

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

Utilising 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

Resumen	3
1. Introducción	4
2. Estado del Arte	6
2.1 Contextualización	6
2.1.1 La Industria Turística en Canarias	6
2.1.2 Aplicación de la Ciencia de Datos a la Industria Turística	6
2.1.3 TripAdvisor's reviews	7
2.2 Conceptualización técnica	8
2.2.1 Recommender Systems	8
2.2.2 Natural Language Processing	9
2.3 Related works	9
2.3.1 Logros de las últimas investigaciones	9
2.3.2 Técnicas transformadoras	11
3. Objetivos	12
4. Solución	13
4.1 Fase 1: TripAdvisor data ingestion	14
4.1.1 Elección de la muestra	14
4.1.2 Procesos de extracción:	15
4.1.3 Exploratory Data Analysis and cleaning	16
4.2 Fase 2: Natural Language Processing	17
4.3 Fase 3: Recommender Systems	18
5. Evaluación	19
5.1 Evaluación del proceso de ingesta de datos	19
5.2 Evaluación del proceso de NLP	20
5.3 Evaluación de la creación de RS	21
5.4 Evaluación de visualización de datos	22
6. Resultados (6 páginas)	24
6.1 Octoparse as the scrapping solution (Q1)	25
6.2 Sentiment Analysis with roBERTa (Q2)	26
6.3 Exploring Collaborative Filtering (Q3)	27
6.4 Topic Modelling and Data Visualisation report (Q4)	28
7. Conclusión y trabajos futuros	30
References	31
Appendixes	32

Resumen

Las notorias cifras de PIB y de generación de empleo que colocan la Industria Turística como la principal en Canarias (España) y la recesión que estas han tenido debido a la pandemia Covid-19, muestran una oportunidad de reconstrucción hacia un turismo sostenible. Para ello, esta investigación plantea la utilización de datos de TripAdvisor y su procesamiento con state-of-the-art techniques en Data Science. Todo esto bajo el paradigma de que la utilización de NLP o Recommender Systems en estos datos pueden constituir la creación de herramientas comerciables para el turismo sostenible, basado en los éxitos que estas tecnologías han tenido en otras industrias (Netflix, Amazon, Spotify, etc.). La investigación cumple sus objetivos orientados a la obtención de datos de TripAdvisor, el procesamiento de Natural Language Processing (mediante la extracción de Topics y de Sentiment Analysis) y de creación de Recommender Systems de tipo Collaborative Filtering model-based, así como la interpretación de estos datos mediante visualizaciones y reportes. Por último, se propone para un futuro proyecto, la ingestión de datos en streaming y un Hybrid Recommender System colaborando a la creación de un producto comerciable para la generación de turismo sostenible.

1. Introducción

La Industria Turística es una de las grandes generadoras de trabajo y PIB en las Islas Canarias de España con cifras que en los últimos años se han aproximado al 40% en ambos casos. Debido a la pandemia Covid-19, estas cifras se han visto reducidas un 65% y comienzan a recuperarse en 2022. En este periodo, esta investigación se plantea posibilidades a la hora de reconstrucción turística, atendiendo al contexto y valorando los datos desde el momento del comienzo de estas recesiones, por la importancia que se le debe otorgar al presente y a las transformaciones que exige.

Teniendo en cuenta este contexto, entra en escena el concepto de e-Travel y la importancia de plataformas como TripAdvisor en la planificación de los viajes, así como en la búsqueda de atractivos turísticos para las personas que planean sus viajes. De igual modo, en las diferentes investigaciones trabajadas en el marco teórico, se valora este tipo de plataformas y el contenido que se comparte en ellas (sobretudo los ratings y reviews a las actividades) para su uso en la transformación del turismo. Aunque, por supuesto, la búsqueda del aumento de consumo y de la generación de capital es vital, este proyecto valora esta regeneración como una posibilidad de creación de Travel Sustainability. La cual no tiene que significar reducción, sino un replanteamiento del turismo y una distribución más inteligente.

Son significantes y múltiples las posibilidades desde la Ciencia de Datos para la consecución de este tipo de transformaciones. Por ello, esta investigación trae a la primera línea las state-of-the-art techniques de Natural Language Processing (NLP) y los Recommender Systems o Recommendation Systems (RS) mediante el tratamiento de datos de portales como TripAdvisor, buscando la creación de nuevas características dentro de ellos, o de la extracción de conocimiento de los mismos para Online Travel Agencies (OTA) o Destination Management Organizations (DMO).

Esta investigación plantea dos objetivos principales, uno académico, orientado hacia la exploración de las técnicas nombradas, aportando al entorno académico, y otro de tipo business, con la intención de construir herramientas y reportes que puedan ser el comienzo a un proyecto mayor, comercializable, para aportar en una deseada transición sostenible en manos de la tecnología y la Inteligencia Artificial.

La metodología consiste en la extracción de datos de TripAdvisor mediante las técnicas conocidas de Scraping, donde se prueban *Scrapy*, *Beautifulsoup* y *Selenium*, así como la API de *Octoparse*. A continuación se realiza una preparación de los datos y, tras esto, las siguientes fases de procesamiento de NLP, creación de RS y la extracción de tablas y visualizaciones. Todo esto se realizan en diferentes Jupyter Notebooks con Python.

Como conclusiones principales, se recomienda el uso de la API de *Octoparse* para la extracción de datos de TripAdvisor, debido a las facilidades que muestra para superar las trabas que la plataforma TripAdvisor a la hora de interpretar su html. También se recomienda el uso de transformers para el procesamiento de NLP, en específico destacan BERTopic en el Topic Modelling y roBERTa en el Sentiment Analysis (SA). En adición, se concluye que el uso de RS de tipo Collaborative Filtering model-based es adecuado para el procesamiento de este tipo de información, teniendo en cuenta la capacidad de mejora en los modelos desarrollados mediante la inclusión de más datos y la exploración de mejoras en el aprendizaje de los mismos.

Por último, se concluye que los datos de TripAdvisor son representativos y validos, coincidiendo el descenso de reviews con las bajadas de las cifras en la Industria Turísticas anteriormente nombradas. Se interpreta que los usuarios de estos portales aportan al crecimiento del e-Travel, al fomentar el turismo activo y consciente, a la vez de aportando un feedback utilizable tanto por la red de viajeros, como por las entidades que rodean la industria, OTAs y DMOs.

Como trabajo futuro, se plantea la exploración de nuevos modelos de RSs junto con la inyección de datos en forma de streaming con su correspondiente pipeline o sistema de flujo de datos, para el desarrollo de un RS comerciable con fines de Travel Sustainability.

2. Estado del Arte

2.1 Contextualización

2.1.1 La Industria Turística en Canarias

Según las estadísticas publicadas en la web del Gobierno de Canarias, the Tourism Industry ha significado en los últimos años en las Islas Canarias al rededor de un 35% del empleo y un 35% del PIB, lo cual demuestra una economía diseñada hacia este sector y en cierto punto dependiente del mismo a día de hoy. Con la pandemia de Covid-19 comenzada el 1 de diciembre, pero realmente impactando sus medidas de restricción de libertades por estado de alarma en España a mediados de marzo de 2020, estas cifras se vieron castigadas: el PIB Turístico descendió a un 11%, al igual que el empleo.

