values and broken down by each class. The four different values are true positives and true negatives and false positives and false negatives. True positives and true negatives are the rightly predicted labels for the record while false positives and false negatives are the incorrectly predicted labels for the records respectively. A false positive is a record labelled positive but is supposed to be labelled negative and vice versa for false negative.



Precision

Precision is defined as the number of predictions labelled positive that is actually positive or the proportion of positive identifications labelled correctly. The precision metric should be a value closer to 1. It is defined by the following equation.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall is similar to precision but answers a different question, What proportion of actual positives was identified correctly? It is defined as the number of predictions labelled positive over the total number of actual positives. Like precision, recall should be a value closer to 1 and is defined by the following equation.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

F1-score is a derived metric from both precision and recall and is defined as the harmonic mean of precision and recall. In the F1 score, we compute the average of precision and recall. A model will obtain a high F1 score if both Precision and Recall are high. A model will obtain a low F1 score if both Precision and Recall are low. A model will obtain a medium F1 score if one of Precision and Recall is low and the other is high. F1-score is defined by the following equation.

$$F1\,score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy

Accuracy is a simple classification matrix and will be used heavily for our models. It is defined as the number of rightly predicted labels over the total number of predictions.

Sentiment Analysis

<u>Subjectivity detection</u>

The aim of subjectivity detection is to detect whether a piece of text is subjective, that is to say, whether a piece of text is neutral or polarising. From the input data received, the data that would be polarising would be data that are positive or negative sentiment with ratings of 1-2 star and 4-5 star while neutral would represent a 3 star rating. However, this brings about a problem. The data is too heavily skewed to polarising data as it consists of 4 out 5 ratings, not to mention the majority of data are positive to begin with.

To tackle this issue, the SMOTE technique, as described in the preprocessing section was used to oversample neutral data points. If SMOTE was not used, due to the imbalance of the training dataset, the model will simply predict almost all of the data points to be a polarising data point as this will lead to a high accuracy. The final model used for prediction is a bi-directional lstm model with the following architecture.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 8)           476008

 bidirectional (Bidirectiona  (None, 64)               10496
 l)

 dropout (Dropout)           (None, 64)                0

 dense (Dense)               (None, 16)                1040

 dropout_1 (Dropout)         (None, 16)                0

 dense_1 (Dense)             (None, 1)                 17

=================================================================
Total params: 487,561
Trainable params: 487,561
Non-trainable params: 0
```

The validation accuracy for the model after training is 0.893.

The model makes use of a bi-directional lstm model and a dense layer with dropout included to prevent overfitting. The model uses the Adam optimizer and binary cross-entropy loss. Other models have been tested on the dataset and the results will be discussed in the ensemble classification section.

Polarity detection

A pre-trained global vectors for word representation (GloVe) word embedding file, 'glove.6B.300d.txt', was loaded to obtain word embeddings for our dataset. The embeddings have been obtained from a preconstructed word vectorization model on a 6 billion word corpus (Wikipedia 2014 + Gigaword 5). It consists of 400000 pretrained 300 dimensional vectors (each glove vector representing one unique word). Specifically, it was obtained by constructing a large co-occurrence matrix representing how often each word (row) is seen in each context (column). This matrix is then reduced into a lower-dimensional matrix by minimising a reconstruction loss to encourage the reduced representation to capture the variance of the original matrix. The loaded embedding matrix is then used to lookup the words in the embedding layer of the neural network.

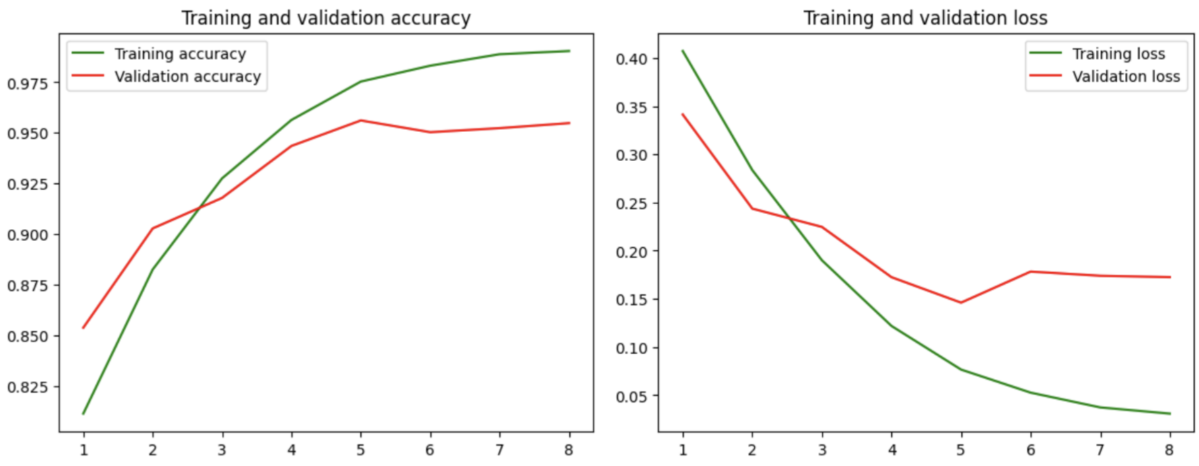
The architecture of the network is as such, with a bidirectional LSTM and dense layer following the embedding layer. The model was trained on the Adam optimizer and binary cross-entropy loss, along with early-stopping mechanism to prevent over-fitting.

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 25, 300)           20897700

bidirectional (Bidirectional) (None, 128)               186880

dense (Dense)                (None, 1)
129
=================================================================
Total params: 21,084,709
Trainable params: 187,009
Non-trainable params: 20,897,700
```

