Exploring Natural Language Processing and Recommender Systems on TripAdvisor.com data

Utilizing state-of-the-art techniques for Travel Sustainability.

[12]

Master's Thesis

Guillermo Daniel Calvo Altesor

Master of Science in Data Science & Big Data

Innovation & Entrepreneurship Business School (IEBS)

Barcelona 2022

Summary

The notorious GDP (gross domestic product) and job creation figures that place the Tourism Industry as the main one in the Canary Islands (Spain) and the recession that these have had due to the Covid-19 pandemic, show an opportunity to rebuild towards sustainable tourism. To do this, this research proposes the use of data from TripAdvisor and its processing with state-of-the-art techniques in Data Science. All of these under the paradigm that the use of NLP or Recommender Systems in these data can constitute the creation of marketable tools for sustainable tourism, based on the successes that these technologies have had in other industries (Netflix, Amazon, Spotify, etc.). The research meets its objectives aimed at obtaining data from TripAdvisor, the use of Natural Language Processing (by extracting Topics and Sentiment Analysis) and the creation of Recommender Systems of the Collaborative Filtering model-based type, as well as the interpretation of this data through visualizations and reports. Finally, it is proposed for a future project, the capture of streaming data and a Hybrid Recommender System collaborate in the creation of a marketable product for the generation of sustainable tourism.

1. Introduction

The Tourism Industry is one of the great generators of work and GDP in the Canary Islands of Spain with figures that in recent years have approached 40% in both cases. Due to the Covid-19 pandemic, these figures have been reduced by 65% and begin to recover in 2022. In this period, this research considers possibilities when it comes to tourist reconstruction, taking into account the context and evaluating the data from the beginning of these recessions, due to the importance that must be given to this topic and to the transformations that it demands.

Taking this context into account, the concept of e-Travel comes into play and the importance of platforms such as TripAdvisor in travel planning, as well as in the search for tourist attractions for people who plan their trips. Similarly, in the different investigations carried out in the theoretical framework, this type of platforms and the content shared in them (especially the ratings and reviews of the activities) are valued for their use in the transformation of tourism. Although, of course, the search for increased consumption and the generation of capital is vital, this project values this transformation as a possibility for creating Travel Sustainability. Which does not have to mean reduction, but a rethinking of tourism and a more intelligent distribution.

The possibilities from Data Science to achieve this type of transformation are significant and multiple. Therefore, this research brings to the forefront the state-of-the-art techniques of Natural Language Processing (NLP) and Recommender Systems or Recommendation Systems (RS) by processing data from portals such as TripAdvisor, seeking the creation of new features within them, or the extraction of knowledge from them for Online Travel Agencies (OTA) or Destination Management Organizations (DMO).

This research raises two main objectives, one academic, oriented towards the exploration of the named techniques, contributing to the academic environment, and another business type, with the intention of building tools and reports that can be the beginning of a larger, marketable project, to contribute to a desired sustainable transition in the hands of technology and Artificial Intelligence.

The methodology consists of extracting data from TripAdvisor using known Scraping techniques such as *Scrapy*, *Beautifulsoup, Selenium*, and *Octoparse API*. Next, preparation of the data is carried out and, after this, the following phases of NLP processing, creation of

RS and the extraction of tables and visualizations. All of these are done in different Jupyter Notebooks with Python.

In the main conclusion, the use of *Octoparse* for the extraction of data from TripAdvisor is recommended, due to the facilities, it shows to overcome the obstacles that the TripAdvisor platform has when interpreting its HTML. The use of transformers for NLP processing is also recommended, specifically, BERTopic stands out in Topic Modelling and roBERTa in Sentiment Analysis (SA). In addition, it is concluded that the use of model-based Collaborative Filtering RS is adequate for the processing of this type of information, taking into account the ability to improve the models developed by including more data and exploring improvements in its learning.

Finally, it is concluded that the TripAdvisor data is representative and valid, coinciding with the decrease in reviews with the previously mentioned Tourism Industry recession. It is interpreted that the users of these portals contribute to the growth of e-Travel, by promoting active and conscious tourism while providing feedback that can be used both by the traveller network, as well as by the entities that surround the industry, OTAs and DMOs.

In future work, the exploration of new models of RS is proposed together with the injection of data in the form of streaming with its corresponding pipeline or data flow systems, for the development of a marketable RS for Travel Sustainability purposes.

2. State of the Art

2.1 Contextualization

2.1.1 The Tourism Industry in the Canary Islands

According to statistics published on the Canary Islands Government website (gobiernodecanarias.org/istac/), the Tourism Industry has meant in recent years in the Canary Islands around 35% of employment and a 35% of GDP, which shows an economy designed towards this sector and to a certain point, dependent on it. With the Covid-19 pandemic, beginning on December of 2019, but really impacting its measures to restrict freedoms due to a state of alarm in Spain, in mid-March 2020, these figures were punished: Tourism GDP droped to 11%, as well as employment. We can also see this in the following graphic, representing the amount of tourists who visited the Canary Islands lately:

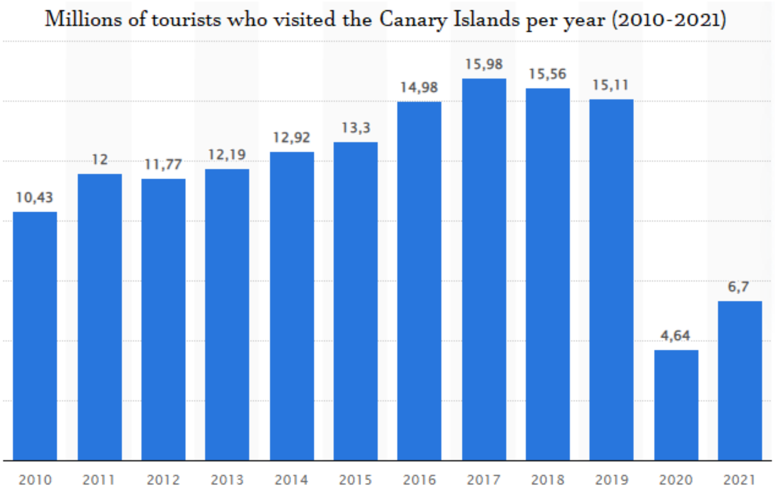


Figure 1. Source: es.statista.com

2.1.2 Application of Data Science to the Tourism Industry

Today, e-Travel is one of the preferred ways of planning trips by tourists. There are countless webs, portals and applications to investigate (Trivago, TripAdvisor, Expedia, Booking, etc.). These portals collect a quantity of information about users and destinations that is becoming more extensive and precise every day (photographs, ratings, definitions, prices, locations, reviews, etc.), which makes them an ideal source of data for research and creation of Machine Learning and Deep Learning models.

Among the most valued variables for the practice of these disciplines are the *reviews*, where users freely and extensively comment on their opinion on a Point of Interest (POI) (eg “El Teide Volcano” or “Timanfaya Park”). Information that can be processed in NLP and RS techniques and marketable through the creation of Travel Apps, for use in OTAs or DMOs as well as for Travel Sustainability reasons: to manage and reduce the impact of tourism or to find nearby similar alternatives, reducing the emissions. [1]

2.1.3 TripAdvisor's Reviews

According to the study "The Power of Reviews: How TripAdvisor Reviews Lead to Booking and Better Travel Experiences" [2], 3 out of 4 respondents value reviewing reviews before taking a trip as extremely or very important in travel decisions. The portal has more than 1 billion of them, 26 million submitted in 2021 with an average of 633 characters (3 times larger than other OTAs), which means that most of the reviews are long-form, something very attractive from the point of view of text processing or NLP, because the greater the content, the more possibility of extracting knowledge from it and of creating models that interpret them.

TripAdvisor reviews have the attributes shown below in Figure 1 and can be seen in Appendix 1. It should be emphasized that this type of platform has profoundly changed the way tourism is consumed and the way in which travel information is searched and shared [3 and 4].

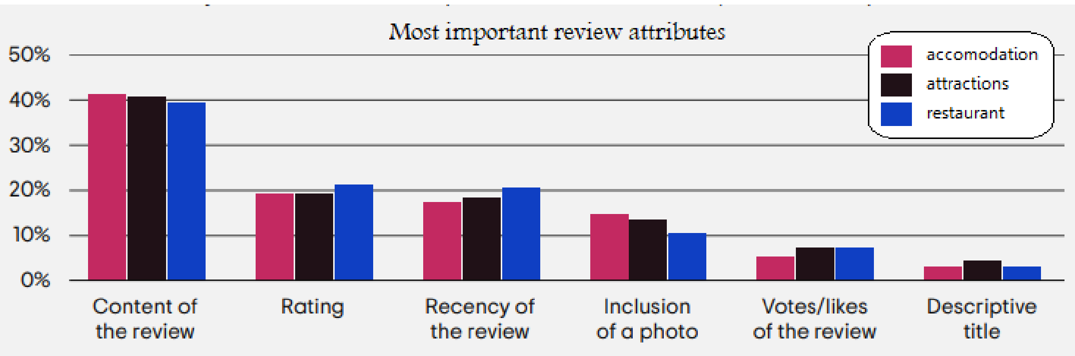


Figure 2. Source: Tripadvisor.com

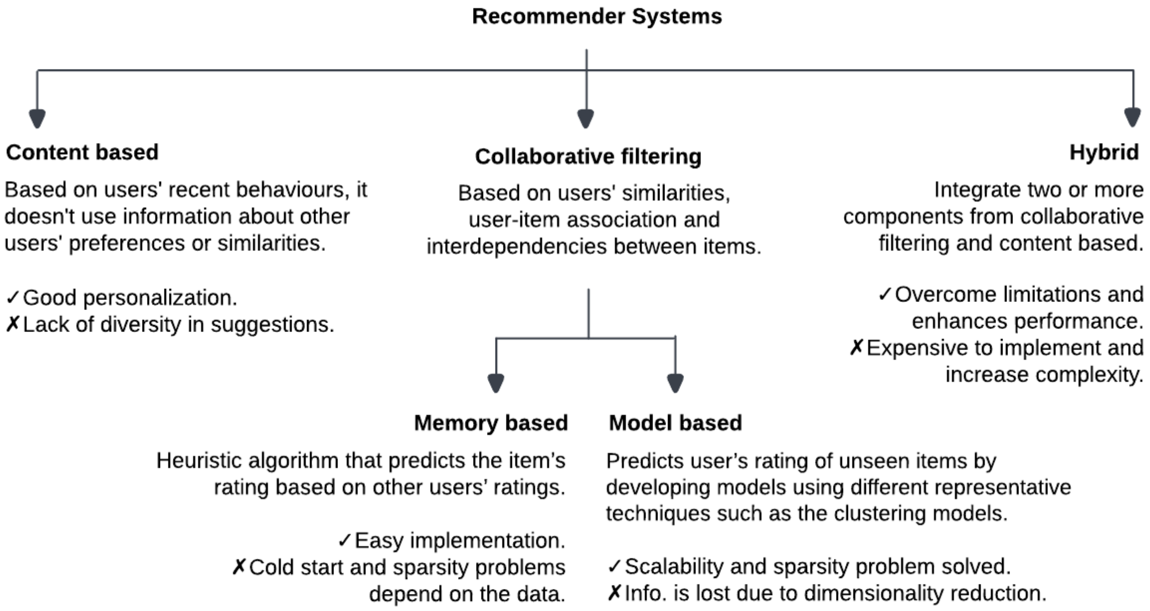
2.2 Technical conceptualization

This section defines the main concepts that will be dealt with throughout the investigation. In the first order, the different types of RS will be defined, and then the scope of the NLP techniques will be explained.

2.2.1 Recommender Systems

A large amount of information available causes a challenge to users, making the decision-making more complex, the RS filter this information and solve information overload. Although they first burst were by changing the way users consume movies or music (Netflix and Spotify), RS are taking a leading role across a number of leading industries and platforms (Amazon, TripAdvisor, Facebook, etc.). The type of navigation that the user performs, his profile and similarity with other profiles, or even his feedback are the bases of these successful instruments. Its function is to suggest to the user what is defined as the "best next action", in other words, what the algorithms predict that the user might like using logic and relationships between items (e.g., movie or POI) and users.

RS algorithms face challenges of *cold start* (lack of data to recommend certain users) or *sparsity* (polarization of user profiles, for example) that different models try to reduce and shovel. RS models or approaches can be summarized as follows [5]:



Recommender Systems

Content based

Based on users' recent behaviours, it doesn't use information about other users' preferences or similarities.

✓Good personalization.
✗Lack of diversity in suggestions.

Collaborative filtering

Based on users' similarities, user-item association and interdependencies between items.

Hybrid

Integrate two or more components from collaborative filtering and content based.

✓Overcome limitations and enhances performance.
✗Expensive to implement and increase complexity.

Memory based

Heuristic algorithm that predicts the item's rating based on other users' ratings.

✓Easy implementation.
✗Cold start and sparsity problems depend on the data.

Model based

Predicts user's rating of unseen items by developing models using different representative techniques such as the clustering models.

✓Scalability and sparsity problem solved.
✗Info. is lost due to dimensionality reduction.

Figure 3. Own source

2.2.2 Natural Language Processing

As its name suggests, it consists of a set of techniques that have their roots in teaching computers to understand and speak natural languages, which have different productive branches. Among them are SA*,* where it is interpreted if, for example, a review has a positive, neutral or black sentiment; *document classification,* where the texts are also classified according to classes; *autocomplete*, which can complete sentences, as is commonly the case when writing emails*;* or *intent classification,* which is used by chatbots to identify the intention of users with their texts; and also *text summarization* [6]*.*

With one of its most used and complete libraries, the *Natural Language Toolkit (NLTK),* we can prepare the data in which they are structured to be processed later. Firstly, tokenization is performed (splitting the sentence into words), lower casing them, removing stop words (non-important words like a, an, the, etc.), stemming them (transforming words to their root form, eg changing to change) or lemmatization (to a word existing in the language, changing to change).

2.3 Related works

The academic publications on RS to date, we must highlight the contributions to hybrid or multi-criteria RS [4, 5, 7, 8, 9, 10]. In addition, only one of those named does not talk about data from TripAdvisor.com [4]. This is the type of RS preferred by platforms that have item-user logic, and the one that is proposed to be applied to add improvements to the platform or to extract knowledge from it in terms of NLP. Which demonstrates the impact of TripAdvisor platform on Data Science related to the possibilities that it brings.