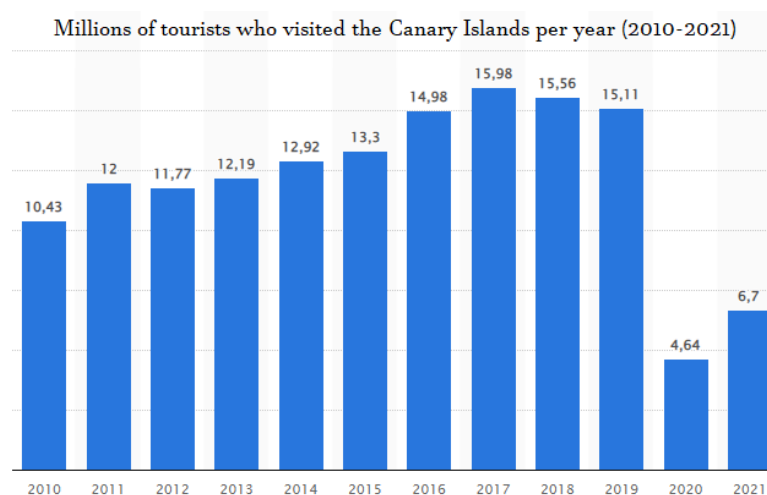


Figure 1. Source: es.statista.com

2.1.2 Aplicación de la Ciencia de Datos a la Industria Turística

Hoy en día el e-Travel es una de las formas preferidas de planificar los viajes por parte de los turistas. Existen infinidad de portales web y aplicaciones en los que investigar (Trivago, TripAdvisor, Expedia, Booking, etc.). Estos portales recaban una cantidad de información sobre usuarios y destinos cada día más extensa y precisa (fotografías, ratings, definiciones, precios, localizaciones, reviews, etc.), lo cual los convierte en una ideal fuente de datos para la investigación y la creación de modelos de Machine Learning y Deep Learning.

Entre las variables más valorados para la práctica de estas disciplinas están las *reviews*, donde los usuarios comentan con libertad y extensión su opinión sobre un Punto de Interés (POI) (e.g. “Volcan El Teide” o “Parque Timanfaya”). Información procesable en técnicas de NLP y RS y comerciable mediante la creación de Travel Apps, para su uso en OTA o DMO así con motivos de Travel Sustainability: para gestionar y reducir el impacto del turismo o para la encuentra de alternativas similares cercanas, reduciendo las emisiones. [1]

2.1.3 TripAdvisor’s reviews

Según el estudio “The Power of Reviews: How TripAdvisor Reviews Lead to Booking and Better Travel Experiences” [2], 3 de cada 4 encuestados valoran como extremadamente o muy importante la revisión de reviews antes de tomar decisiones de viaje. El portal cuenta con más de 1 billón de las mismas, 26 millones submitted en 2021 con una media de 633 caracteres (3 veces más grandes que otros OTAs), lo cual quiere decir que la mayoría de las reviews son long form, algo muy atractivo desde el punto de vista del procesamiento de texto o NLP, pues a mayor contenido, más posibilidad de extraer conocimiento de ella y de crear modelos que las interpreten.

Las reviews de TripAdvisor cuentan con los atributos que se muestran a continuación en la Figura 1 y se pueden visualizar en el Appendix 1. Corresponde subrayar que este tipo de plataformas han cambiado profundamente la forma en que se consume el turismo y la forma en la que se busca y comparte información de viaje [3, 4].

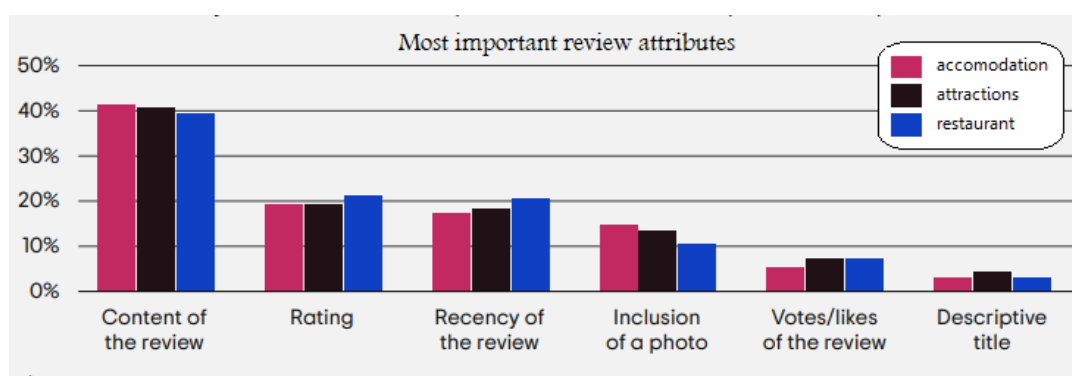


Figure 2. Source: Tripadvisor.com

2.2 Conceptualización técnica

En este apartado se definen los conceptos principales que se tratarán a lo largo de la investigación. En primer orden se definirán los diferentes tipos de RS, y luego se explicarán los alcances de las técnicas de NLP.

2.2.1 Recommender Systems

La gran cantidad de información disponible causa un reto a los usuarios, making the decision-making more complex, los RS filtran esta información y solucionan la sobrecarga de información. Aunque en primer lugar irrumpieron cambiando la forma en que los usuarios consumen películas o música (Netflix and Spotify), los RS están teniendo un papel protagonista a través de cantidad de industrias y plataformas líderes (Amazon, TripAdvisor, Facebook, etc.). El tipo de navegación que el usuario realiza, su perfil y similitud con otros perfiles, o hasta su feedback son las bases de estos exitosos instrumentos. Su función es sugerir al usuario lo que se define como la “best next action”, en otros términos, lo que los algoritmos predicen que al usuario le puede gustar utilizando lógicas y relaciones entre items (película or POI, e.g.) y usuarios.

Los algoritmos de los RS enfrentan retos de *cold start* (falta de datos para recomendar a ciertos usuarios) o *sparsity* (polarización de perfiles de usuarios, por ejemplo) que diferentes modelos intentan reducir y paliar. Los modelos o approaches de RS se pueden resumir de la siguiente manera [5]:

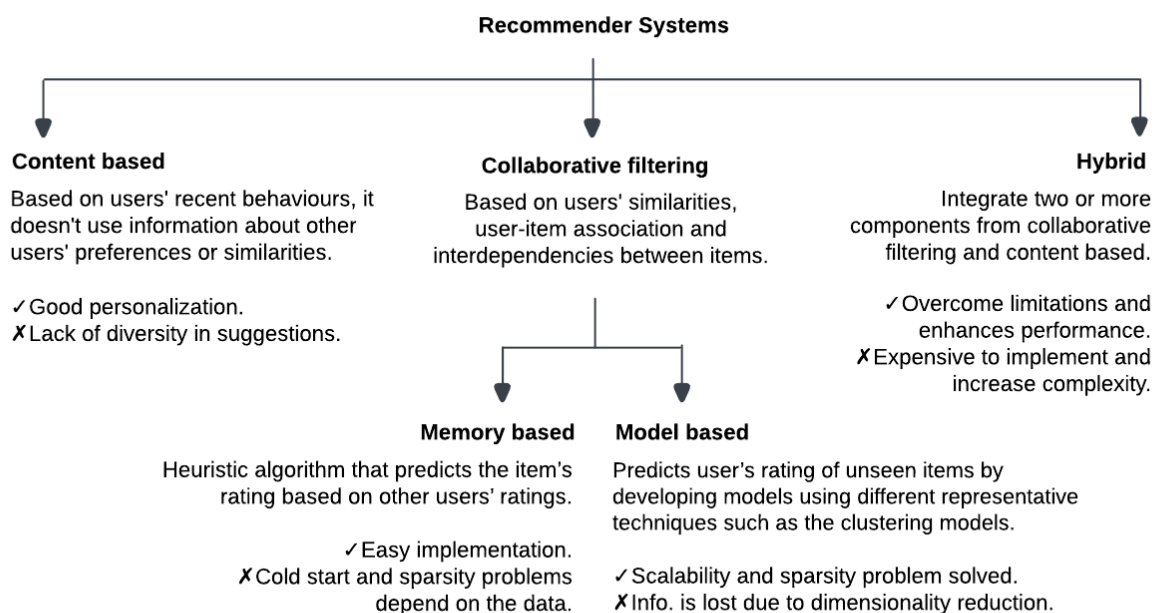


Figure 3. Own source

2.2.2 Natural Language Processing

Como su nombre sugiere, consiste en un conjunto de técnicas que tienen su raíz en enseñar a las computadoras a entender y hablar lenguajes naturales, lo cual tiene diferentes ramas productivas. Entre ellas están SA, donde se interpreta si por ejemplo una review tiene un sentimiento positivo, neutro o negativo; *document classification*, donde también se clasifican los textos según classes; la técnica de *autocomplete*, que puede completar frases, como ocurre comunmente en la redacción de mails; también es parte de NLP la *language translation*; por otra parte, *intent classification* es usado por chatbots para identificar la intención de los usuarios con sus textos; y, también, se puede realizar *text summarization* [6].

Con una de sus librerías más utilizadas y completas, *Natural Language Toolkit (NLTK)* podemos realizar la preparación de los datos en en los que se estructuran para ser procesados posteriormente. En primer orden, se realiza una tokenización (splitting the sentence into words), lower casing them, removing stop words (non important words like a, an, the, etc.), stemming them (transforming words to its root form, e.g. changing to chang) or lemmatization (to a word existing in the language, changing to change).

2.3 Related works

De las publicaciones académicas sobre RS hasta la fecha, destacan las aportaciones a RS híbridos o multi-criteria [4, 5, 7, 8, 9, 10]. Además, sólo uno de los nombrados no habla de datos de TripAdvisor.com [4]. Siendo este el tipo de RS preferido por las plataformas que cuentan con lógicas ítem-user y el que se propone aplicar para añadir mejoras a la plataforma o para extraer conocimiento de ella en términos de NLP. Y demostrando la repercusión de la plataforma de TripAdvisor en la Ciencia de Datos.

2.3.1 Logros de las últimas investigaciones

Específicamente sobre la aplicación de NLP sobre reviews, sin tratar RS se destaca la investigación de Ali et al, 2021, se trata la “Negative e-reputation using aspect based SA approach” en Marrakech. Se tratan 39,216 TripAdvisor reviews con NLP y se realiza Topic Modelling (extraction of main topics in the text) mediante Latent Dirichet Allocation (LDA), técnica que trata la similitud de los térinos según su posicionamiento y considering several dimensions in the document, donde cada topic representa a un et of words. Así consiguen

“extract hidden aspects and dimensions from feedbacks” los cuales pueden ser interpretados para mejorar el turismo en esta ciudad [3]

Volviendo a investigaciones que culminan su proceso NLP con RS, en la investigación de Al-Ghuribi & Mohd Noah de 2019 sobre “Multi-Criteria Review-Based Recommender System–The State of the Art” se critica que, hasta el momento, “most of the RSs approaches rely on a single-criterion” y aportan una solución a los RS en cuanto hay falta de datos de los usuarios. Esto es, cuando un usuario da un rating sobre la limpieza y el localización, se utilizan las reviews para completar esta información (e.g. el usuario hace un comentario sobre el staff y el wifi), luego esto se procesa por Hybrid RS que tienen en cuenta las lógicas item-user nombradas en anteriores apartados [5]. Por otro lado, Logesh & Subramaniaswamy publican en 2019 “Exploring Hybrid Recommender System for Personalized Travel Applications” donde proponen un *Personalized Context-aware Hybrid Travel Recommender System* (PCAHTRS), el cual va un paso más adelante en la obtención de datos de usuarios y quiere tener en cuenta datos contextuales del usuario sobre su movilidad y temporalidad a través de sus datos móviles durante sus viajes (Figure 5). Con esto, se superarían los tradicionales problemas de cold start y sparsity a la vez que se mejora la prediction accuracy [7].

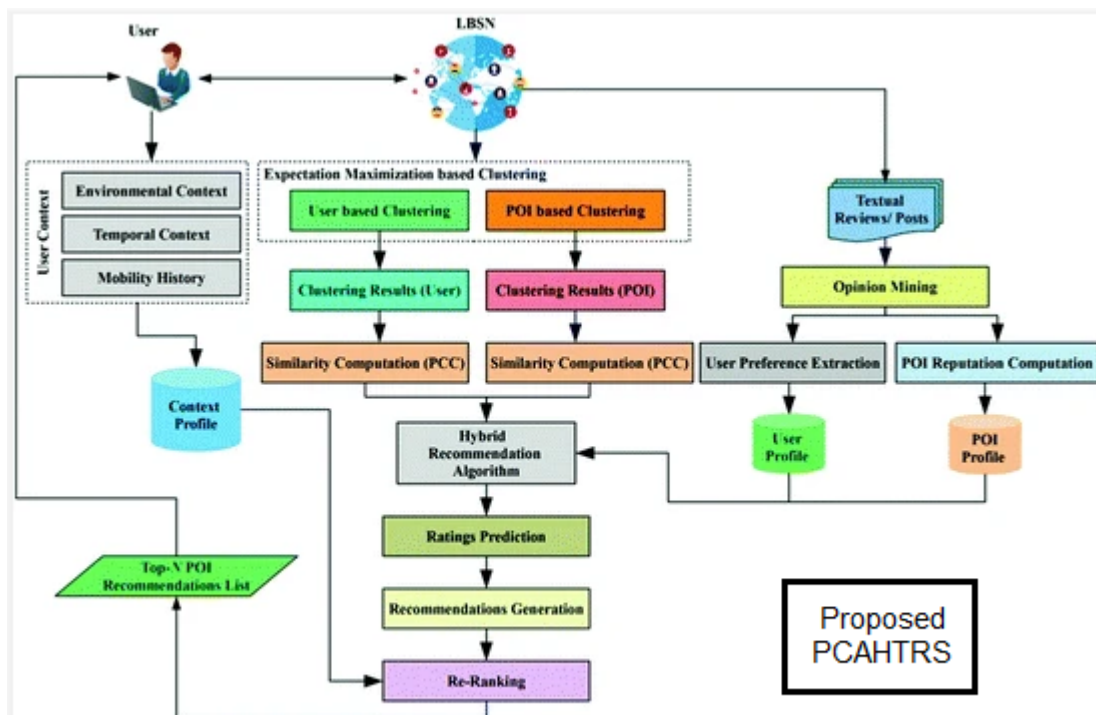


Figure 4. Source: Logesh & Subramaniaswamy (2019)

Una de las investigaciones más relevantes para este trabajo es la de Nadezhda (2020), "Recommendation System for Travellers Based on TripAdvisor.com Data". En el mismo, la autora utiliza LDA para extraer tópicos y con ellos configura una nueva clasificación de tipos de Actividades. TripAdvisor.com distingue, en el momento de la investigación nombrada, 51 tipo de actividades en la ciudad de Londres (e.g. Museums, Fun & Games, Nature & Parks) y la autora los reduce a 6 (Art, Food, Nature, Performing arts, Landmarks and Tours). Otro de los logros de esta investigación es la generación de un RS de tipo híbrido, que es además desarrollada en una interactive web-based service application [8].