Other architectures have also been experimented for this task, including a uni-directional LSTM and the addition of an attention layer.

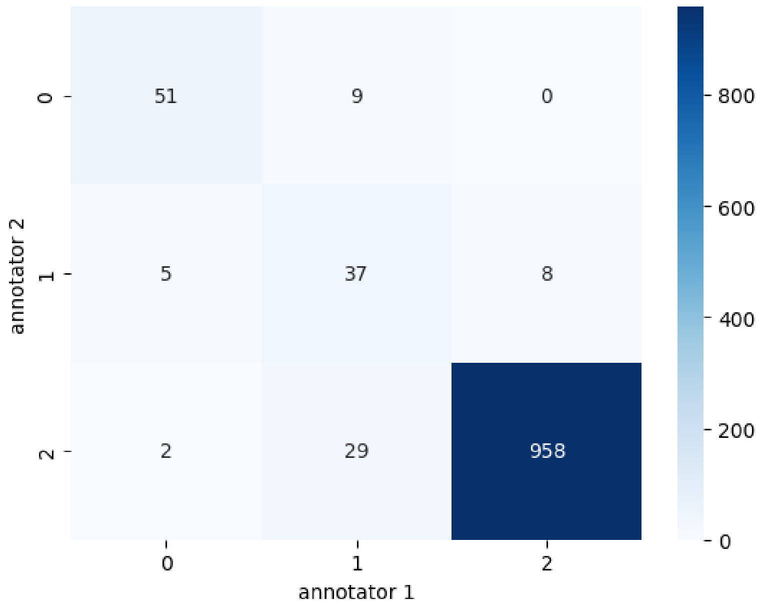


The model achieved a validation accuracy of 95.6%

Evaluation of Models

Evaluation dataset

Two of our annotators annotated 1100 crawled reviews, with 0 being neutral, 1 being negative and 2 being positive. We achieved an inter-annotator agreement of 95.18%. This will form our evaluation dataset for this section.