2.3.1 Achievements of the latest research

Specifically on the application of NLP on reviews, without dealing with RS, the research by Ali et al, 2021, is highlighted, dealing with the "Negative e-reputation using aspect-based Sentiment Analysis approach" in Marrakech. 39,216 TripAdvisor reviews are treated with NLP and Topic Modeling (extraction of main topics in the text) is performed using Latent Dirichlet Allocation (LDA), a technique that treats the similarity of the terms according to their positioning and considering several dimensions in the document, where each topic represents an et of words. Thus they get "extract hidden aspects and dimensions from feedback" which can be interpreted to improve tourism in this city [3]

Returning to investigations that culminate their NLP process with RS, in the investigation by Al-Ghuribi & Mohd Noah of 2019 on “Multi-Criteria Review-Based Recommender System–The State of the Art” criticises that, to date, “most of the RS approaches rely on a single-criterion” and provide a solution to RS when there is a lack of data of the users. That is, when a user gives a rating on cleanliness and location, the reviews are used to complete this information (eg the user makes a comment about the staff and wifi), and then this is processed by Hybrid RS that takes into account the item-user logics named in previous sections [5]. On the other hand, Logesh & Subramaniyaswamy published 2019 "Exploring Hybrid Recommender System for Personalized Travel Applications" where they propose a *Personalised Context-aware Hybrid Travel Recommender System* (PCAHTRS), which goes one step further in obtaining user data and wants to take into account contextual data of the user about their mobility and temporality through their mobile data during their trips (Figure 5). With this, the traditional problems of cold start and sparsity would be overcome while prediction accuracy is improved [7].

.

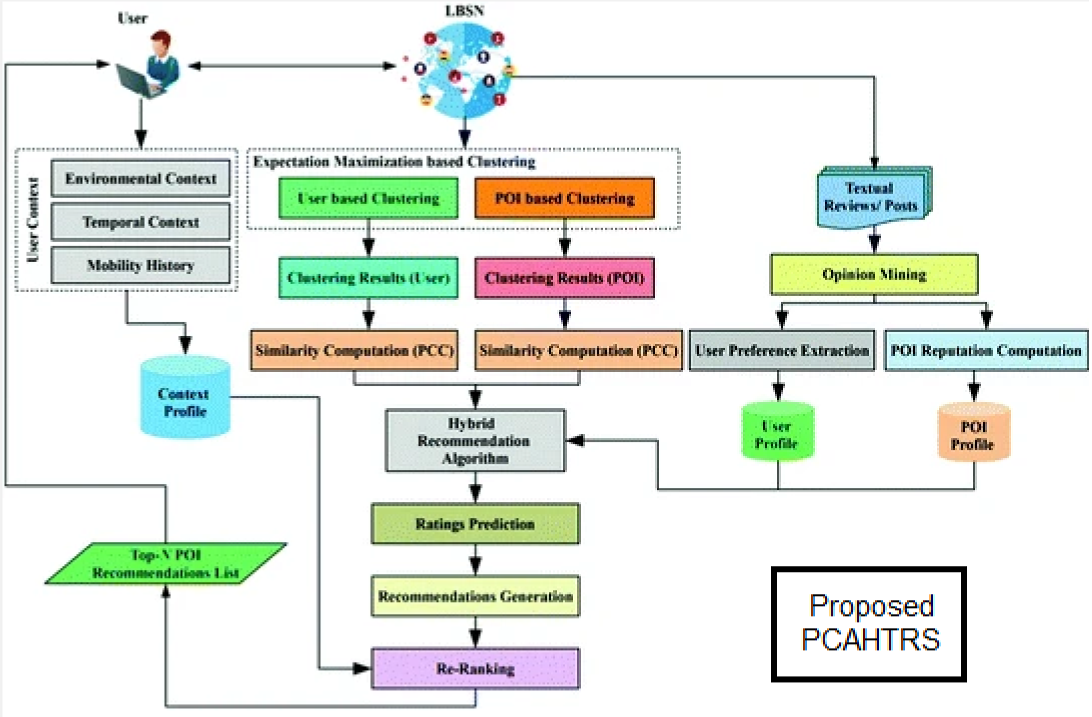

User
LBSN
User Context
Environmental Context
Temporal Context
Mobility History
Expectation Maximization based Clustering
User based Clustering
POI based Clustering
Textual Reviews/ Posts
Clustering Results (User)
Clustering Results (POI)
Opinion Mining
Similarity Computation (PCC)
Similarity Computation (PCC)
User Preference Extraction
POI Reputation Computation
Context Profile
Hybrid Recommendation Algorithm
User Profile
POI Profile
Ratings Prediction
Top-N POI Recommendations List
Recommendations Generation
Re-Ranking
Proposed PCAHTRS

Figure 4. Source: Logesh & Subramaniyaswamy (2019)

One of the most relevant investigations for this work is that of Nadezhda (2020), "Recommendation System for Travelers Based on TripAdvisor.com Data". The author uses LDA to extract topics and with them configures a new classification of types of Activities. TripAdvisor.com distinguishes, at the time of the aforementioned investigation, 51 types of activities in the city of London (e.g., Museums, Fun & Games, Nature & Parks) and the author reduces them to 6 (Art, Food, Nature, Performing arts, Landmarks and Tours). Another achievement of this research is the generation of a hybrid RS, which is also developed in an interactive web-based service application [8].

Although the LDA technique is widely used in Topic Modeling [3 and 8], there are critical investigations with it that propose better models. This is the case of Arenas-Márquez et al. (2021), in their publication "Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor" conclude that "the paper demonstrates that the neural encoding fitted as part of a classifier maximizes the discrimination of documents when compared to her encoding schemes" such as LDA [9].

23.2 Transformative Techniques

In 2017, Google's Brain Team (artificial intelligence and deep learning) introduced Transformers, which just like Recurrent Neural Networks (RNN) are capable of processing languages, but unlike RNN, Transformers process the entire input at once. Transformers have made great strides in NLP, interpreting languages significantly better and making the process cheaper. The transformers have a set of pre-trained data or embeddings (vectorisation techniques) to which fine-tuning is performed (this learning is applied to a new text). Added to this are the Transformers from the Huggingface group (https://huggingface.co/) that further reduce the cost of the transformers by training them with half the words and maintaining 97% performance, in words of the opium research group "reduces the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster" [10].

In Zhuang, Y. (2021), "A BERT-Based Multi-Criteria Recommender System for Hotel Management", the author highlights the limitations of current SA methods. Therefore, it proposes a process that uses BERT to compute customer aspect ratings and overall ratings. Although the SA is more accurate, the results of the RS do not improve [11].

3. Objectives

Academic: develop an end-to-end Machine Learning (ML) project using text extracted from TripAdvisor, including data harvesting, the use of NLP techniques for feature extraction & language understanding, as well as conducting further analysis of the results to finally consolidate the findings in charts and reports.

- Q1. What are the possibilities of scrapping data from TripAdvisor.com using open-source methods?
- Q2. Can we develop NLP models with state-of-the-art techniques and using Transformers with this data?

Business**:** find general usage trends in touristic language on the Canary Islands on TripAdvisor.com (real-world data), to propose a post-pandemic sensitive marketing strategy, creating a new feature space derived from social data.

- Q3. Can we recommend new content/Attractions based onTripAdvisor.com data?
- Q4. What are the topics prioritized and the performance of reviews in the last 2 years in this context?

4. Solution

The techniques chosen to achieve the set objectives will be described below. Noting the reasons why these have been chosen and describing the different phases that the project has had. The phases of this methodology are represented in Figure 5. The process is divided into 6 Notebooks (Figure 6).

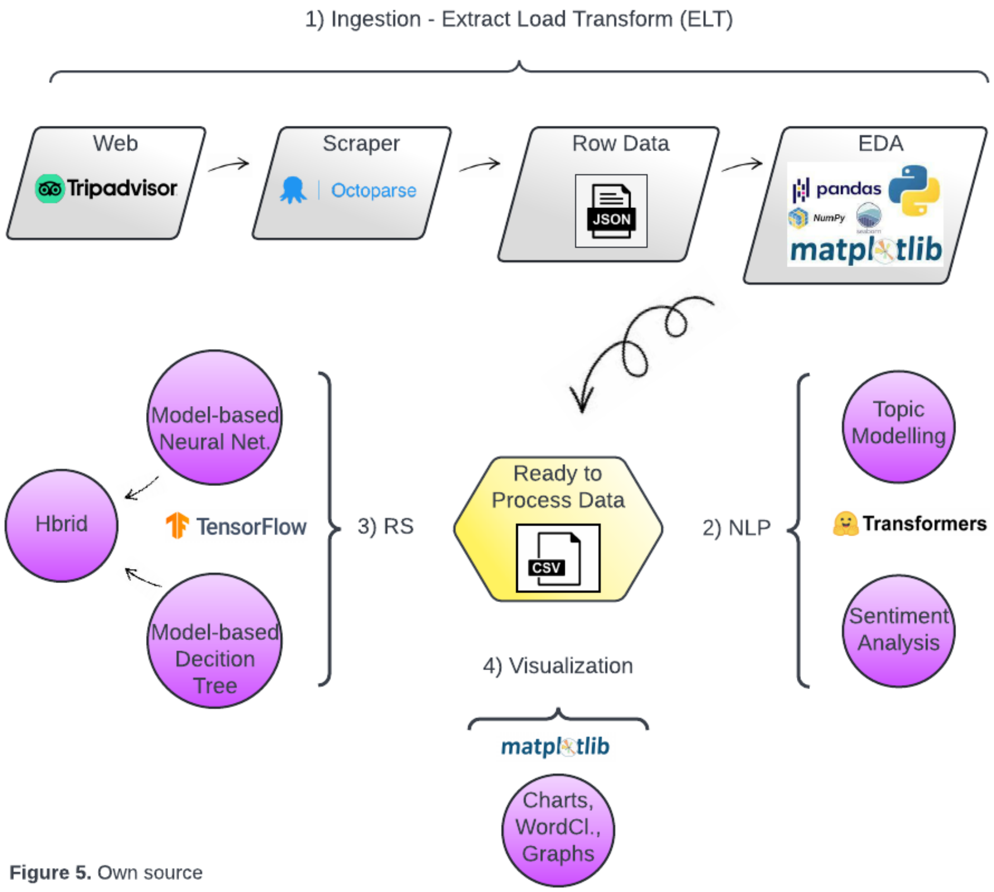
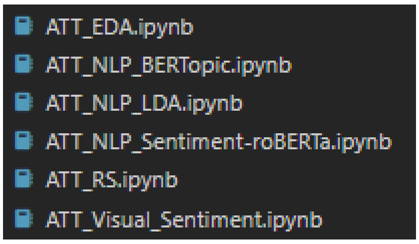


Figure 5. Own source

Steps 1-4 will be written in python on a Jupyter Notebook using Visual Studio Code's IDE. The Project was largely developed using an Intel Core 7 CPU (2.8 GHz, RAM 16GB) Asus machine. The compute-intensive workload was handled by using Google Colab's Runtime Engine through the ColabCode web, thereby leveraging their Tensor Processing Unit (TPU) or Graphic Processing Unit (GPU) processors.

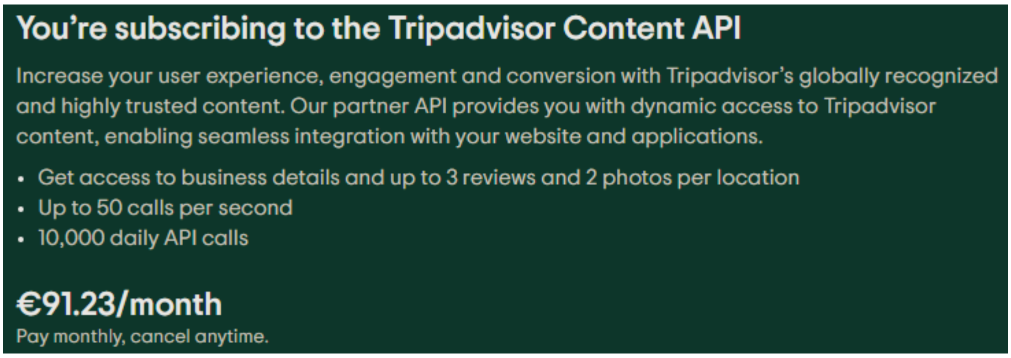


```
ATT_EDA.ipynb
ATT_NLP_BERTopic.ipynb
ATT_NLP_LDA.ipynb
ATT_NLP_Sentiment-roBERTa.ipynb
ATT_RS.ipynb
ATT_Visual_Sentiment.ipynb
```

Figure 6. Jupyter Notebooks, Own source

4.1 Phase 1: TripAdvisor data ingestion

TripAdvisor.com presents all the necessary conditions to carry out NLP and RS. It has complex access to data, public data is rarely found, and it has an API for businesses where data can be accessed for an approximate cost of €90 per month.



You're subscribing to the Tripadvisor Content API

Increase your user experience, engagement and conversion with Tripadvisor's globally recognized and highly trusted content. Our partner API provides you with dynamic access to Tripadvisor content, enabling seamless integration with your website and applications.

- Get access to business details and up to 3 reviews and 2 photos per location
- Up to 50 calls per second
- 10,000 daily API calls

€91.23/month

Pay monthly, cancel anytime.

Figure 7. Source: https://www.tripadvisor.com/developerscheckout

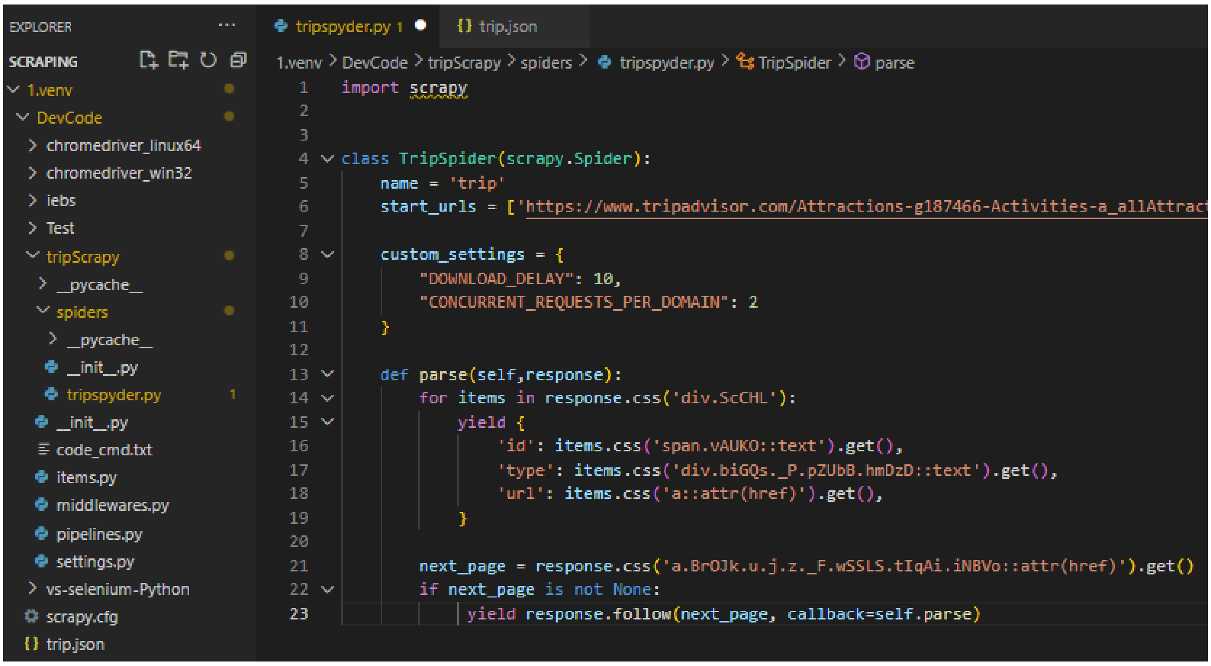
4.1.1 Choice of sample

Given the reasons mentioned in the previous point 2.1.1, it seems pertinent to this study to concentrate its sample in the Canary Islands of Spain. Which have a great diversity of reliefs, divided into 8 different territories/islands: La Palma, El Hierro, La Gomera, Tenerife, Gran Canaria, Fuerteventura, Lanzarote and La Graciosa. In addition to being a point of mass tourism for sun and gastronomic and nocturnal consumption for several decades, the Canary Islands of eternal spring present countless unique corners of nature, cultural activities, history and gastronomy. And, following what was commented on in 2.1.2, specifically on the use of data science for the sustainable management of tourism and travel, this work decides to investigate the information that the TripAdvisor portal can provide about this popular European tourist destination, with an economy built around tourism, it could certainly take advantage of improvements in tourism management, after the reopening of the borders.