Aunque la técnica del LDA es muy utilizada en Topic Modelling [3 y 8], existen investigaciones críticas con la misma y que proponen mejores modelos. Es el caso de Arenas-Márquez et al. (2021), en su publicación "Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor" concluyen que "the paper demonstrates that the neural encoding fitted as part of a classifier maximizes the discrimination of documents when compared to other encoding schemes" such as LDA [9].

2.3.2 Técnicas transformadoras

En el año 2017, Google's Brain Team (artificial intelligence and deep learning) introduced Transformers, which just as Recurrent Neural Networks (RNN) son capaces de procesar lenguajes, pero unlike RNNs, Transformers procesan the entire input at once. Transformers han generado grandes avances en NLP, interpretando significativamente mejor los lenguajes y abaratando el proceso. Los transformers cuentan con un conjunto de datos ya pre-entrenados o embeddings (vectorisation techniques) a los que se les realiza fine-tuning (se aplica ese aprendizaje a un nuevo texto). A esto, se le añaden los Transformers del grupo de Huggingface (<https://huggingface.co/>) que reducen aun más el costo de los transformers al entrenarlos con la mitad de palabras y manteniendo un 97% de rendimiento, en palabras de la investigación del opio grupo "reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster" [10].

En la investigación de Zhuang, Y. (2021), "A BERT-Based Multi-Criteria Recommender System for Hotel Management", el autor destaca las limitaciones que tienen los métodos actuales de SA. Por ello, propone un proceso que use BERT to compute customer aspect ratings and overall rating. Aunque el SA is more accurate, los resultados del RS no mejoran [11].

3. Objetivos

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?

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 on the data that we have available on TripAdvisor.com?
- Q4. What are the topics prioritised and the performance of reviews in the last 2 years in this context?

4. Solución

A continuación se describirán las técnicas elegidas para la consecución de los objetivos marcados. Haciendo notar las razones por las que estas han sido elegidas y describiendo las diferentes fases que ha tenido el proyecto. La fases de esta metodología del proyecto se representa en la Figura 5. El proceso se reparte en 6 Notebooks (Figura 6).

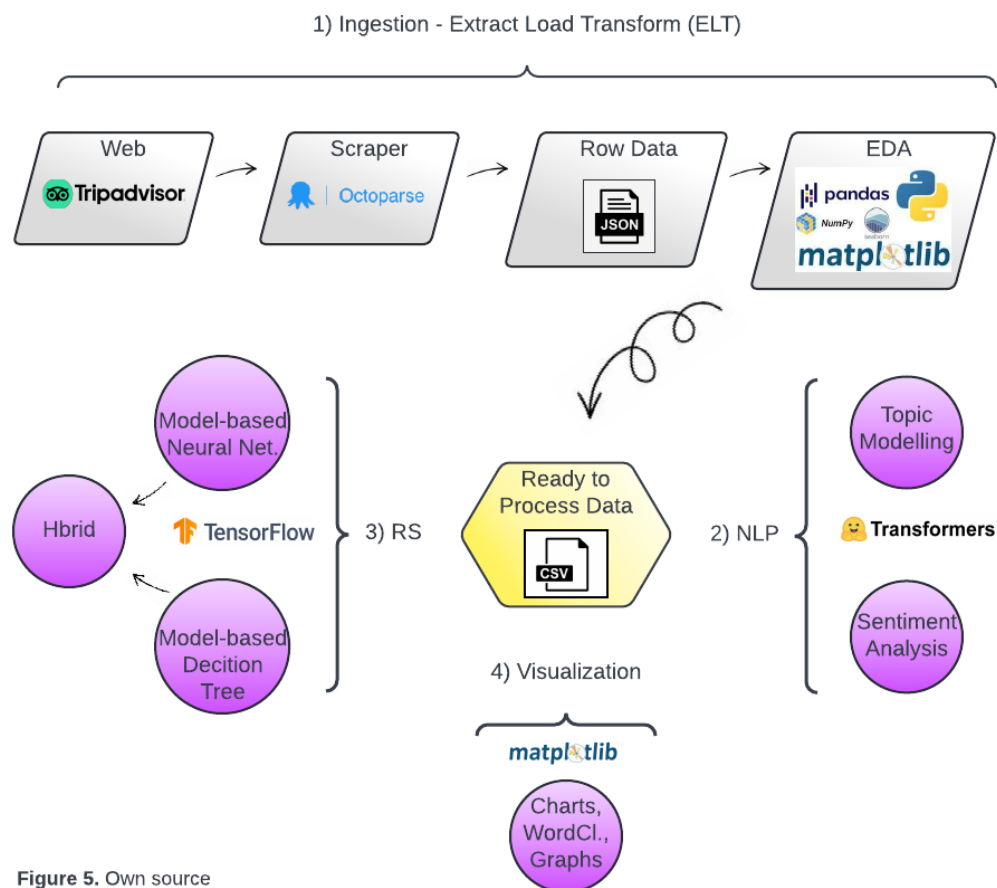


Figure 5. Own source

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.

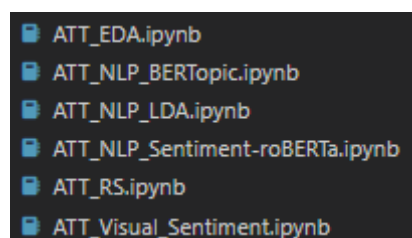


Figure 6. Jupyter Notebooks, Own source

4.1 Fase 1: TripAdvisor data ingestion

TripAdvisor.com presenta todas las condiciones necesarias para realización de NLP y de RS. Tiene un acceso complejo a los datos, raramente se encuentran datos públicos, y cuenta con una API para negocios en donde se pueden acceder a los datos por un costo de 90€ mensuales.