Inter-annotator agreement = 95.18%

Confusion Matrix

Subjectivity Model

2	48
0	1049

Polarity Model

29	29
22	944

Subjectivity Model

	Precision	Recall	f1-score	support
0	1.00	0.04	0.08	50
1	0.96	1.00	0.98	1049
accuracy			0.96	1099
Macro avg	0.98	0.52	0.53	1099
Weighted avg	0.96	0.96	0.94	1099

Polarity Model

	Precision	Recall	f1-score	support
0	0.57	0.5	0.53	58
1	0.96	1.00	0.98	966
accuracy			0.95	1024
Macro avg	0.77	0.74	0.75	1024
Weighted avg	0.95	0.95	0.95	1024

Subjectivity Prediction

From the results above we can see that the subjectivity model despite having high accuracy and high precision has a rather low f1-score. This is due to the fact that the data is heavily imbalanced with only 50 neutral reviews as opposed to 1049 polarising reviews. The model did not manage to accurately predict neutral reviews despite the use of SMOTE giving the model more training data, only predicting 2 neutral reviews accurately but managed to predict all polarising reviews accurately as it managed a recall score of 1.00. Overall, the model managed a 0.94 weighted average f1-score.

Several random test cases:

Test sentence	Model prediction	Model prediction score
A lovely ambience. Good food. Enjoyed the high tea. There was good balance of savoury and sweet dishes	opinionated	61.8786
Food was normal. Could have been better.	opinionated	59.8784
Manager was playing loud shooting game on phone in restaurant. He had to finish his game before he could be bothered to take our payment. Really small portions.	opinionated	62.1914
It's alright, nothing spectacular. Food was just about average.	opinionated	51.1484
Waiter was rude and food was not fresh. Don't come here :-(	opinionated	60.7621

Polarity prediction

For the polarity model, despite having a very imbalanced dataset, only having 51 negative reviews compared to 973 positive reviews, it managed to perform decently. The weighted average for its f1-score was at an impressive 0.95, in line with the accuracy of the model as well. While the subjectivity model scored a low f1-score for its neutral reviews, the polarity model managed a 0.53 f1-score for its negative reviews and 0.98 f1-score for its positive reviews.

Several random test cases:

Test sentence	Model prediction	Model prediction score
A lovely ambience. Good food. Enjoyed the high tea. There was good balance of savoury and sweet dishes	positive	99.9999
I was there celebrating my wife's birthday. The food and service are excellent! Will definitely visit again.	positive	99.9999
Manager was playing loud shooting game on phone in restaurant. He had to finish his game before he could be bothered to take our payment. Really small	negative	2.5443

portions.		
Delicious food. Good drinks. Must visit! Authentic flavours. Good vibes.	positive	99.9998
Waiter was rude and food was not fresh. Don't come here :-(	negative	0.2164

In summary, both models have the ability to predict reviews to a certain extent with high weighted average f1-score for both models.

Question 5

Ensemble classification

We have tested the use of ensemble learning in an attempt to improve the overall accuracy of the prediction. The idea behind ensemble learning is that by combining the predictions of multiple models, we can reduce the impact of individual model's weaknesses and improve the overall accuracy of the final prediction. This is because each model in the ensemble can capture different aspects of the problem or make different errors, and by combining their predictions, we can take advantage of their complementary strengths and reduce the impact of their weaknesses. The models used for the ensemble network are as follows: Random Forest, Logistic Regression, XGBoost and the previously mentioned bi-directional LSTM model. For our ensemble network, we made use of the stacking technique whereby we use the results of the various base models as inputs to a meta model. The meta model is a simple feedforward neural network. The results are shown in the table below.

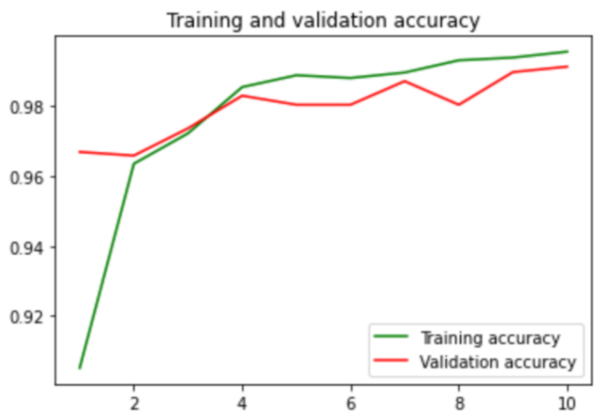
Model	Accuracy
Bi-Directional LSTM	0.893
XGBoost	0.872
Random Forest	0.873
Logistic Regression	0.818
Ensemble Network	0.893