Most of the academic works published on TripAdvisor focus on Accommodations/hotels as the items to be processed in their RS and from which to obtain knowledge, with an obvious commercial reason, only the work of Nadezhda (2020) deals with Activities, which also have a commercial interest since they allude to tours, restaurants, casinos and other businesses. For this reason, this work intends to differentiate itself by treating only Attractions, which have free admission or are characteristic to the point of defining the area where they are located (we are referring, eg, to landmarks, museums, or internationally recognized water parks).

4.1.2 Extraction processes:

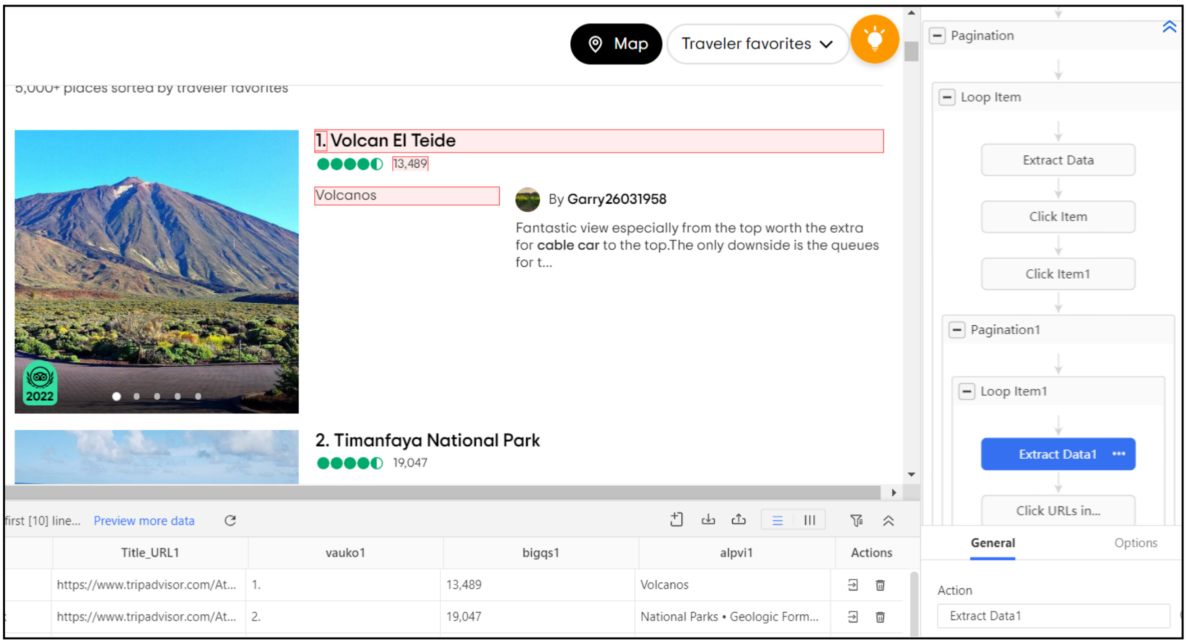
Faced with the challenge of obtaining data and the objective of dealing with real problems (Q3). In the first place, this project developed a local virtual environment for scraping, and interpreting the HTML of TripAdvisor.com, with spiders that would execute the collection of thousands of reviews within hundreds of Attractions. It was also tested with the *selenium* and *beautifulsoup* application *chromedriver*:



```python
import scrapy


class TripSpider(scrapy.Spider):
    name = 'trip'
    start_urls = ['https://www.tripadvisor.com/Attractions-g187466-Activities-a_allAttract

    custom_settings = {
        "DOWNLOAD_DELAY": 10,
        "CONCURRENT_REQUESTS_PER_DOMAIN": 2
    }

    def parse(self,response):
        for items in response.css('div.ScCHL'):
            yield {
                'id': items.css('span.vAUKO::text').get(),
                'type': items.css('div.biGQs._P.pZUbB.hmDzD::text').get(),
                'url': items.css('a::attr(href)').get(),
            }

        next_page = response.css('a.BrOJk.u.j.z._F.wSSLS.tIqAi.iNBVo::attr(href)').get()
        if next_page is not None:
            yield response.follow(next_page, callback=self.parse)
```

Figure 8. Scrapy.Spider, Own source

Second, the Octoparse API was used:



1. Volcan El Teide
●●●●◐ 13,489
Volcanos

By Garry26031958

Fantastic view especially from the top worth the extra for **cable car** to the top.The only downside is the queues for t...

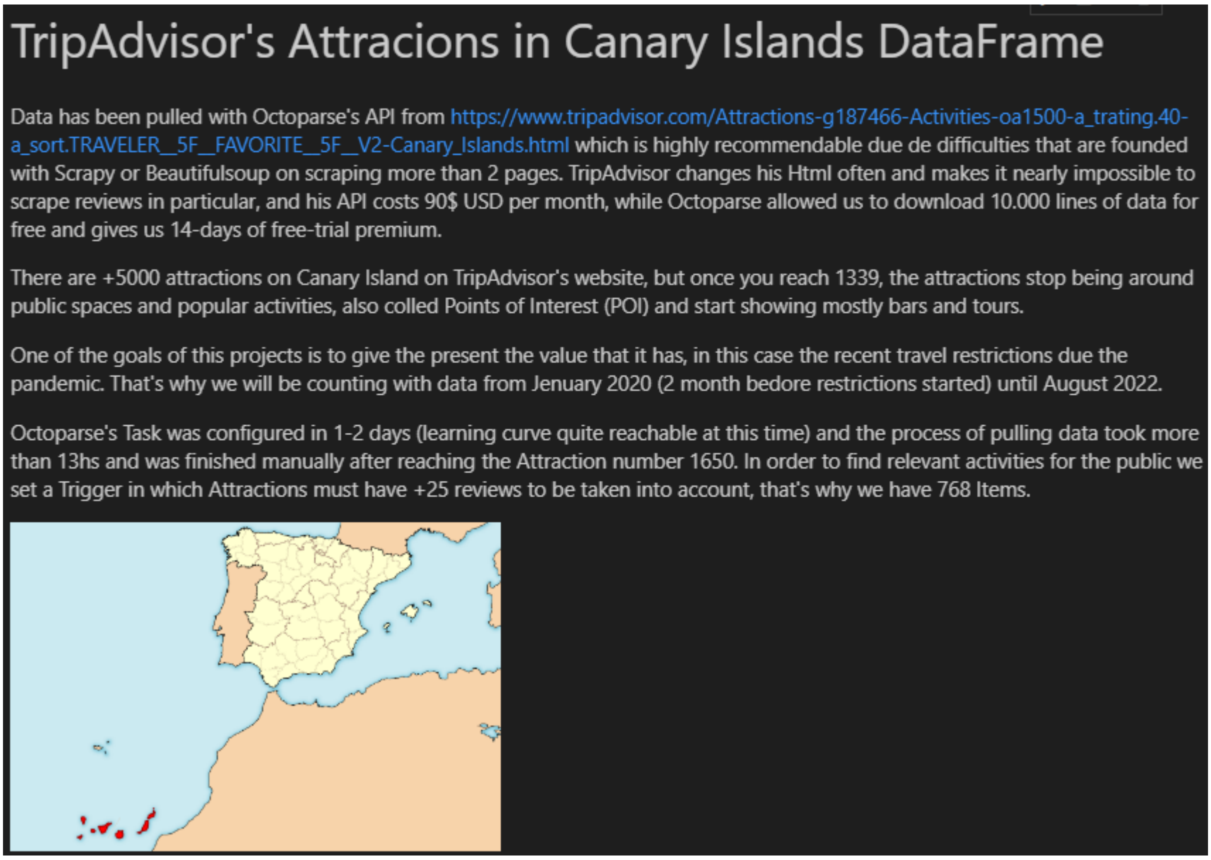
2. Timanfaya National Park
●●●●◐ 19,047

Title_URL1	vauko1	bigqs1	alpvi1	Actions
https://www.tripadvisor.com/At...	1.	13,489	Volcanos	
https://www.tripadvisor.com/At...	2.	19,047	National Parks • Geologic Form...	

Figure 9. Octoparse scraping, Own source

4.1.3 Exploratory Data Analysis and cleaning

Once the data was obtained in a JSON format, the data was processed by the *pandas, NumPy and Matplotlib libraries* where a large number of corrections were made, since the row data required it in 75 lines of code (“ATT_EDA.ipynb”).



TripAdvisor's Attracions in Canary Islands DataFrame

Data has been pulled with Octoparse's API from https://www.tripadvisor.com/Attractions-g187466-Activities-oa1500-a_trating.40-a_sort.TRAVELER__5F__FAVORITE__5F__V2-Canary_Islands.html which is highly recommendable due de difficulties that are founded with Scrapy or Beautifulsoup on scraping more than 2 pages. TripAdvisor changes his Html often and makes it nearly impossible to scrape reviews in particular, and his API costs 90$ USD per month, while Octoparse allowed us to download 10.000 lines of data for free and gives us 14-days of free-trial premium.

There are +5000 attractions on Canary Island on TripAdvisor's website, but once you reach 1339, the attractions stop being around public spaces and popular activities, also colled Points of Interest (POI) and start showing mostly bars and tours.

One of the goals of this projects is to give the present the value that it has, in this case the recent travel restrictions due the pandemic. That's why we will be counting with data from Jenuary 2020 (2 month bedore restrictions started) until August 2022.

Octoparse's Task was configured in 1-2 days (learning curve quite reachable at this time) and the process of pulling data took more than 13hs and was finished manually after reaching the Attraction number 1650. In order to find relevant activities for the public we set a Trigger in which Attractions must have +25 reviews to be taken into account, that's why we have 768 Items.

Figure7. ATT_EDA.ipynb, Own source

It was in this notebook where corrections of formats, typos, data in mixed columns, the combination of columns, elimination or creation of new ones, and multiple fixes were treated. Also many visualisations that completed an exhaustive cleaning.

Finally, 9,860 records will be taken into account (each with a review), from more than 6,000 different users, on more than 650 different Attractions/items categorized into approximately 150 types.

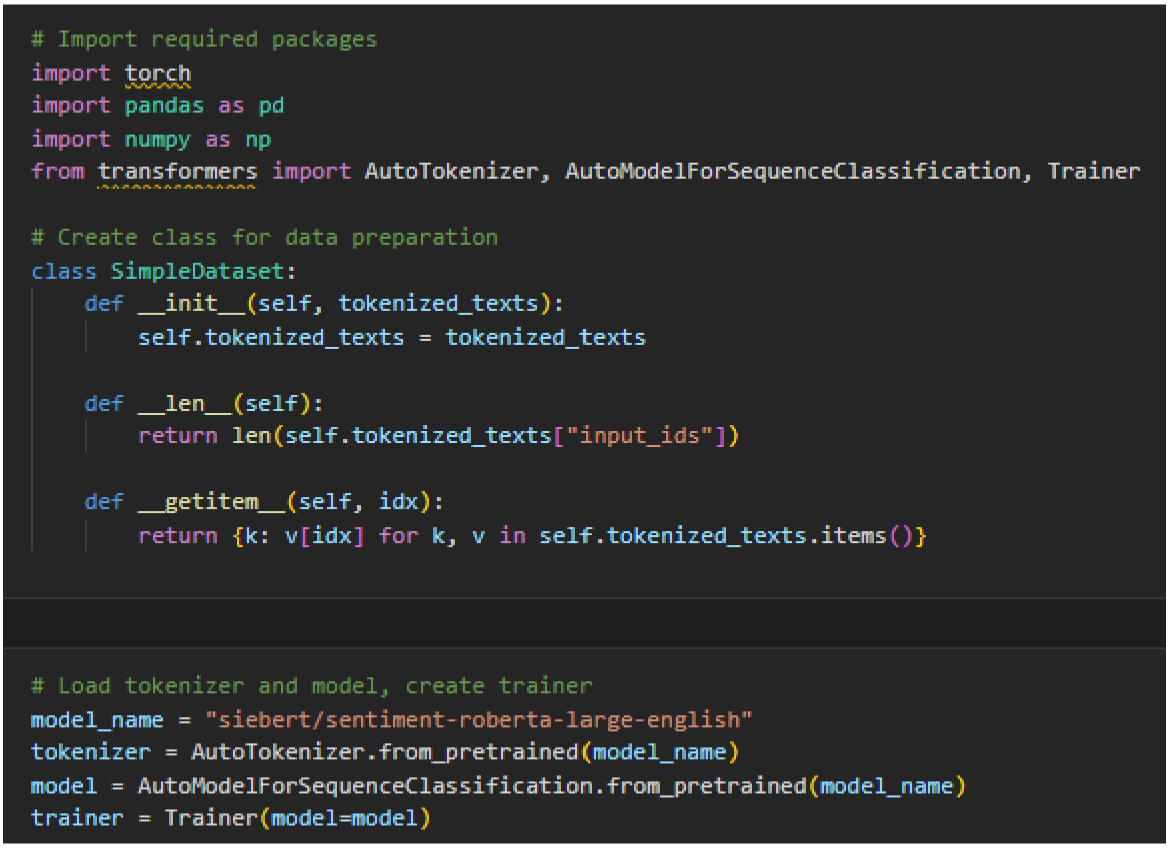
4.2 Phase 2: Natural Language Processing

Although the SA reviews included in RS have not been successful in previous research, it does not mean that it is not of great importance in understanding the data. The ability to extract Topics, or to visualize that Sentiment provides perspective and tools for improvement and proposal of solutions. Therefore, as was marked in its objectives, this work intends to apply Transformers, among other techniques.

Techniques without recurrence for this project such as Named Entity Recognition or Keyphrase Extraction have been ruled out. And SA with roBERTa, TensorFlow pretrained embedding created by the Hugging Face group with its origin in the aforementioned Distill-BERT[2] has been applied. Also, this research proposes to work with BERTopic for Topic Modeling and compare it with LDA since it has been used in several investigations. The *NLKT* package and its SentimentIntensityAnalyzer library will be used for text preprocessing and subsequent visualization (phase 4) of SA with *Matplotlib* and *seaborn*.