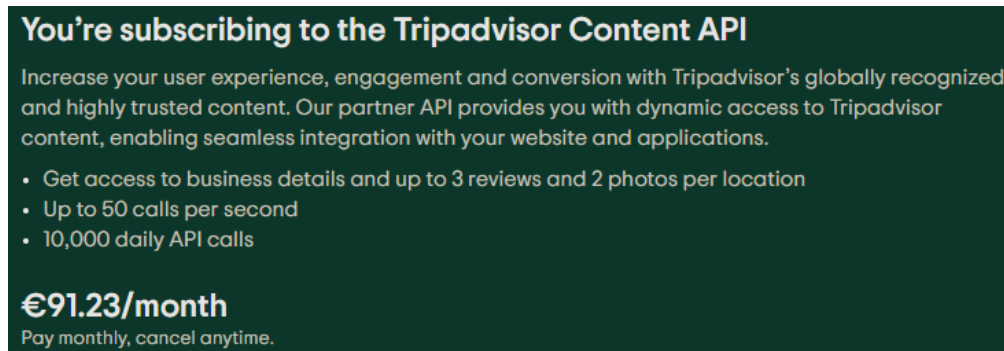


Figure 7. Source: <https://www.tripadvisor.com/developerscheckout>

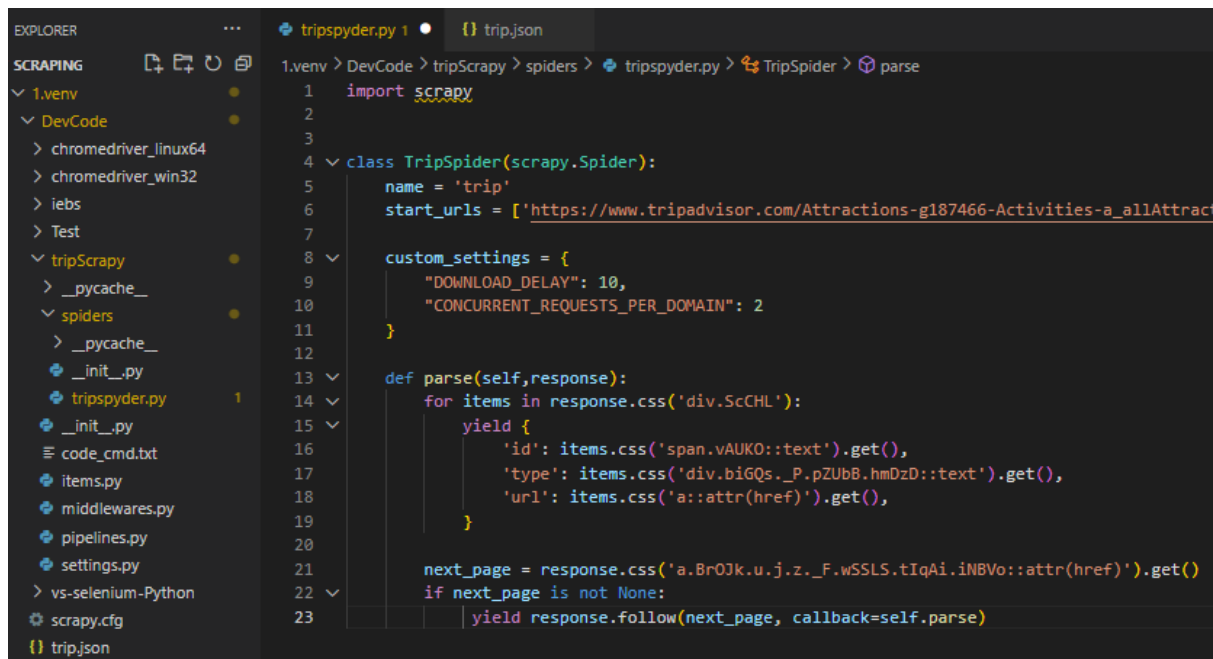
4.1.1 Elección de la muestra

Dada las razones comentadas en el anterior punto 2.1.1, a este estudio le parece pertinente concentrar su muestra en las Islas Canarias de España. Las cuales cuentan con una gran diversidad de relieves, divididos en 8 diferentes territorios/islas: La Palma, El Hierro, La Gomera, Tenerife, Gran Canaria, Fuerteventura, Lanzarote y La Graciosa. Además de ser un punto de turismo masivo de sol y consumo gastronómico y nocturno desde varias décadas, las Islas Canarias de la eterna primavera presentan infinidad de rincones únicos de naturaleza, actividades culturales, historia y gastronomía. Y, siguiendo lo comentado en el punto 2.1.2, específicamente sobre el uso de la ciencia de datos para la gestión sostenible del turismo y los viajes, este trabajo decide indagar en la información que el portal de TripAdvisor puede aportar sobre este destino turístico europeo tan popular, con una economía construida al rededor del turismo, pues sin duda podría aprovechar mejoras en la gestión del turismo, tras la reapertura de las fronteras.

La mayoría de los trabajos académicos publicados sobre TripAdvisor se centran en Alojamientos/hoteles como los ítems a procesar en sus RS y de donde sacar conocimiento, con una evidente razón comercial, sólo en trabajo de Nadezhda (2020) trata Actividades, las cuales tienen también un interés comercial, pues aluyen a tours, restaurantes, casinos y demás comercios. Por ello, este trabajo pretende diferenciarse tratando únicamente Atracciones, las cuales cuentan con libre entrada, o son características hasta el punto de definir el área donde se sitúan (nos referimos, e.g. a landmarks, museos, o a parques acuáticos de reconocimiento internacional).

4.1.2 Procesos de extracción:

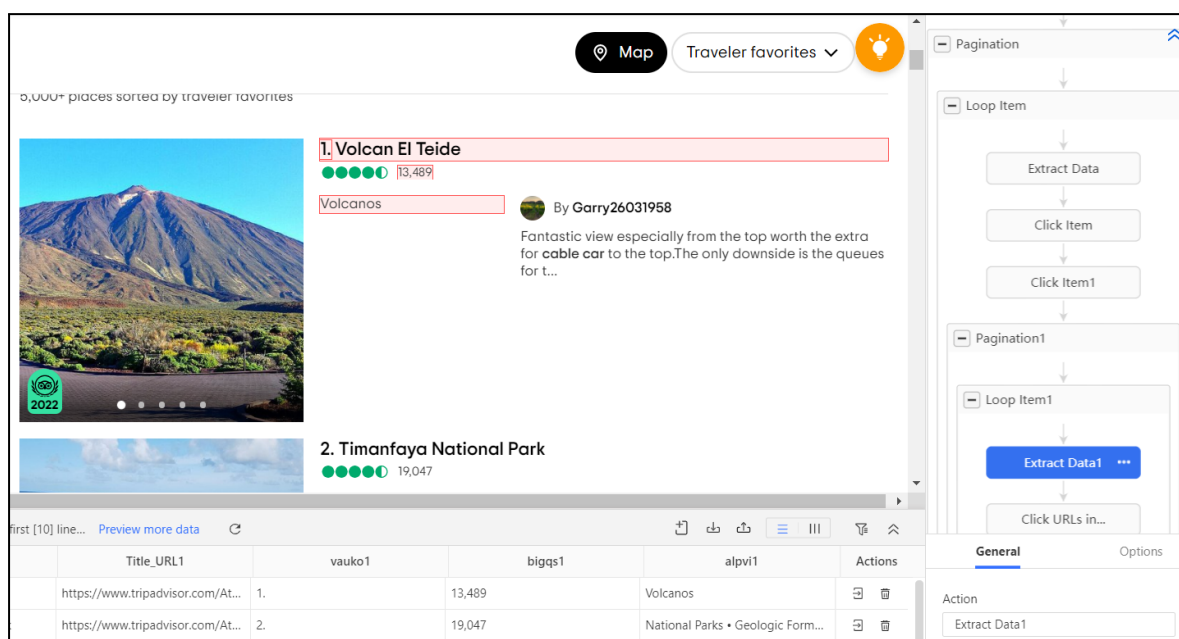
Ante el reto de la obtención de datos y el objetivo de lidiar con problemáticas reales (Q3). En primer lugar, este proyecto desarrolló un entorno virtual local para la realización de scraping, interpretando el html de TripAdvisor.com, con spiders que ejecutaran la ejecución de la recogida de miles de reviews dentro de cientos de Atracciones. También se probó con la librerías *selenium* y *beautifulsoup* y la aplicación *chromedriver*.



```
1.venv > DevCode > tripScrapy > spiders > tripspyder.py > TripSpider > parse
1  import scrapy
2
3
4  class TripSpider(scrapy.Spider):
5      name = 'trip'
6      start_urls = ['https://www.tripadvisor.com/Attractions-g187466-Activities-a_allAttrac
7
8      custom_settings = {
9          "DOWNLOAD_DELAY": 10,
10         "CONCURRENT_REQUESTS_PER_DOMAIN": 2
11     }
12
13     def parse(self, response):
14         for items in response.css('div.ScCHL'):
15             yield {
16                 'id': items.css('span.vAUKO::text').get(),
17                 'type': items.css('div.biGQs._P.pZUb8.hmDzD::text').get(),
18                 'url': items.css('a::attr(href)').get(),
19             }
20
21         next_page = response.css('a.BrOJk.u.j.z._F.wSSLs.tIqAi.iNBVo::attr(href)').get()
22         if next_page is not None:
23             yield response.follow(next_page, callback=self.parse)
```

Figure 8. Scrapy.Spider, Own source

En segundo lugar, se utilizó la API Octoparse:



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

Una vez obtenidos los datos en un formato JSON, los datos fueron procesados por las librerías *pandas*, *numpy* y *matplotlib* en donde se realizaron una numerosa cantidad de correcciones, pues the row data así lo requería en 75 líneas de código (“ATT_EDA.ipynb”).