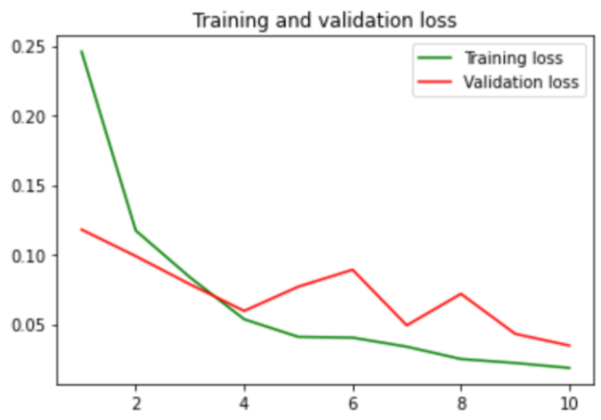
From the results above, we conclude that the ensemble network did not manage to improve the accuracy of the predictions, instead, the result was that the network has the exact same accuracy as the LSTM model, the one with the best results. One of the reasons we think is the cause is that all the various models have very similar predictions in that the predictions are more skewed towards the polarising data. Therefore, the model cannot be simply improved by making use of ensemble learning as there is no variance in the predictions despite making use of various models. With no variance in the predictions by its base models, the ensemble network will only be able to give its prediction equal to the base model with the highest accuracy, that being the LSTM model. Hence, the LSTM model was chosen.

Spam detection

Spam detection aims to identify unsolicited advertisements, illegitimate comments, or those which can post security threats to users (eg. scams). Web spams / ads are still present even on popular social media sites today, and they can significantly deteriorate the quality of results displayed. In an actual application, spam detection could be used to downgrade the ranking of spam comments or eliminate it from the index. This section will focus on the machine learning techniques used to classify spam comments, which yielded compelling results.

An external labelled dataset was used for spam detection training (https://www.kaggle.com/datasets/washingtongold/spam-or-ham-emp-week-2-ml-hw-dataset).
The dataset consisted of 5169 uncleaned SMS text data from sources including the SMS Spam Corpus v.0.1, NUS SMS Corpus and Grumbletext website. Since the dataset consisted of about 87% non-spam texts, the dataset was balanced using a similar approach as described in question 4.
A combination of deep learning and decision tree classifier was used for the spam detection. For the deep learning model, the preprocessing steps and architecture is similar to that used in polarity detection, eventually obtaining a validation accuracy of almost 99%.



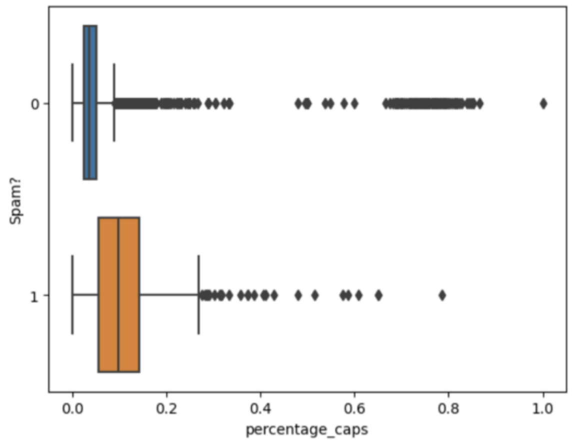


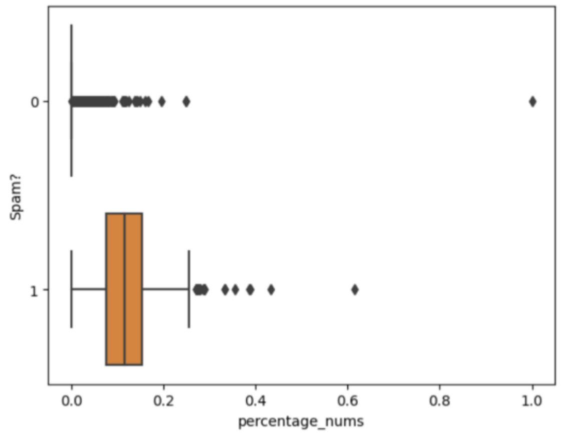
Several custom test cases:

Test sentence	Model prediction	Model prediction score
Welcome! Please reply with your AGE and GENDER to begin. e.g 24M	Spam	98.7303
YOU HAVE WON! As a valued Vodafone customer our computer has picked YOU to win a £150 prize. To collect is easy. Just call 09061743386	Spam	99.9907

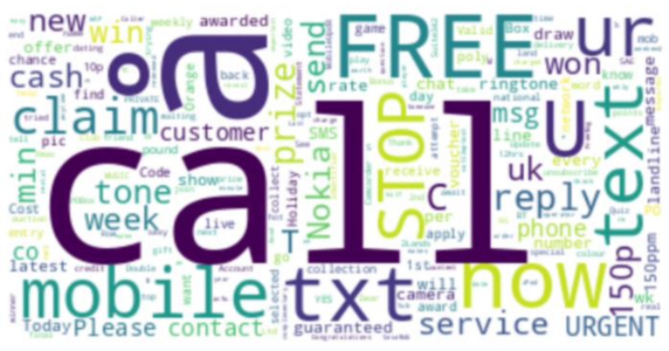
The waitress was attentive to detail and was very helpful in every aspect. The food was delicious as usual! Would come back here again.	Not spam	0.0354
Superb service, nice ambience at their new north point outlet.	Not spam	0.0623
Food was affordable, but I saw them taking it out from the microwave.	Not spam	0.1170

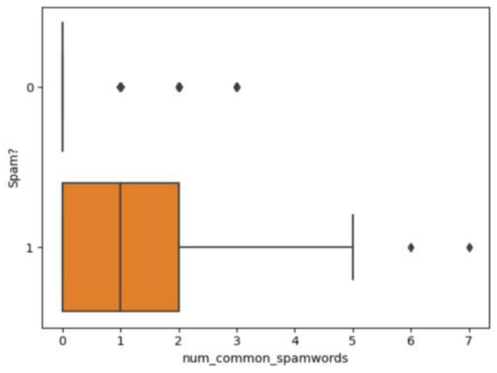
However, since aspects of the text such as capitalization, special symbols and hyperlinks have to be normalised or removed in order to convert the text into vector embeddings for network training, they cannot be captured when using the neural network alone.





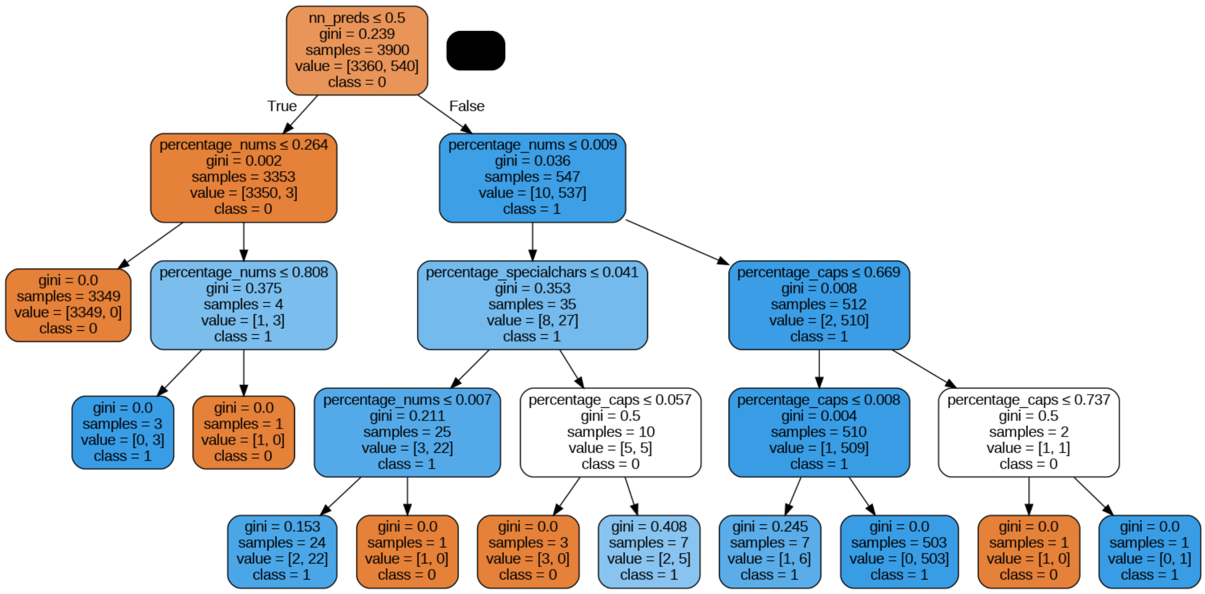
Nonetheless, such information can be important for the task of spam detection. From the boxplots, spam texts tend to have a significantly high proportion of capital letters and numbers, for instance. Certain words also occur much more commonly in spam texts. We define the list of most common spam words to be : `common_spamwords = ['free', 'call', 'text', 'mobile', 'now', 'txt', 'won', 'prize', 'claim', 'cash', 'stop', 'reply']`