[1] https://huggingface.co/siebert/sentiment-roberta-large-english

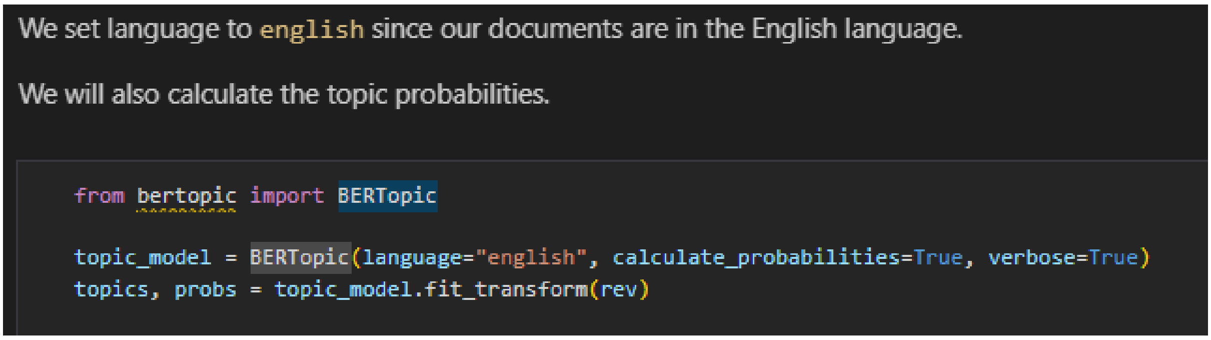
two https://huggingface.co/docs/transformers/model_doc/distilbert



```
# Import required packages
import torch
import pandas as pd
import numpy as np
from transformers import AutoTokenizer, AutoModelForSequenceClassification, Trainer

# Create class for data preparation
class SimpleDataset:
    def __init__(self, tokenized_texts):
        self.tokenized_texts = tokenized_texts

    def __len__(self):
        return len(self.tokenized_texts["input_ids"])

    def __getitem__(self, idx):
        return {k: v[idx] for k, v in self.tokenized_texts.items()}
```

```
# Load tokenizer and model, create trainer
model_name = "siebert/sentiment-roberta-large-english"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name)
trainer = Trainer(model=model)
```

Figure 8. ATT_Sentiment-roberta-prediction.ipynb, Own source

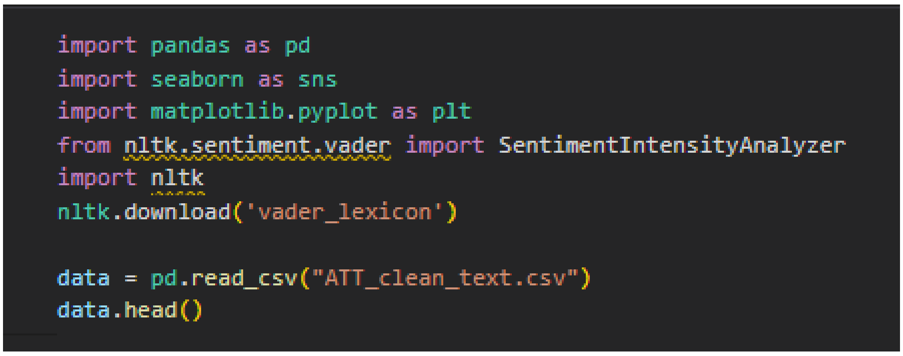


We set language to `english` since our documents are in the English language.

We will also calculate the topic probabilities.

```
from bertopic import BERTopic

topic_model = BERTopic(language="english", calculate_probabilities=True, verbose=True)
topics, probs = topic_model.fit_transform(rev)
```

Figure 9. ATT_BERTopic.ipynb, Own source



```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import nltk
nltk.download('vader_lexicon')

data = pd.read_csv("ATT_clean_text.csv")
data.head()
```

Figure 10. ATT_Sentiment_visul.ipynb, Own source

4.3 Phase 3: Recommender Systems

In this last phase, the research proposes the creation of three RS: the first two will be Collaborative Filtering and model-based, since they are the ones that give the best results for this type of problem according to the consulted investigations. The intention of this project is to develop a multi-criteria RS by also creating a Hybrid RS, which will consist of a combination of the two previous RS. In addition, data visualisations (phase 4) are also performed here.

Research reviewed in point 2 suggests that the inclusion of SA results in user ratings does not create an improvement in RS [11]. Therefore, this technique has been discarded.

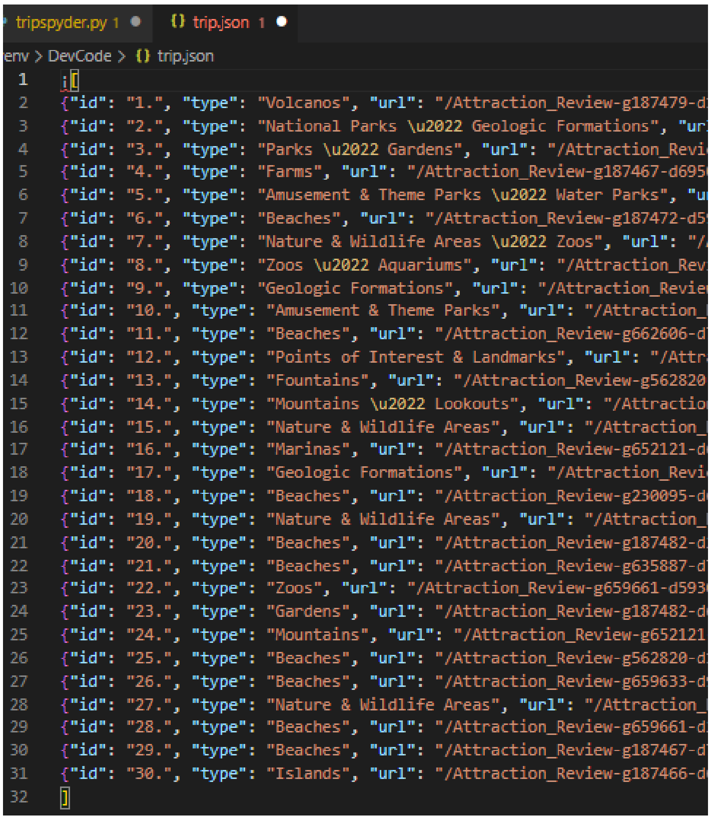
To work with RS, 3 Dataframes will be created: user_df, item_df and rating_df. In order to make use of the categorical variables of Item_df, they will be processed by One Hote Encoding of the *Scikit-learn* library. After preparing the data, the two RS will be developed, one based on Neural Networks (NN) + Nearest Neighbors and the other on Decision Tree + OneHotEncoder + XGBoost_recommender.

5. Evaluation

This section proceeds to describe the interpretations of the proposed methodology, the drawbacks found in the process, and the ways in which the results are reached.

5.1 Evaluation of the data intake process

First of all, the research found the difficulties mentioned above in terms of data collection. In the scraping process with *Scrapy*, with the previous creation of a virtual environment from the Terminal and executing the script from it (Figure 8), the code was not able to overcome more than two data extraction pages, entering a loop from the second. And gathering only 30 lines:



```
[
{"id": "1.", "type": "Volcanos", "url": "/Attraction_Review-g187479-d
{"id": "2.", "type": "National Parks \u2022 Geologic Formations", "ur
{"id": "3.", "type": "Parks \u2022 Gardens", "url": "/Attraction_Revi
{"id": "4.", "type": "Farms", "url": "/Attraction_Review-g187467-d695
{"id": "5.", "type": "Amusement & Theme Parks \u2022 Water Parks", "u
{"id": "6.", "type": "Beaches", "url": "/Attraction_Review-g187472-d5
{"id": "7.", "type": "Nature & Wildlife Areas \u2022 Zoos", "url": "/
{"id": "8.", "type": "Zoos \u2022 Aquariums", "url": "/Attraction_Rev
{"id": "9.", "type": "Geologic Formations", "url": "/Attraction_Revie
{"id": "10.", "type": "Amusement & Theme Parks", "url": "/Attraction_
{"id": "11.", "type": "Beaches", "url": "/Attraction_Review-g662606-d
{"id": "12.", "type": "Points of Interest & Landmarks", "url": "/Attr
{"id": "13.", "type": "Fountains", "url": "/Attraction_Review-g562820
{"id": "14.", "type": "Mountains \u2022 Lookouts", "url": "/Attractio
{"id": "15.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
{"id": "16.", "type": "Marinas", "url": "/Attraction_Review-g652121-d
{"id": "17.", "type": "Geologic Formations", "url": "/Attraction_Revi
{"id": "18.", "type": "Beaches", "url": "/Attraction_Review-g230095-d
{"id": "19.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
{"id": "20.", "type": "Beaches", "url": "/Attraction_Review-g187482-d
{"id": "21.", "type": "Beaches", "url": "/Attraction_Review-g635887-d
{"id": "22.", "type": "Zoos", "url": "/Attraction_Review-g659661-d593
{"id": "23.", "type": "Gardens", "url": "/Attraction_Review-g187482-d
{"id": "24.", "type": "Mountains", "url": "/Attraction_Review-g652121
{"id": "25.", "type": "Beaches", "url": "/Attraction_Review-g562820-d
{"id": "26.", "type": "Beaches", "url": "/Attraction_Review-g659633-d
{"id": "27.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
{"id": "28.", "type": "Beaches", "url": "/Attraction_Review-g659661-d
{"id": "29.", "type": "Beaches", "url": "/Attraction_Review-g187467-d
{"id": "30.", "type": "Islands", "url": "/Attraction_Review-g187466-d
]
```

Figure 11. First JSON, Own source

The construction of the virtual environment did not mean difficulty, but the attempts to create a Spider capable of collecting the data did. The investment of time in this part of the project, together with the fact that another solution had to be found, will be part of the learning process and meant a delay in the dates set.

On the other hand, as already explained in the notebook displayed in Figure 7 (EDA_ATT.ipynb), using the Octoparse API it was possible to speed up this process, not without some difficulty, and with a high need to clean the data.

5.2 Evaluation of the NLP process

The NLP has been divided into three parts, Topic Modelling with LDA, BERTopic and SA (roBERTa). All parties have had a satisfactory process, without problems in processing, following the established packages and libraries. Visualizations have also been made throughout the processing as a way of understanding the data, which adheres to phase 4.

In the case of LDA, this study has been able to see the potential of this technique, using the LDAvis library for visualization. visualization of the results, which is interactive, being able to select the different Topics bubbles for in-depth analysis.

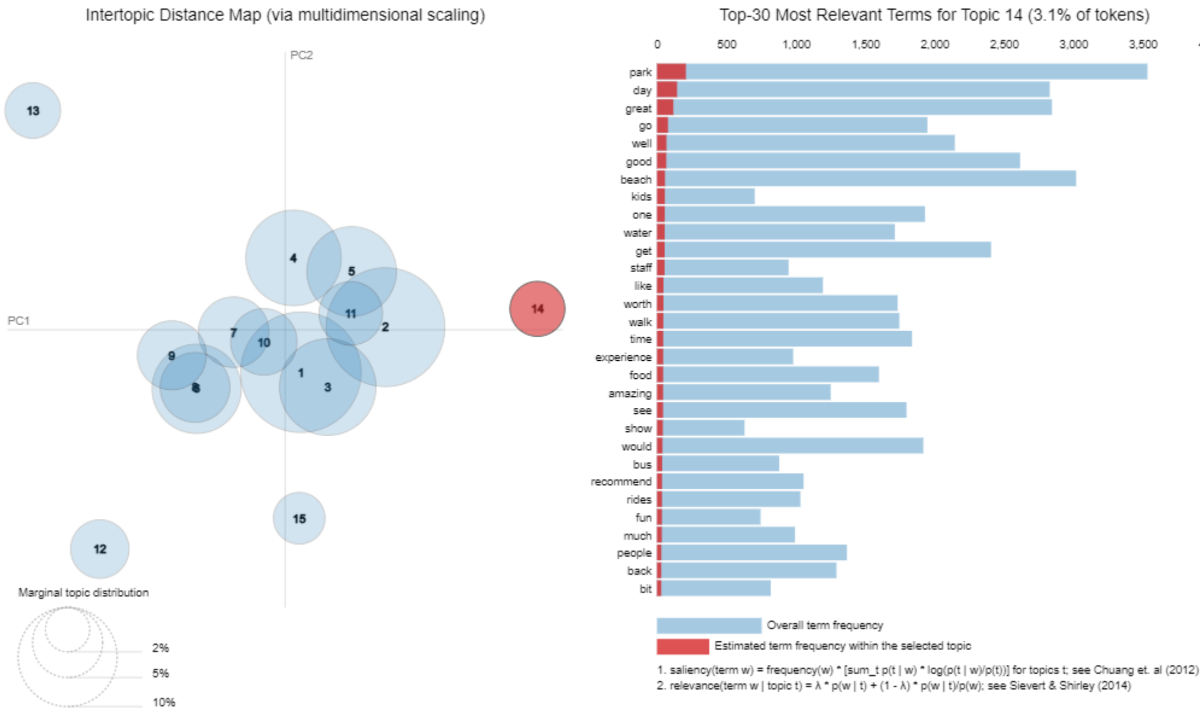
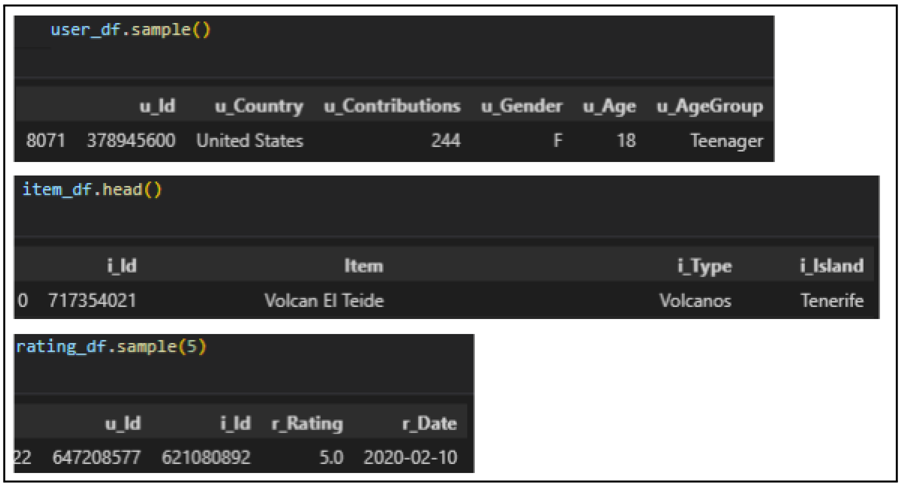


Figure 11. LDA visualization. Own source

These visualizations have allowed us, as in previous investigations [8], the possibility of differentiating some main Topics, with *Parks* and *Beaches* being the most referenced and important contents in this superficial analysis.

5.3 Evaluation of the RS

The creation of the RS is based on the formation of the three Dataframes (user_df, item_df, rating_df). Since the data extraction and subsequent transformation for its use were successful, there was no problem in forming them. Next, visualizations were made to understand the data.



```
user_df.sample()
```

	u_Id	u_Country	u_Contributions	u_Gender	u_Age	u_AgeGroup
8071	378945600	United States	244	F	18	Teenager

```
item_df.head()
```

	i_Id	Item	i_Type	i_Island
0	717354021	Volcan El Teide	Volcanos	Tenerife

```
rating_df.sample(5)
```

	u_Id	i_Id	r_Rating	r_Date
22	647208577	621080892	5.0	2020-02-10

Figure 12. User_df, item_df, rating_df, Own source

The validation of the first RS has been carried out with BinaryCrossentropy(), giving a loss of 0.36 in training and 0.59 in validation, in the best of cases, given that in each execution are obtained different results. Here, a need to process more data is distinguished, for better training and an RS with less variation in each execution. At the same time, the validation of XGBoost was intended to be carried out with the RMSE (Root mean square error), which was not possible due to errors in the scripting. Confirming the need for improvements in this process and in the adaptation of the data.