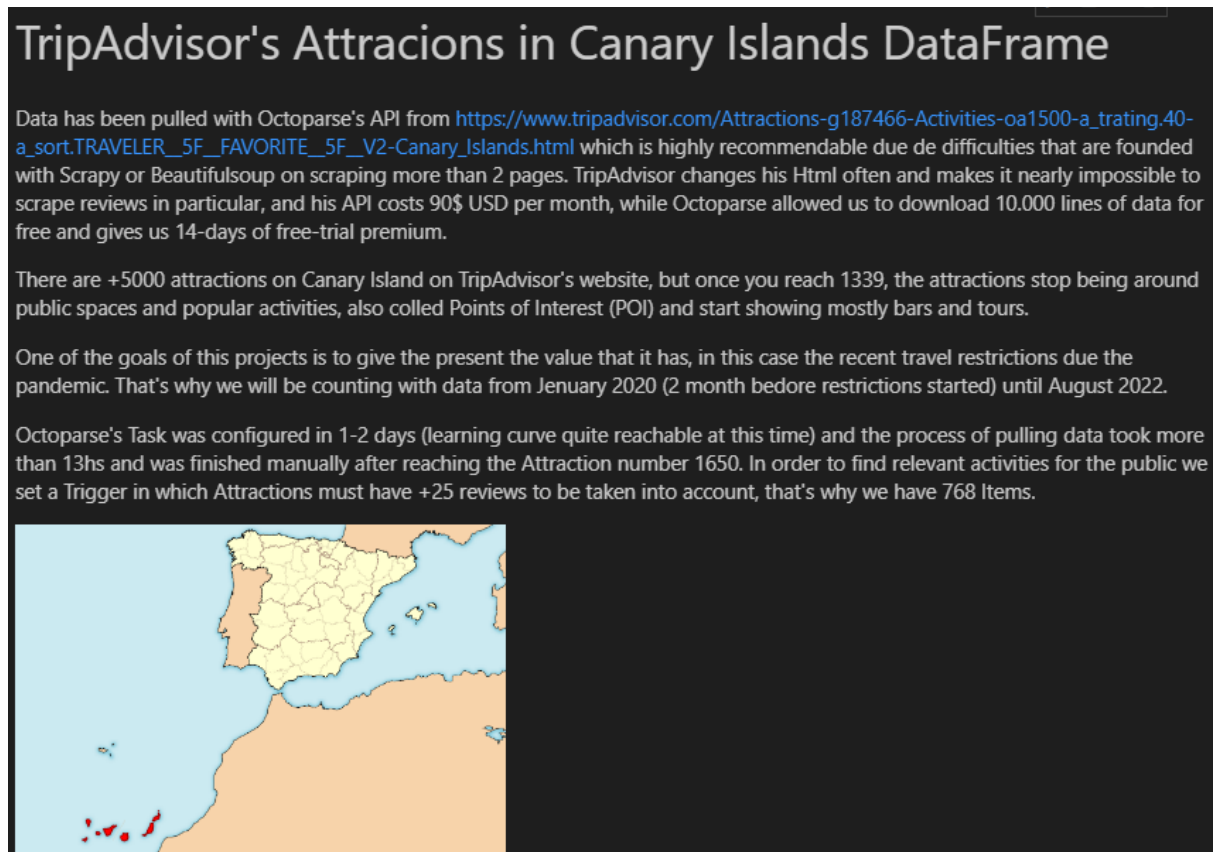


Figure 7. ATT_EDA.ipynb, Own source

Fue en este notebook donde se trataron correcciones de formatos, tipos, datos en columnas mezcladas, combinación de columnas, eliminación o creación de nuevas y múltiples comprobaciones, también a modo de visualización que completaron una limpieza y exhaustiva.

Finalmente, se tendrán en cuenta 9860 registros (cada uno con una review), de más de 6000 usuarios diferentes, sobre más de 650 Attractions/items diferentes categorizados dentro de aproximadamente 150 tipos.

4.2 Fase 2: Natural Language Processing

Aunque el SA a reviews incluido en RS no hayan dado resultado en anteriores investigaciones, no quiere decir que no tenga una gran importancia en la comprensión de los datos. La capacidad de extraer Topics, o de visualizar ese Sentiment aporta perspectiva y herramientas para la mejora y propuesta de soluciones. Por ello, este trabajo, como fue marcado en sus objetivos, pretende aplicar Transformers, entre otras técnicas.

Se han descartado técnicas sin recurrencia para investigación como Name Entity Recognition o Keyphrase Extraction, y se ha apostado por SA con roBERTa, pretrained embedding de TensorFlow creado por el grupo Hugging Face con su origen en el ya comentado Distill-BERT². También, esta investigación propone trabajar con BERTopic para el Topic Modelling y compararlo con LDA pues ha sido utilizado en varias investigaciones. El paquete NLKT y su librería SentimentIntensityAnalyzer será utilizado para el preprocesamiento de texto y la posterior visualización (fase 4) de SA con *matplotlib* y *seaborn*.

¹ <https://huggingface.co/siebert/sentiment-roberta-large-english>

² 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 Fase 3: Recommender Systems

En esta última fase, la investigación se propone la creación de tres RS: los dos primeros serán de tipo Collaborative Filtering y model-based, dado que son los que mejor resultado dan para este tipo de problemáticas según las investigaciones consultadas. La intención de este proyecto es desarrollar un multi-criteria RS al crear, también, un Hybrid RS, el cual consistirá de la combinación de los dos anteriores RS. Además también se realizan visualizaciones de datos (fase 4) aquí.

Investigaciones revisadas en punto 2, sugieren que la inclusión de los resultados del SA a los ratings de los usuarios no crea una mejora en el RS [11]. Por ello, se ha descartado esta técnica.

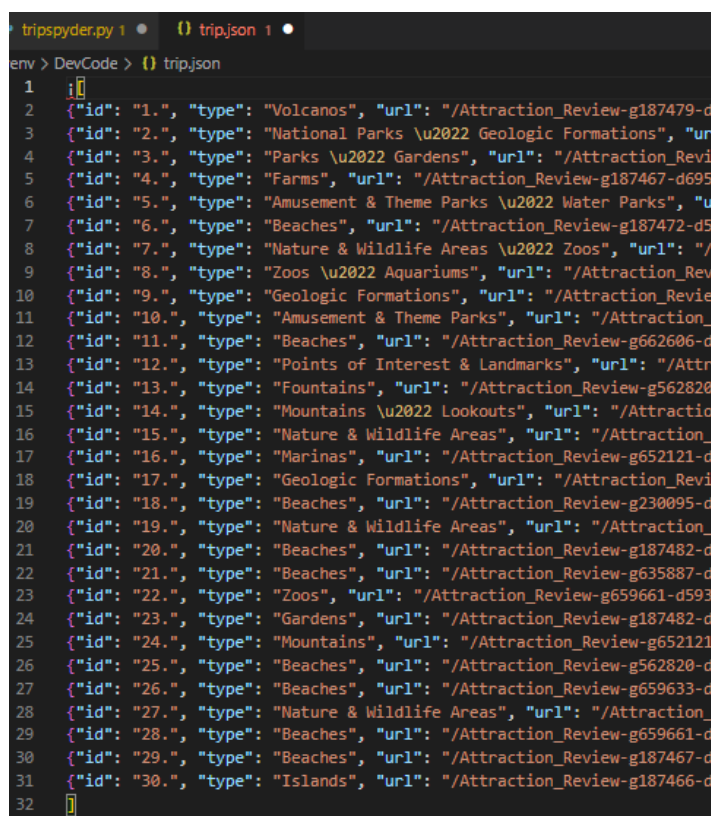
Para trabajar con RS se crearán 3 DataFrames: `user_df`, `item_df` y `rating_df`. Para poder hacer uso de las variables categóricas de `Item_df`, serán procesadas por One Hot Encoding de scikit-learn library. Tras la preparación de los datos, se desarrollarán los dos RS, uno basado en Redes Neuronales (NN) + Nearest Neighbours y otro en Decision Tree + OneHotEncoder + XGBoost_recommender.