The motivation is to account for these indicative factors, that would have otherwise not been captured in word embeddings, into the task of spam detection as well. A decision tree was being fitted on the following predictor variables:

- Number of common spam words: occurrences of common spam words text
- Percentage caps: total capitalised characters / total characters
- Percentage numbers: total numeric characters / total characters
- Percentage special characters: total non alpha-numeric characters (excluding ‘ . ’ and ‘ , ’) / total characters
- Number of urls: number of hyperlinks in text (detected using regex)
- Model predictions: float value of deep learning model output



nn_preds ≤ 0.5 gini = 0.239 samples = 3900 value = [3360, 540] class = 0
True
False
percentage_nums ≤ 0.264 gini = 0.002 samples = 3353 value = [3350, 3] class = 0
percentage_nums ≤ 0.009 gini = 0.036 samples = 547 value = [10, 537] class = 1
gini = 0.0 samples = 3349 value = [3349, 0] class = 0
percentage_nums ≤ 0.808 gini = 0.375 samples = 4 value = [1, 3] class = 1
percentage_specialchars ≤ 0.041 gini = 0.353 samples = 35 value = [8, 27] class = 1
percentage_caps ≤ 0.669 gini = 0.008 samples = 512 value = [2, 510] class = 1
gini = 0.0 samples = 3 value = [0, 3] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
percentage_nums ≤ 0.007 gini = 0.211 samples = 25 value = [3, 22] class = 1
percentage_caps ≤ 0.057 gini = 0.5 samples = 10 value = [5, 5] class = 0
percentage_caps ≤ 0.008 gini = 0.004 samples = 510 value = [1, 509] class = 1
percentage_caps ≤ 0.737 gini = 0.5 samples = 2 value = [1, 1] class = 0
gini = 0.153 samples = 24 value = [2, 22] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.408 samples = 7 value = [2, 5] class = 1
gini = 0.245 samples = 7 value = [1, 6] class = 1
gini = 0.0 samples = 503 value = [0, 503] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1

The final decision tree model achieved an accuracy of 99.7% on the test set for spam detection.

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1465
           1       0.99      1.00      0.99       207

    accuracy                           1.00      1672
   macro avg       0.99      1.00      0.99      1672
weighted avg       1.00      1.00      1.00      1672

Confusion matrix:
[[1462    3]
 [   1  206]]
```

Links

Video Presentation YouTube Link:
https://www.youtube.com/watch?v=52HXfjVqK5I

Data Files Link:
https://drive.google.com/drive/folders/1f2owZIm92et0j68PFGW56dT88EeyO4SU?usp=sharing

Source Code Link:
https://drive.google.com/drive/folders/1PlaX4C79mrZsqbq-BA1qZkd9I9I7U_il?usp=sharing