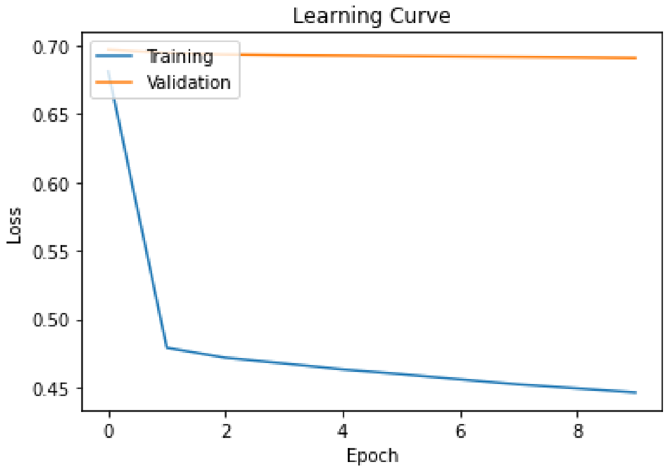


Figure 13. Neural Network Learning Curve, Own source

5.4 Evaluation of data visualization

The visualization process was effective in all phases, a multitude of graphs, charts and Wordclouds were displayed that invited the understanding of the data and their correction from phase 1 (ATT_EDA.ipynb).

In the EDA we were able to distinguish particularities of our data, such as the polarity in the data of the number of contributions that a user has made on TripAdvisor (total number of reviews), having a large bulk with dozens of reviews, but some users with 5,000, 10,000 or up to 60,000. We also differentiate the proportion of ratings that the total items of our dataset have.

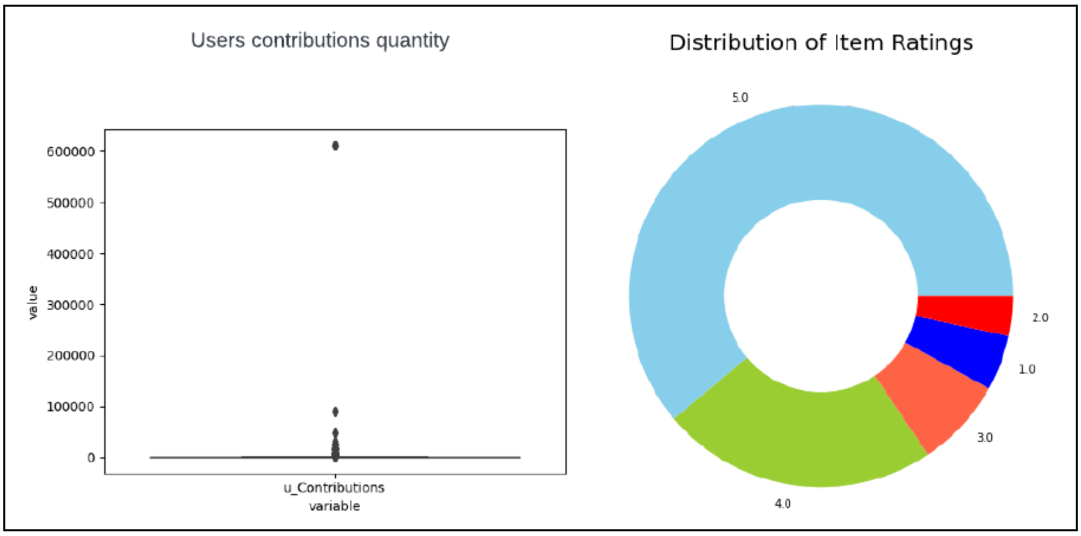


Figure 14. EDA visual evaluation, Own source

In the NLP phase, different visualizations were extracted. In the Topic Modeling processing, in addition to a similarity matrix between items (Appendix 2), the following Wordcloud was displayed that shows a relationship between the importance and recurrence of the use of words in reviews according to sizes and colours:

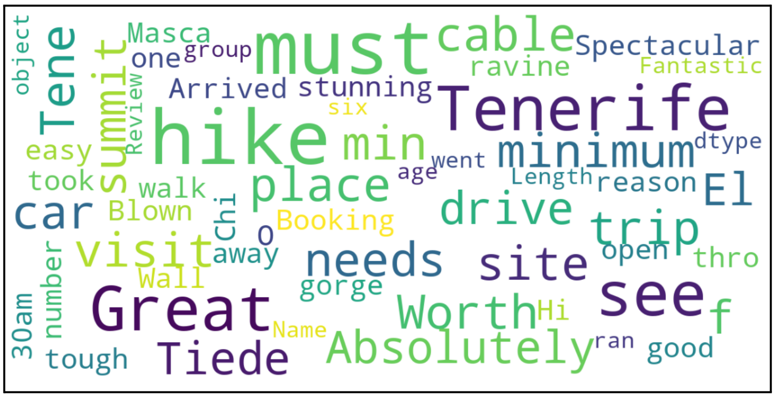


Figure 15. BERTopic visual evaluation, Own source

Also in the NLP processing, specifically SA with roBERTa, we were able to evaluate the density of the reviews, in a number of characters, depending on the rating that the users gave to the Attractions/items.

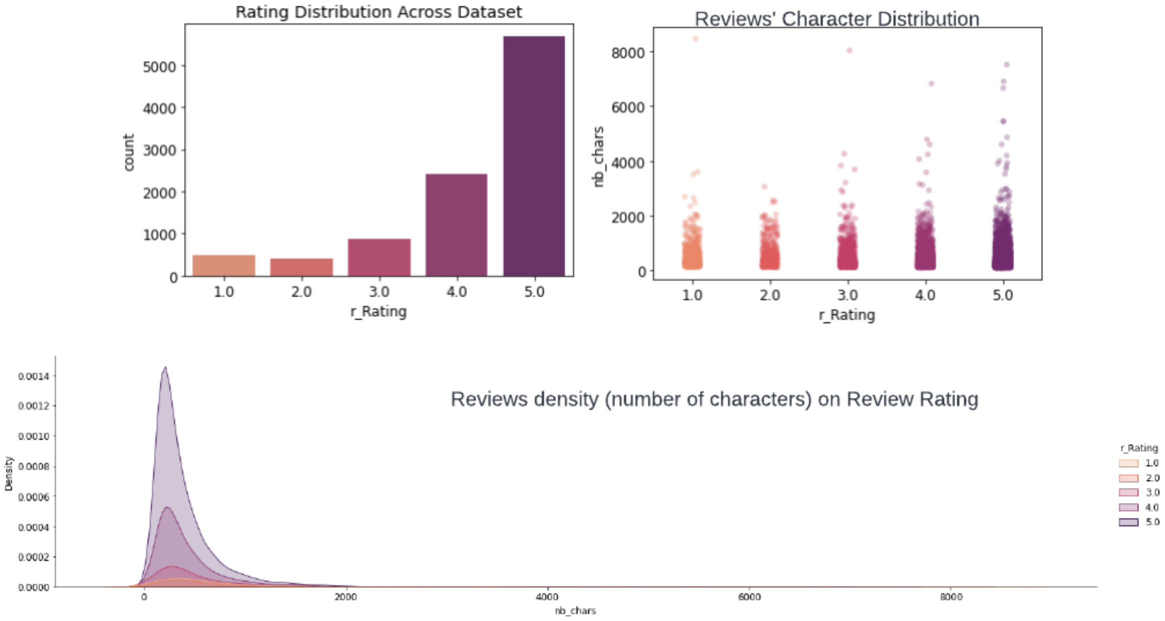


Figure 16. roBERTa visual evaluation, Own source

Finally, with the Notebook ATT_Visual_Sentiment.ipynb, dedicated specifically to printing visualizations relevant to research, we were able to interpret variables such as the *type of trip-u_Trip* (alone, accompanied and with what way, or unknown) as well as the percentages of users by age group. It stands out that, of the users who have this information, the majority travel in pairs or in a group of friends, and there is a small percentage that travel alone (and make reviews on Tripadvisor). A bar graph is also displayed in which the items with the most reviews are differentiated (Appendix 3): beaches, Landmarks, Museums, Malls and Parks lead the list.

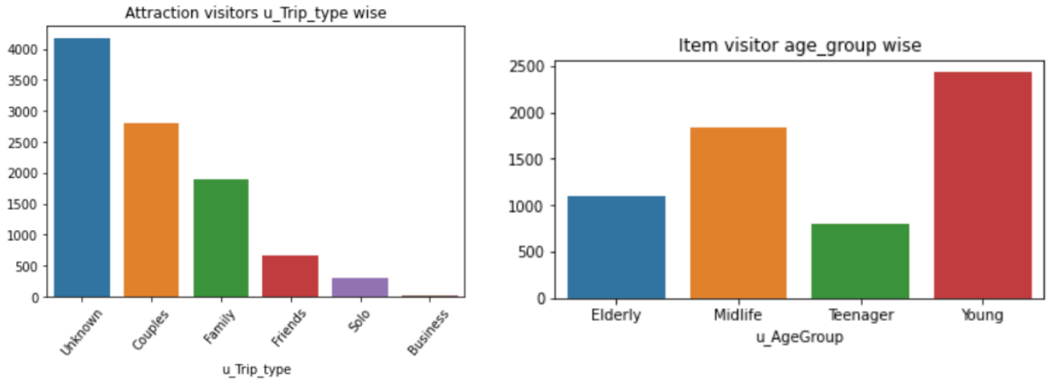
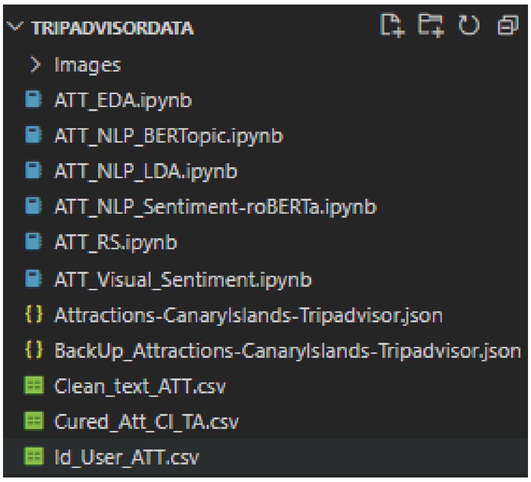


Figure 17. Visual_Sentiment, Own source

6. Results

In this chapter, the results obtained with the different techniques used will be analysed. It is aimed to explain the achievement or not of the objectives and the discoveries of the work. The result of all the explorations carried out is displayed in 2 JSON with the original data obtained by scraping with Octoparse, 6 notebooks with the Python code, and 3 CSV files with the Dataframes clean and ready to be processed. This research has published on the GitHub portal [13] so that others can take advantage of the data collected and the techniques used, and thus contribute to the community of Data Scientists and the treatment of Big Data.



```
TRIPADVISORDATA
  > Images
  ATT_EDA.ipynb
  ATT_NLP_BERTopic.ipynb
  ATT_NLP_LDA.ipynb
  ATT_NLP_Sentiment-roBERTa.ipynb
  ATT_RS.ipynb
  ATT_Visual_Sentiment.ipynb
  {} Attractions-CanaryIslands-Tripadvisor.json
  {} BackUp_Attractions-CanaryIslands-Tripadvisor.json
  Clean_text_ATT.csv
  Cured_Att_CI_TA.csv
  Id_User_ATT.csv
```

Figure 18. Investigation files, Own source

This research started with an academic objective oriented towards the ingestion of data with open source tools and its treatment with NLP, and with another business-oriented one based on the search for opportunities through the creation of RS and the interpretation of the data behaviour from the beginning of the Covid-19 pandemic to the present day. The following sections respond to the four questions that this research raised within its objectives.

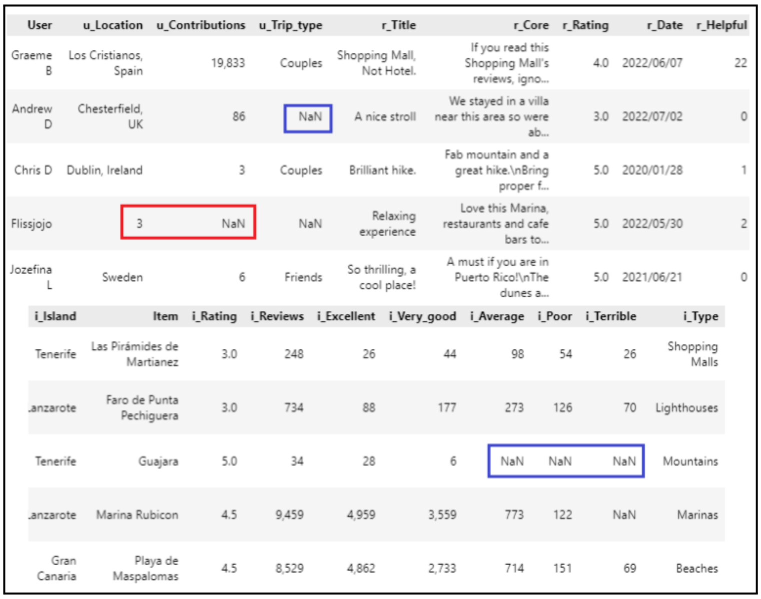
- Q1. What are the possibilities of scrapping data from TripAdvisor.com using open-source methods?
- Q2. Can we use this data and develop NLP models with state-of-the-art techniques and using Transformers with this data?
- Q3. Can we recommend new content/Attractions based on TripAdvisor.com data?
- Q4. What are the topics prioritised and the performance of reviews in the last 2 years in this context?

6.1 Octoparse as the scrapping solution (Q1)

TripAdvisor presented severe difficulties in obtaining data with open-source tools. Although with the *scrapy* to obtain precise and comfortable results on most websites, in this case, it was not possible to obtain the data in this way. Neither with *beautifulsup* and *selenium*. What was effective, and with a short learning curve, was the Octoparse tool. With this application, the user can teach the spider where to click and where to collect data without having to interpret HTML for much of the process. The data collection can be done on a local machine (as long as there are less than 10,000 registrations), or in the Octoparse cloud itself with a Premium account (unlimited registrations).

Data collection was stopped manually after 13 hours, as around 1,650 Attractions had been processed (with more than 25 reviews, for reliability reasons). Thus, 12,460 records were collected on 768 items/Attractions. Each record had information on Users (u_), Reviews (r_) and Items (i_). This process is defined as ELT (Extract, Load, Transform), since there is no streaming data input (adding records daily), so there was no need to create a database. They were Loaded in Local and then Transformed in Jupyter Notebooks with Python.

As is visible in Figure 19, there were imperfections in the data, with several columns showing NaN values, and others showing the information in the wrong column. At first glance, through the Exploratory Data Analysis (EDA) other variables were found that needed multiple corrections.