5. Evaluación

En este apartado se procede a describir las interpretaciones a la metodología propuesta, los inconvenientes encontrados en el proceso, y la formas en las que se llega a los resultados.

5.1 Evaluación del proceso de ingesta de datos

En primer lugar, la investigación encontró las dificultades anteriormente mencionadas en cuanto a la obtención de datos. En el proceso de scraping con *scrapy*, con la previa creación de un entorno virtual desde el Terminal y ejecutando desde el mismo el script (Figura 8), el código no fue capaz de superar más de dos páginas de extracción de datos, entrando en bucle a partir de la segunda. Y reuniendo únicamente 30 líneas:



```
tripspyder.py 1  {} trip.json 1
env > DevCode > {} trip.json
1  i
2  {"id": "1.", "type": "Volcanos", "url": "/Attraction_Review-g187479-d
3  {"id": "2.", "type": "National Parks \u2022 Geologic Formations", "ur
4  {"id": "3.", "type": "Parks \u2022 Gardens", "url": "/Attraction_Revi
5  {"id": "4.", "type": "Farms", "url": "/Attraction_Review-g187467-d695
6  {"id": "5.", "type": "Amusement & Theme Parks \u2022 Water Parks", "u
7  {"id": "6.", "type": "Beaches", "url": "/Attraction_Review-g187472-d5
8  {"id": "7.", "type": "Nature & Wildlife Areas \u2022 Zoos", "url": "/.
9  {"id": "8.", "type": "Zoos \u2022 Aquariums", "url": "/Attraction_Rev
10 {"id": "9.", "type": "Geologic Formations", "url": "/Attraction_Revie
11 {"id": "10.", "type": "Amusement & Theme Parks", "url": "/Attraction_
12 {"id": "11.", "type": "Beaches", "url": "/Attraction_Review-g662606-d
13 {"id": "12.", "type": "Points of Interest & Landmarks", "url": "/Attr
14 {"id": "13.", "type": "Fountains", "url": "/Attraction_Review-g562820-
15 {"id": "14.", "type": "Mountains \u2022 Lookouts", "url": "/Attractio
16 {"id": "15.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
17 {"id": "16.", "type": "Marinas", "url": "/Attraction_Review-g652121-d
18 {"id": "17.", "type": "Geologic Formations", "url": "/Attraction_Revi
19 {"id": "18.", "type": "Beaches", "url": "/Attraction_Review-g230095-d
20 {"id": "19.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
21 {"id": "20.", "type": "Beaches", "url": "/Attraction_Review-g187482-d
22 {"id": "21.", "type": "Beaches", "url": "/Attraction_Review-g635887-d
23 {"id": "22.", "type": "Zoos", "url": "/Attraction_Review-g659661-d593
24 {"id": "23.", "type": "Gardens", "url": "/Attraction_Review-g187482-d
25 {"id": "24.", "type": "Mountains", "url": "/Attraction_Review-g652121
26 {"id": "25.", "type": "Beaches", "url": "/Attraction_Review-g562820-d
27 {"id": "26.", "type": "Beaches", "url": "/Attraction_Review-g659633-d
28 {"id": "27.", "type": "Nature & Wildlife Areas", "url": "/Attraction_
29 {"id": "28.", "type": "Beaches", "url": "/Attraction_Review-g659661-d
30 {"id": "29.", "type": "Beaches", "url": "/Attraction_Review-g187467-d
31 {"id": "30.", "type": "Islands", "url": "/Attraction_Review-g187466-d
32
```

Figure 11. First JSON, Own source

La construcción del entorno virtual no significó una dificultad, pero sí los intentos de creación de un Spider capaz de recoger los datos. La inversión de tiempo en esta parte del proyecto, unida a que se tuvo que buscar otra solución, será parte del aprendizaje y significó un atraso en las fechas marcadas.

Por otro lado, como ya es explicado en el notebook visualizado en la Figura 7 (EDA_ATT.ipynb), utilizando la API de Octoparse se consiguió acelerar este proceso, no sin cierta dificultad, y con una alta necesidad de realizar limpieza de los datos.

5.2 Evaluación del proceso de NLP

El NLP se ha dividido en tres partes, Topic Modellin con LDA, BERTopic y SA (roBERTa). Todas las partes han tenido un proceso satisfactorio, sin inconvenientes en el procesamiento, siguiendo los paquetes y librerías establecidas. También se han realizado visualizaciones a lo largo del procesamiento como forma de comprensión de los datos, las cuales se adhieren a la fase 4.

En el caso de LDA, este estudio ha podido ver el potencial de esta técnica, utilizando la librería LDAvis para la visualización de los resultados, la cual es interactiva, pudiendo seleccionar las diferentes burbujas de Topics para su análisis en profundidad.