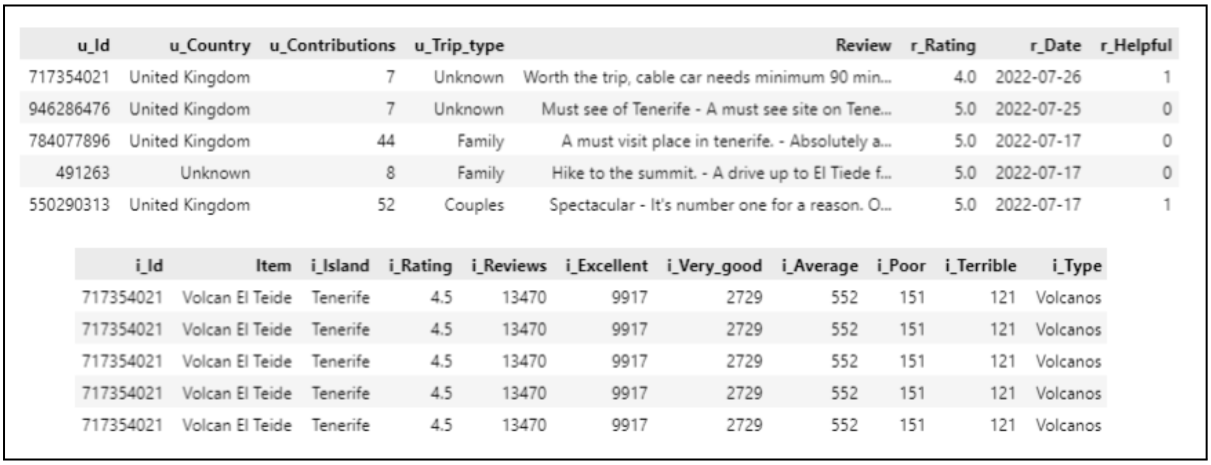


User	u_Location	u_Contributions	u_Trip_type	r_Title	r_Core	r_Rating	r_Date	r_Helpful
Graeme B	Los Cristianos, Spain	19,833	Couples	Shopping Mall, Not Hotel.	If you read this Shopping Mall's reviews, igno...	4.0	2022/06/07	22
Andrew D	Chesterfield, UK	86	NaN	A nice stroll	We stayed in a villa near this area so were ab...	3.0	2022/07/02	0
Chris D	Dublin, Ireland	3	Couples	Brilliant hike.	Fab mountain and a great hike.\nBring proper f...	5.0	2020/01/28	1
Flissjojo	3	NaN	NaN	Relaxing experience	Love this Marina, restaurants and cafe bars to...	5.0	2022/05/30	2
Jozefina L	Sweden	6	Friends	So thrilling, a cool place!	A must if you are in Puerto Rico!\nThe dunes a...	5.0	2021/06/21	0

i_Island	Item	i_Rating	i_Reviews	i_Excellent	i_Very_good	i_Average	i_Poor	i_Terrible	i_Type
Tenerife	Las Pirámides de Martianez	3.0	248	26	44	98	54	26	Shopping Malls
anzarote	Faro de Punta Pechiguera	3.0	734	88	177	273	126	70	Lighthouses
Tenerife	Guajara	5.0	34	28	6	NaN	NaN	NaN	Mountains
anzarote	Marina Rubicon	4.5	9,459	4,959	3,559	773	122	NaN	Marinas
Gran Canaria	Playa de Maspalomas	4.5	8,529	4,862	2,733	714	151	69	Beaches

Figure 19. Row Dataframe, Own source

After cleaning and reducing the Dataframe by 21% in the ATT_EDA.ipynb file, a copy of the Dataframe was created in the Curet_ATT_CI_TA.csv file. Finally, the result of the Dataframe is displayed in Figure 20.



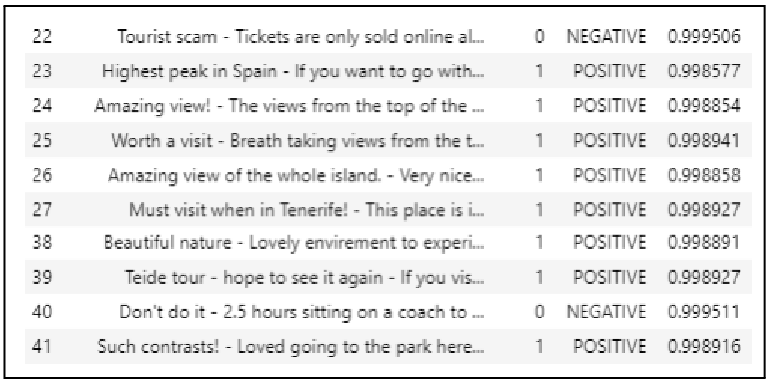
u_Id	u_Country	u_Contributions	u_Trip_type	Review	r_Rating	r_Date	r_Helpful
717354021	United Kingdom	7	Unknown	Worth the trip, cable car needs minimum 90 min...	4.0	2022-07-26	1
946286476	United Kingdom	7	Unknown	Must see of Tenerife - A must see site on Tene...	5.0	2022-07-25	0
784077896	United Kingdom	44	Family	A must visit place in tenerife. - Absolutely a...	5.0	2022-07-17	0
491263	Unknown	8	Family	Hike to the summit. - A drive up to El Tiede f...	5.0	2022-07-17	0
550290313	United Kingdom	52	Couples	Spectacular - It's number one for a reason. O...	5.0	2022-07-17	1

i_Id	Item	i_Island	i_Rating	i_Reviews	i_Excellent	i_Very_good	i_Average	i_Poor	i_Terrible	i_Type
717354021	Volcan El Teide	Tenerife	4.5	13470	9917	2729	552	151	121	Volcanos
717354021	Volcan El Teide	Tenerife	4.5	13470	9917	2729	552	151	121	Volcanos
717354021	Volcan El Teide	Tenerife	4.5	13470	9917	2729	552	151	121	Volcanos
717354021	Volcan El Teide	Tenerife	4.5	13470	9917	2729	552	151	121	Volcanos
717354021	Volcan El Teide	Tenerife	4.5	13470	9917	2729	552	151	121	Volcanos

Figure 20. Ready to process Dataframe, Own source

6.2 Sentiment Analysis with roBERTa (Q2)

Answering Q2, it can be ensured that the preprocessing of the data was adequate for its subsequent use with state-of-the-art techniques. This study successfully processed TripAdvisor reviews with LDA and with BERTopic (for Topic Modelling), as well as with roBERTa (for SA). The results of the Topic Modeling will serve for the report that responds to Q4, as well as the SA visualizations. A sample of the SA results can be seen below (Figure 21), where it can be roughly differentiated that the results are adequate.



22	Tourist scam - Tickets are only sold online al...	0	NEGATIVE	0.999506
23	Highest peak in Spain - If you want to go with...	1	POSITIVE	0.998577
24	Amazing view! - The views from the top of the ...	1	POSITIVE	0.998854
25	Worth a visit - Breath taking views from the t...	1	POSITIVE	0.998941
26	Amazing view of the whole island. - Very nice...	1	POSITIVE	0.998858
27	Must visit when in Tenerife! - This place is i...	1	POSITIVE	0.998927
38	Beautiful nature - Lovely envirement to experi...	1	POSITIVE	0.998891
39	Teide tour - hope to see it again - If you vis...	1	POSITIVE	0.998927
40	Don't do it - 2.5 hours sitting on a coach to ...	0	NEGATIVE	0.999511
41	Such contrasts! - Loved going to the park here...	1	POSITIVE	0.998916

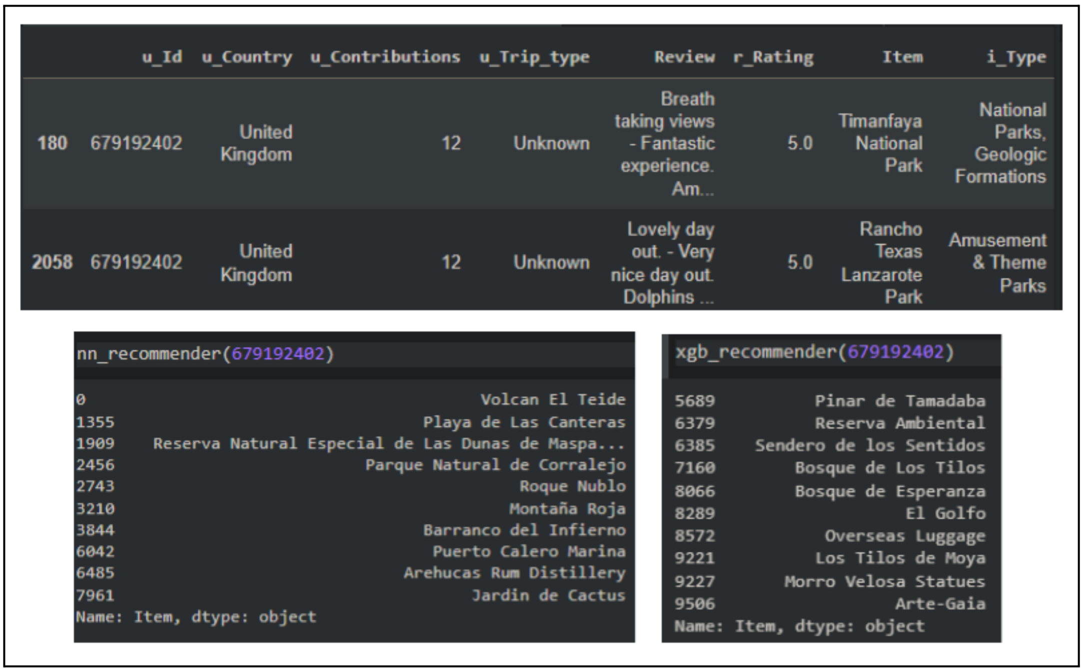
Figure 21. roBERTa results, Own source

These applications can be consulted in the notebooks: ATT_NLP_LDA.ipynb, ATT_NLP_BERTopic.ipynb and ATT_NLP_Sentiment-roBERTa.ipynb.

6.3 Exploring Collaborative Filtering (Q3)

The third research question corresponds to the business side. With this, the research aims to develop tools and expose information that may be useful for OTAs, DMOs and for Travel Sustainability. Therefore, the creation of RS, as a new feature for the Attractions in the Canary Islands, was presented as a solution.

The proposed solution of developing two RS of the Collaborative Filtering Model-based type was carried out without problems, the data could be processed, so they had adequate pre-processing. The research managed to develop an RS with Neural Networks and another with Decision Trees (XGBoost) that had adequate results:



	u_Id	u_Country	u_Contributions	u_Trip_type	Review	r_Rating	Item	i_Type
180	679192402	United Kingdom	12	Unknown	Breath taking views - Fantastic experience. Am...	5.0	Timanfaya National Park	National Parks, Geologic Formations
2058	679192402	United Kingdom	12	Unknown	Lovely day out. - Very nice day out. Dolphins ...	5.0	Rancho Texas Lanzarote Park	Amusement & Theme Parks

```
nn_recommender(679192402)

0                                              Volcan El Teide
1355                                      Playa de Las Canteras
1909    Reserva Natural Especial de Las Dunas de Maspa...
2456                                 Parque Natural de Corralejo
2743                                                Roque Nublo
3210                                               Montaña Roja
3844                                      Barranco del Infierno
6042                                        Puerto Calero Marina
6485                                     Arehucas Rum Distillery
7961                                           Jardin de Cactus
Name: Item, dtype: object
```

```
xgb_recommender(679192402)

5689          Pinar de Tamadaba
6379          Reserva Ambiental
6385     Sendero de los Sentidos
7160          Bosque de Los Tilos
8066          Bosque de Esperanza
8289                   El Golfo
8572            Overseas Luggage
9221           Los Tilos de Moya
9227        Morro Velosa Statues
9506                  Arte-Gaia
Name: Item, dtype: object
```

Figure 22. roBERTa results, Own source

On the other hand, two drawbacks were found: first, the learning curve of the nn_recomender model shows high levels and, therefore, insufficient. On the other hand, the RMSE could not be computed for the RMSE model check. Therefore, it can be affirmed that the research was able to develop adequate RS with the data it has, but its level of reliability must be improved, by obtaining a greater amount of data, better processing of the models or other models that return better results. Secondly, the creation of a hybrid model was presented as a more powerful and recurring solution, uniting the two RS developed. This solution has not been able to be developed satisfactorily, so the need for improvements at this point in the project is confirmed.

6.4 Topic Modeling and Data Visualization report (Q4)

The last result of this research has to do with giving the context the importance it has, in taking into account the socioeconomic position that surrounds the territory to which this research alludes. Likewise, for this report, TripAdvisor data is interpreted as valid to make judgments, in general terms, about tourism in the Canary Islands. Well, as highlighted in different investigations, e-Travel is a form of travel that gains importance year after year due to its impact on businesses and travellers who use it. Therefore, we can differentiate the traveller who uses these platforms as a desired type of tourist, either because of the possibility of analyzing their opinion or because they are characterized as a consumer of Attractions and an active tourist.

Thus, as discussed in point 2.1.1 *The Tourism Industry in the Canary Islands*, the amount of employment and GDP generated by this industry was reduced by around 65%. These data are also reflected in the number of reviews that have been made. Figure 23 shows two breaks, one in mid-March 2021 (beginning of restrictions and closing of borders) and another in mid-January (second package of highly restrictive measures towards the freedoms of citizens), coinciding with government policies to reduce the levels of contagion of the Covid-19 virus. There is also a progressive improvement from the summer of 2021 (measures to restrict freedoms began to be reduced), and as of August 2022, the number of reviews in the Canary Islands on TripAdvisor means 40% of what It was before the pandemic. These data demonstrate the success in the contextualization of the problem.

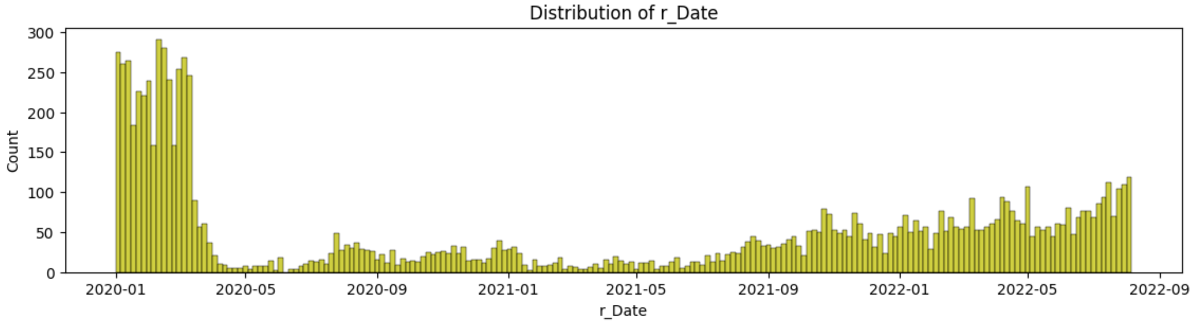


Figure 22. Number of reviews per time, Own source

Going deeper into the way in which these reviews are distributed by travel profile, we have already seen how the type of trip that is made most frequently is with the family, then with a group of friends, then in pairs and, finally, alone (Figure 17). If we were to take this data as a travel census, it could be confirmed that the preferred form of travel to the Canary Islands is in a group and that the free-entry Attractions or attractions of cultural interest that are presented and can be enjoyed are more used as a family, as a couple or alone. Also, from a sustainable travel perspective, there are interesting figures regarding the distribution of tourism in the Canary Islands. The unequal distribution of records on the different islands is notorious:

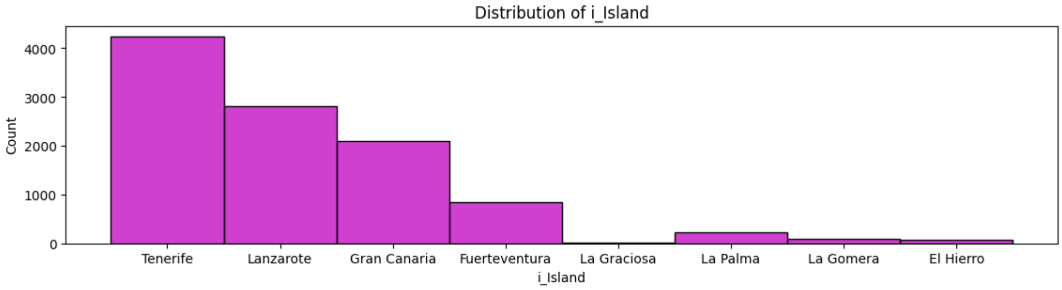


Figure 23. Distribution of reviews per island, Own source

Regarding the knowledge extracted from Topic Modelling, the proposed solutions show interesting findings. As we saw in figure 11 on LDA, we can highlight different topics and their relationship with others, where a strong relationship between the words Park and Beach stands out in all the topics. In the case of BERTopic, in addition to the relationships displayed in the similarity matrix (Appendix 2), the Topic Word Scores highlights Parks, Plants, Museum, Market, Dunes, Church, Golf and Marina:

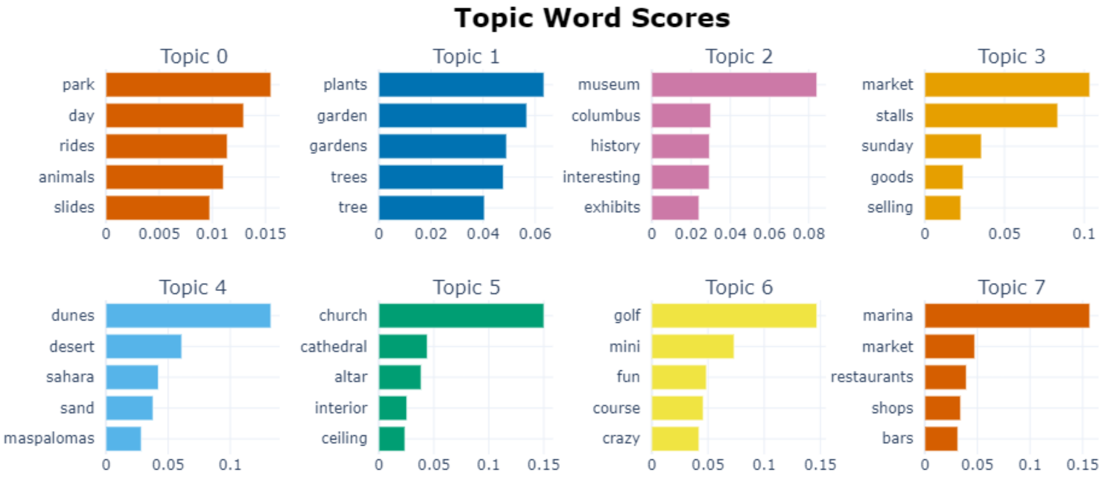


Figure 24. Own source

In addition, another powerful feature of BERTopic is Hierarchical Clustering, where the following relationships can be seen within the same three large clusters:

CLUSTER 1: *Market_sunday & marina_marke // beach_restaurant_bars & place_shos_nice*
CLUSTER 2: *Timanfaya_volcanp & tour_coach // tenerife_teide & cable_car*
CLUSTER 3: *Plants_garden & cactus_garden // night_dancing_show & blankets_movie_film*

7. Conclusion and future work

The research has successfully faced the different challenges posed in its four phases. In the first place, obtaining data from TripAdvisor, a platform that has a payment API for the release of information, as well as limitations when it comes to scraping. Next, the use of NLP techniques with libraries like *nltk*, *pyLDAvis*, *gensim* and transformers like *roBERTa* and *BERTopic*. Third, the construction of *Collaborative Filtering model-based* on TensorFlow. And, finally, the realization of a report with visualizations for understanding the data, to which the *seaborn* and *matplotlib*.

With this in mind, from the first phase, it is concluded that the tool that has given the best results for obtaining data from TripAdvisor is Octoparse, due to its facilities and capabilities. Making scraping easier compared to tools like Scrapy or Beatifulsoup or Selenium. The second conclusion is that the Topics that stand out for their use and relationship among all the reviews are Beach and Parks according to LDA; and Parks, Plants, Museum, Market, Dunes and Church according to BERTopic; giving both useful results for the interpretation of the data. It also highlights the relevance shown by BERTopic's Hierarchical Clustering. Likewise, it is concluded that the use of Collaborative Filtering model-based RS for TripAdvisor data is adequate, but that a review with more data or improvements in the models is necessary to achieve better results, as well as the creation of a Hybrid RS.

Finally, the results of all the applied techniques, the visualization of results and the report made, invite the acceptance of TripAdvisor data as valid and representative. With their reviews, users contribute to the growth of e-Travel and promote active and conscious tourism. In the same way, there is evidence of a possibility of redistribution of the amount of tourism for each island, due to the disparity between them. It is also concluded that as of August 2022, the contributions of TripAdvisor users in Attractions of the Canary Islands is close to 40% of what it was before the Covid-19 pandemic. Showing a large space for tourism recovery, and with it, an opportunity for reconstruction.

In future work, the exploration of new RS models is proposed together with the injection of data in the form of streaming with its corresponding pipeline or data flow system, for the development of a marketable RS for Travel Sustainability purposes.

References

1. Honglui CAO, Phd & Anna Tsolakou (2020). Travel Recommendation System using destination similarity. *Amadeus for Developers*.

2. Foley, Becky (2021). The Power of Reviews: How TripAdvisor Reviews Lead to Booking and Better Travel Experiences. *TripAdvisor.com*

3. Ali, T., Marc, B., Omar, B., Soulaimane, K., & Larbi, S. (2021). Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor. *Tourism Management Perspectives*, *40*, 100892. https://doi.org/10.1016/j.tmp.2021.100892

4. S. Aciar, D. Zhang, S. Simoff, and J. Debenham, "Informed Recommender: Basing Recommendations on Consumer Product Reviews, "in IEEE Intelligent Systems, vol. 22, no. 3, p. 39-47, May-June 2007, doi: 10.1109/MIS.2007.55.

5. SM Al-Ghuribi and SA Mohd Noah, "Multi-Criteria Review-Based Recommender System–The State of the Art," in IEEE Access, vol. 7, p. 169446-169468, 2019, doi: 10.1109/ACCESS.2019.2954861.

6. Raina, V., Krishnamurthy, S. (2022). “Natural Language Processing”. In: Building an Effective Data Science Practice. Press, Berkeley, CA. https://doi.org/10.1007/978-1-4842-7419-4_6.

7. Logesh, R., Subramaniyaswamy, V. (2019). Exploring Hybrid Recommender Systems for Personalized Travel Applications. In: Mallick, P., Balas, V., Bhoi, A., Zobaa, A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore. https://doi.org/10.1007/978-981-13-0617-4_52

8. Nadezhda, D. (2020). Recommendation System for Travelers Based on TripAdvisor.com Data. Saint Petersburg School of Economics and Management. In: 03.38.02 'Management'.

9. Arenas-Marquez, FJ, Martinez-Torres, R., & Toral, S. (2021). Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor. *Information Processing & Management*, *58*(5), 102645. https://doi.org/10.1016/j.ipm.2021.102645

10. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Cornell University. doi: https://doi.org/10.48550/arXiv.1910.01108

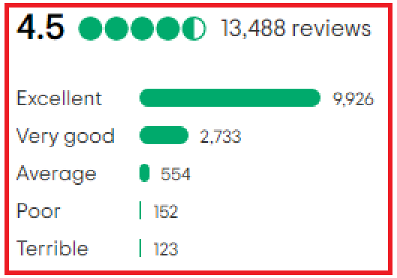
11. Zhuang, Yuanyuan, and Jaekyeong Kim. 2021. "A BERT-Based Multi-Criteria Recommender System for Hotel Promotion Management" Sustainability 13, no. 14: 8039. https://doi.org/10.3390/su13148039

12. https://www.freepik.es/foto-gratis/element-design-icon-recycling-green_15559635.htm#query=sustainability%20icon&position=4&from_view=search (rawpixel.com)

13. https://github.com/GuilleAlte/TripAdvisorData-NLP-RecommenderSystem

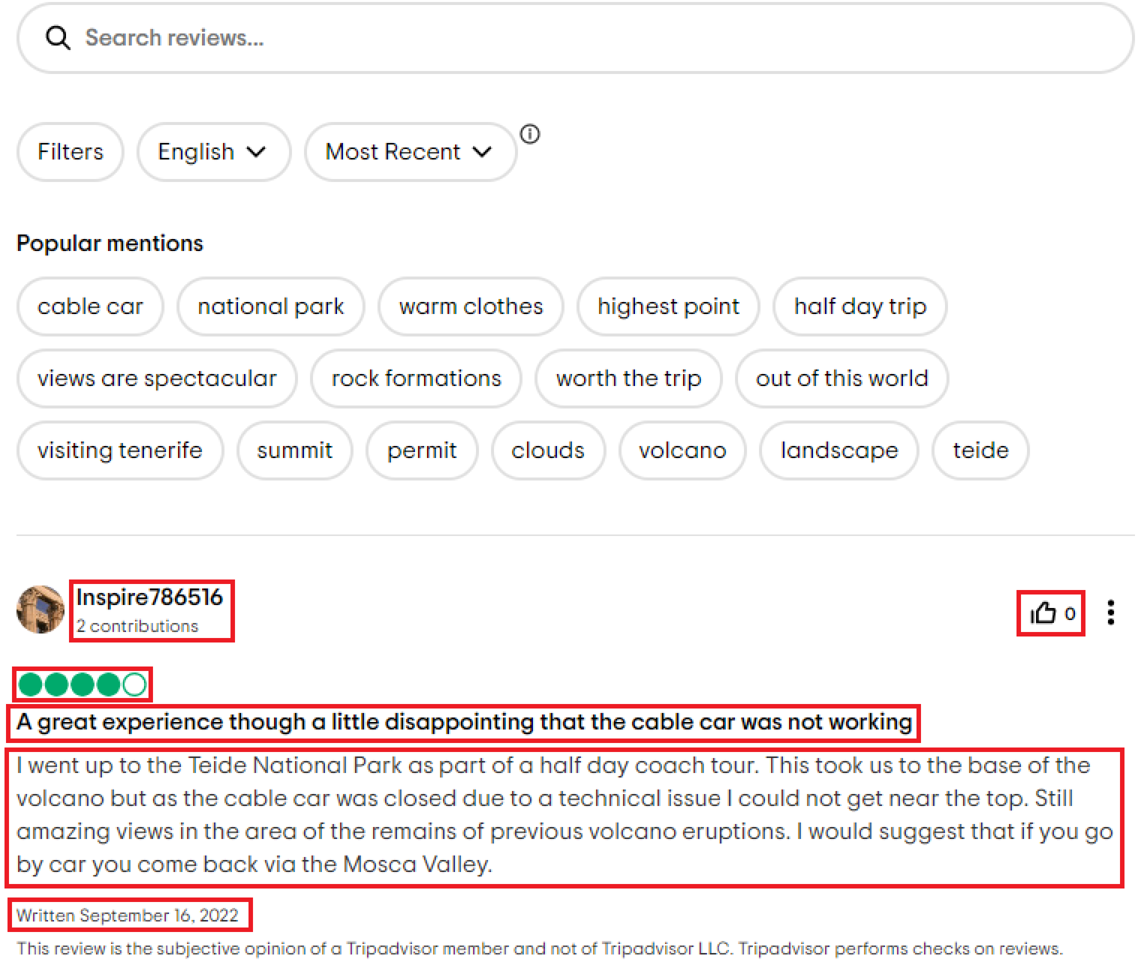
Appendixes

Appendix 1:



4.5 ●●●●◐ 13,488 reviews

Excellent	9,926
Very good	2,733
Average	554
Poor	152
Terrible	123



Popular mentions

Inspire786516
2 contributions

●●●●○

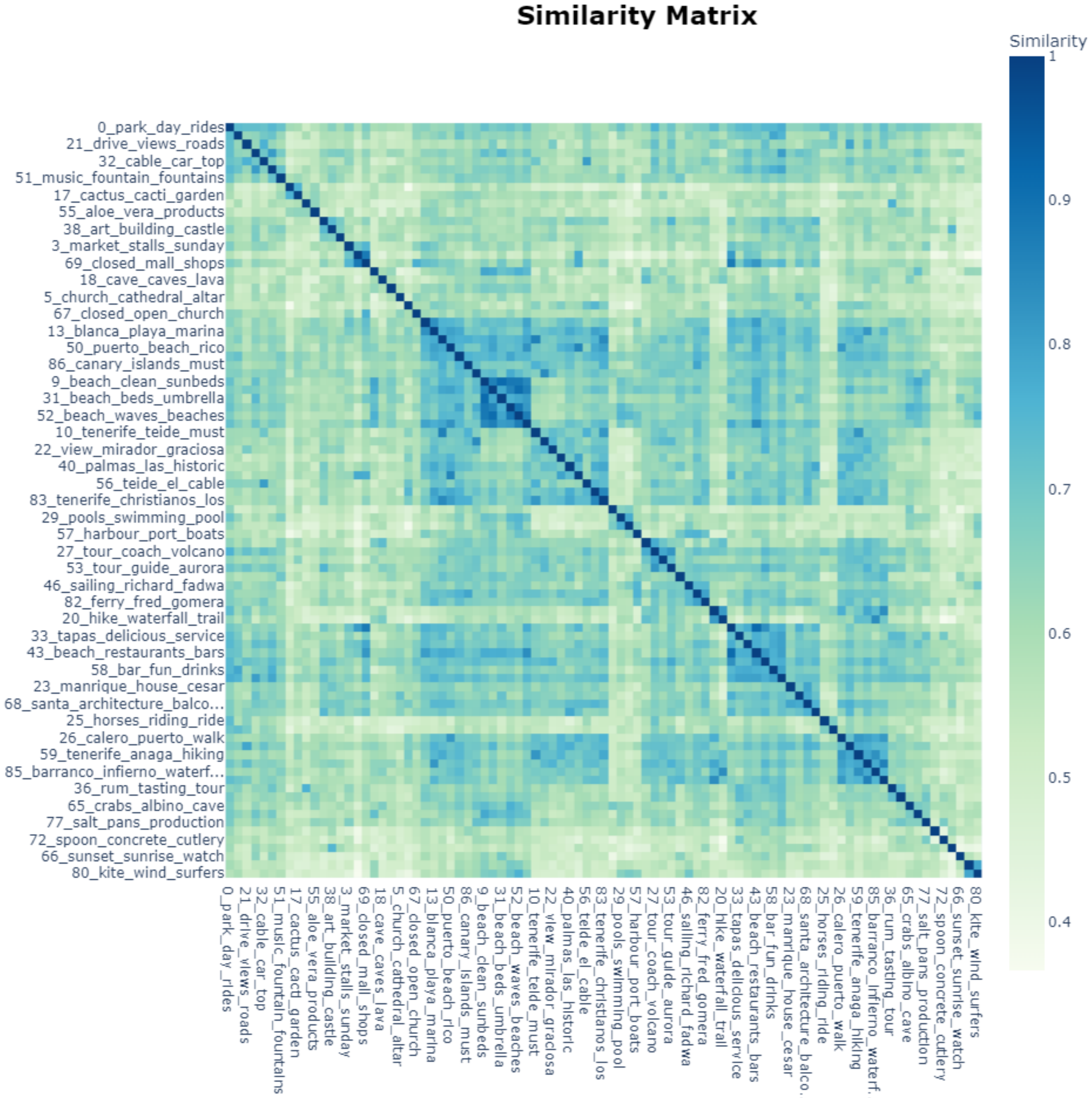
A great experience though a little disappointing that the cable car was not working

I went up to the Teide National Park as part of a half day coach tour. This took us to the base of the volcano but as the cable car was closed due to a technical issue I could not get near the top. Still amazing views in the area of the remains of previous volcano eruptions. I would suggest that if you go by car you come back via the Mosca Valley.

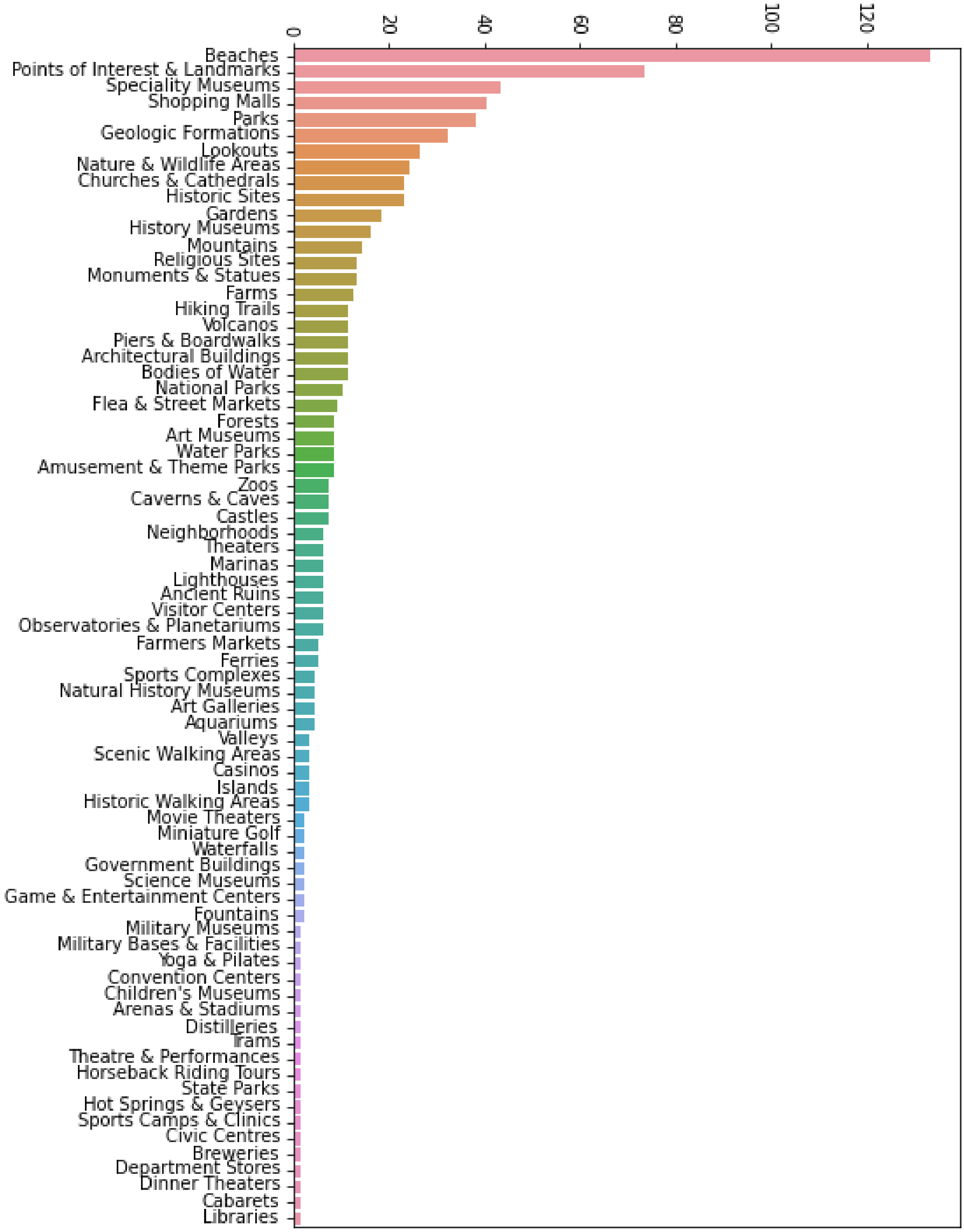
Written September 16, 2022

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

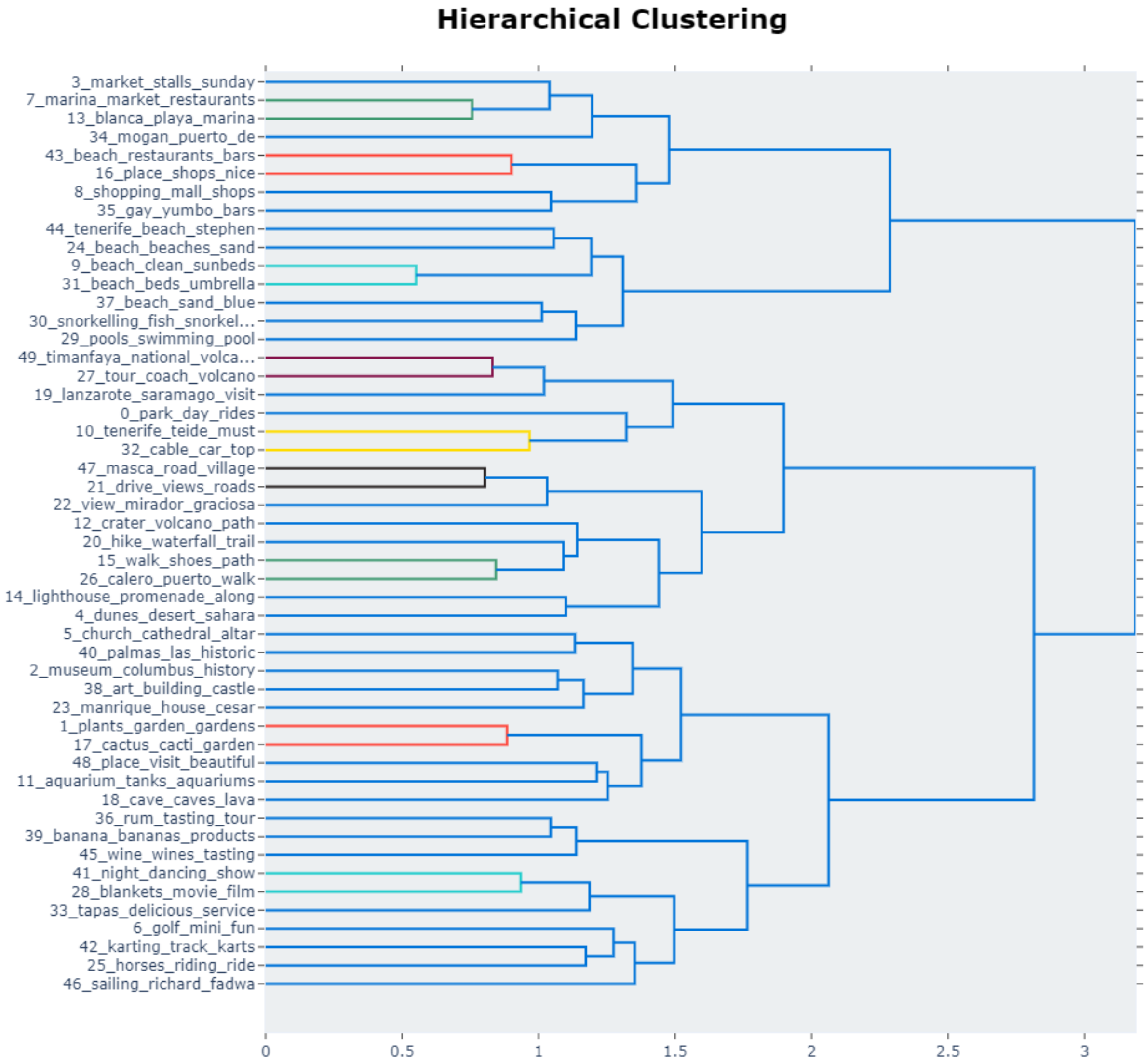
Appendix 2:


Similarity Matrix
Similarity
1
0.9
0.8
0.7
0.6
0.5
0.4
0_park_day_rides
21_drive_views_roads
32_cable_car_top
51_music_fountain_fountains
17_cactus_cacti_garden
55_aloe_vera_products
38_art_building_castle
3_market_stalls_sunday
69_closed_mall_shops
18_cave_caves_lava
5_church_cathedral_altar
67_closed_open_church
13_blanca_playa_marina
50_puerto_beach_rico
86_canary_islands_must
9_beach_clean_sunbeds
31_beach_beds_umbrella
52_beach_waves_beaches
10_tenerife_teide_must
22_view_mirador_graciosa
40_palmas_las_historic
56_teide_el_cable
83_tenerife_christianos_los
29_pools_swimming_pool
57_harbour_port_boats
27_tour_coach_volcano
53_tour_guide_aurora
46_sailing_richard_fadwa
82_ferry_fred_gomera
20_hike_waterfall_trail
33_tapas_delicious_service
43_beach_restaurants_bars
58_bar_fun_drinks
23_manrique_house_cesar
68_santa_architecture_balco...
25_horses_riding_ride
26_calero_puerto_walk
59_tenerife_anaga_hiking
85_barranco_infierno_waterf...
36_rum_tasting_tour
65_crabs_albino_cave
77_salt_pans_production
72_spoon_concrete_cutlery
66_sunset_sunrise_watch
80_kite_wind_surfers
0_park_day_rides
21_drive_views_roads
32_cable_car_top
51_music_fountain_fountains
17_cactus_cacti_garden
55_aloe_vera_products
38_art_building_castle
3_market_stalls_sunday
69_closed_mall_shops
18_cave_caves_lava
5_church_cathedral_altar
67_closed_open_church
13_blanca_playa_marina
50_puerto_beach_rico
86_canary_islands_must
9_beach_clean_sunbeds
31_beach_beds_umbrella
52_beach_waves_beaches
10_tenerife_teide_must
22_view_mirador_graciosa
40_palmas_las_historic
56_teide_el_cable
83_tenerife_christianos_los
29_pools_swimming_pool
57_harbour_port_boats
27_tour_coach_volcano
53_tour_guide_aurora
46_sailing_richard_fadwa
82_ferry_fred_gomera
20_hike_waterfall_trail
33_tapas_delicious_service
43_beach_restaurants_bars
58_bar_fun_drinks
23_manrique_house_cesar
68_santa_architecture_balco...
25_horses_riding_ride
26_calero_puerto_walk
59_tenerife_anaga_hiking
85_barranco_infierno_waterf...
36_rum_tasting_tour
65_crabs_albino_cave
77_salt_pans_production
72_spoon_concrete_cutlery
66_sunset_sunrise_watch
80_kite_wind_surfers

Appendix 3:


0
20
40
60
80
100
120
Beaches
Points of Interest & Landmarks
Speciality Museums
Shopping Malls
Parks
Geologic Formations
Lookouts
Nature & Wildlife Areas
Churches & Cathedrals
Historic Sites
Gardens
History Museums
Mountains
Religious Sites
Monuments & Statues
Farms
Hiking Trails
Volcanos
Piers & Boardwalks
Architectural Buildings
Bodies of Water
National Parks
Flea & Street Markets
Forests
Art Museums
Water Parks
Amusement & Theme Parks
Zoos
Caverns & Caves
Castles
Neighborhoods
Theaters
Marinas
Lighthouses
Ancient Ruins
Visitor Centers
Observatories & Planetariums
Farmers Markets
Ferries
Sports Complexes
Natural History Museums
Art Galleries
Aquariums
Valleys
Scenic Walking Areas
Casinos
Islands
Historic Walking Areas
Movie Theaters
Miniature Golf
Waterfalls
Government Buildings
Science Museums
Game & Entertainment Centers
Fountains
Military Museums
Military Bases & Facilities
Yoga & Pilates
Convention Centers
Children's Museums
Arenas & Stadiums
Distilleries
Trams
Theatre & Performances
Horseback Riding Tours
State Parks
Hot Springs & Geysers
Sports Camps & Clinics
Civic Centres
Breweries
Department Stores
Dinner Theaters
Cabarets
Libraries

Appendix 4:


Hierarchical Clustering
3_market_stalls_sunday
7_marina_market_restaurants
13_blanca_playa_marina
34_mogan_puerto_de
43_beach_restaurants_bars
16_place_shops_nice
8_shopping_mall_shops
35_gay_yumbo_bars
44_tenerife_beach_stephen
24_beach_beaches_sand
9_beach_clean_sunbeds
31_beach_beds_umbrella
37_beach_sand_blue
30_snorkelling_fish_snorkel...
29_pools_swimming_pool
49_timanfaya_national_volca...
27_tour_coach_volcano
19_lanzarote_saramago_visit
0_park_day_rides
10_tenerife_teide_must
32_cable_car_top
47_masca_road_village
21_drive_views_roads
22_view_mirador_graciosa
12_crater_volcano_path
20_hike_waterfall_trail
15_walk_shoes_path
26_calero_puerto_walk
14_lighthouse_promenade_along
4_dunes_desert_sahara
5_church_cathedral_altar
40_palmas_las_historic
2_museum_columbus_history
38_art_building_castle
23_manrique_house_cesar
1_plants_garden_gardens
17_cactus_cacti_garden
48_place_visit_beautiful
11_aquarium_tanks_aquariums
18_cave_caves_lava
36_rum_tasting_tour
39_banana_bananas_products
45_wine_wines_tasting
41_night_dancing_show
28_blankets_movie_film
33_tapas_delicious_service
6_golf_mini_fun
42_karting_track_karts
25_horses_riding_ride
46_sailing_richard_fadwa
0
0.5
1
1.5
2
2.5
3