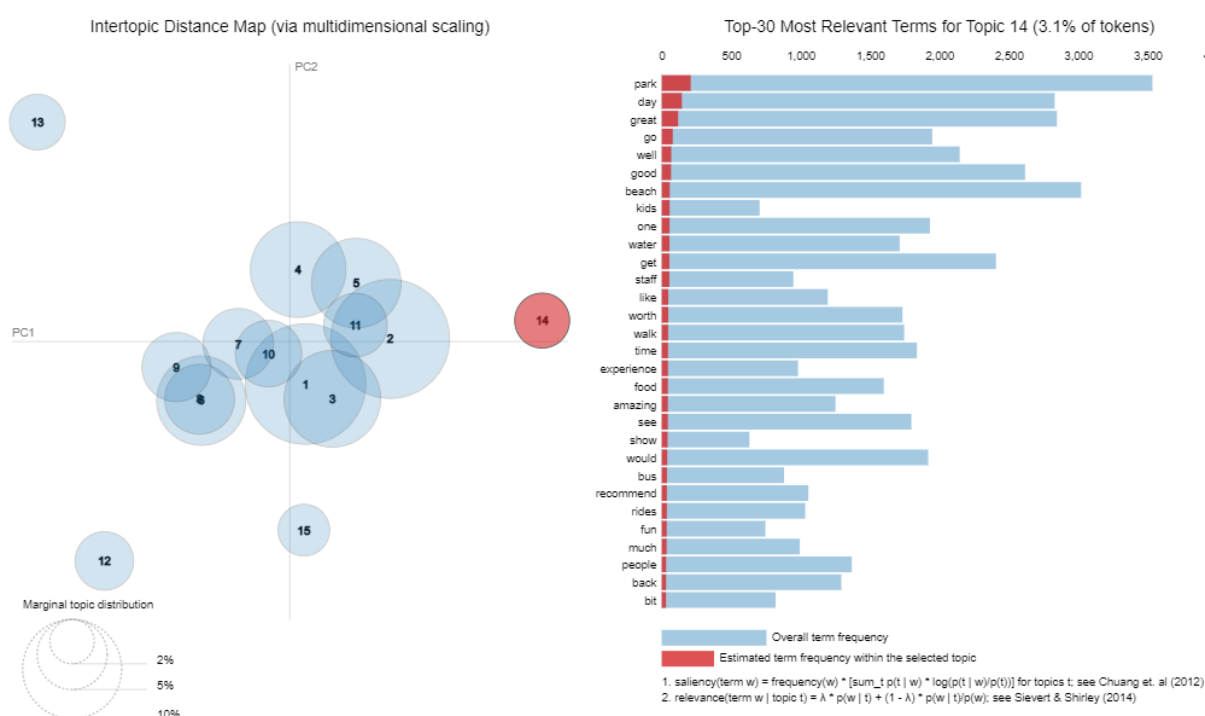


Figure 11. LDA visualization. Own source

Estas visualizaciones nos ha permitido, al igual que en anteriores investigaciones [8], la posibilidad de diferenciar unos Tópicos principales, siendo *Parks* and *Beaches* los contenidos más referenciados e importantes en este análisisi supervisual.

5.3 Evaluación de la creación de RS

La creación de los RS parten de la formación de los tres Dataframes (user_df, item_df, rating_df). Dado que la extracción de datos y su posterior transformación para su uso fue exitosa, no hubo inconveniente a la hora de formarlos. Seguidamente se realizaron visualizaciones para la comprensión de los datos.

```
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

La validación del primer RS se ha realizado con BinaryCrossentropy(), dando una pérdida de 0.36 en entrenamiento y 0.59 en validación, en el mejor de los casos, dado que en cada ejecución se obtienen diferentes resultados. Aquí, se distingue una necesidad de procesar más datos, para un mejor entrenamiento y un RS con menos variación en cada ejecución. Al mismo tiempo, la validación de XGBoost se pretendió realizar con el RMSE (Root mean square error) lo cual no se consiguió realizar por errores en el scripting. Confirmando la necesidad de mejoras en este proceso y en la adaptación de los datos.

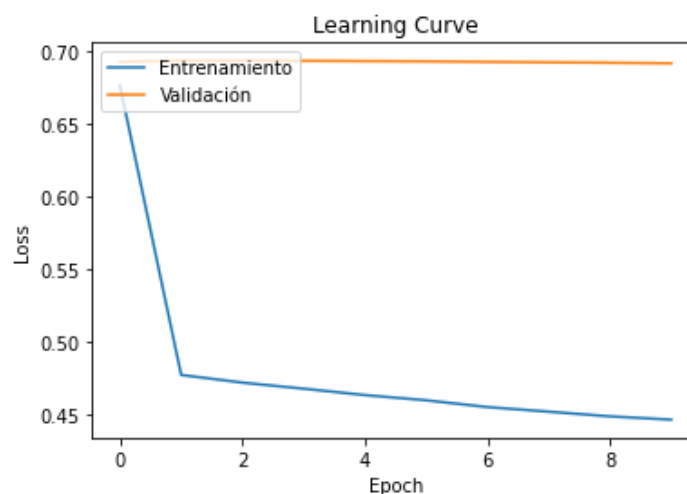


Figure 13. Neural Network Learning Curve, Own source

5.4 Evaluación de visualización de datos

El proceso de visualización fue efectivo en todas las fases, se desplegaron multitud de gráficas, charts y Wordclouds que invitaron a la comprensión de los datos y a la corrección de los mismos desde la fase 1 (ATT EDA.ipynb).

En el EDA pudimos distinguir particularidades de nuestros datos, como la polaridad en los datos de cantidad de contribuciones que un usuario a realizado en TripAdvisor (número total de reviews), habiendo un gran grueso con decenas de reviews, pero algunos usuarios con 5.000, 10.000 o hasta 60.000. También diferenciamos la proporción de ratings que tienen el total de ítems de nuestro dataset.

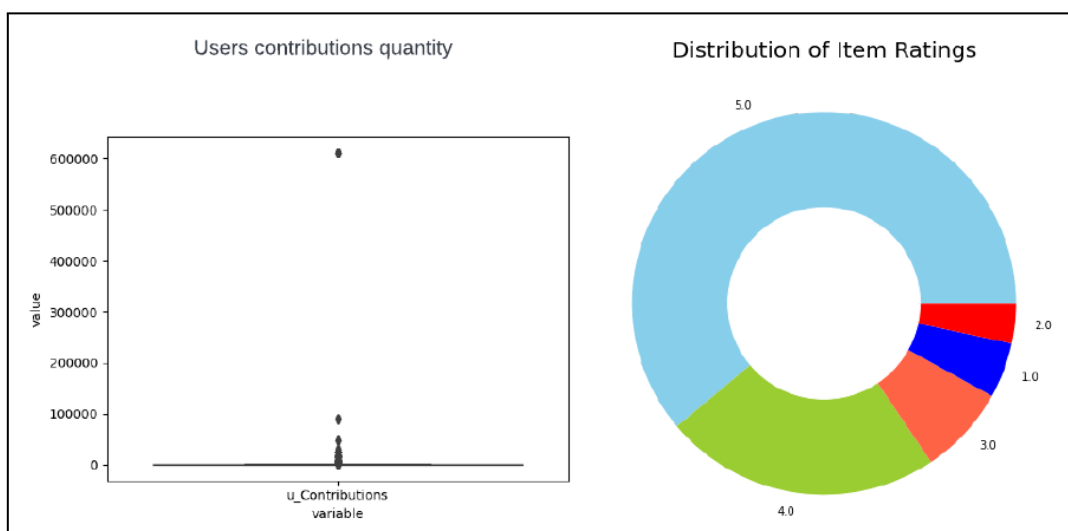


Figure 14. EDA visual evaluation. Own source

En la fase de NLP, se extrajeron diferentes visualizaciones. En el procesamiento de Topic Modelling, además de una similarity matrix entre ítems (Appendix 2), se desplegó la siguiente Wordcloud que muestra una relación de la importancia y recurrencia del uso de palabras en reviews según tamaños y colores:

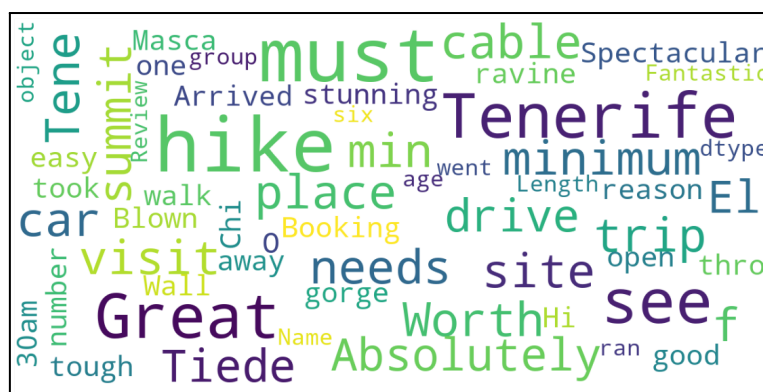


Figure 15. BERTopic visual evaluation, Own source

También en el procesamiento de NLP, en específico de SA con roBERTa, pudimos evaluar la densidad de las reviews, en cantidades de caracteres, dependiendo del rating que los usuarios daban a las Atracciones/items.

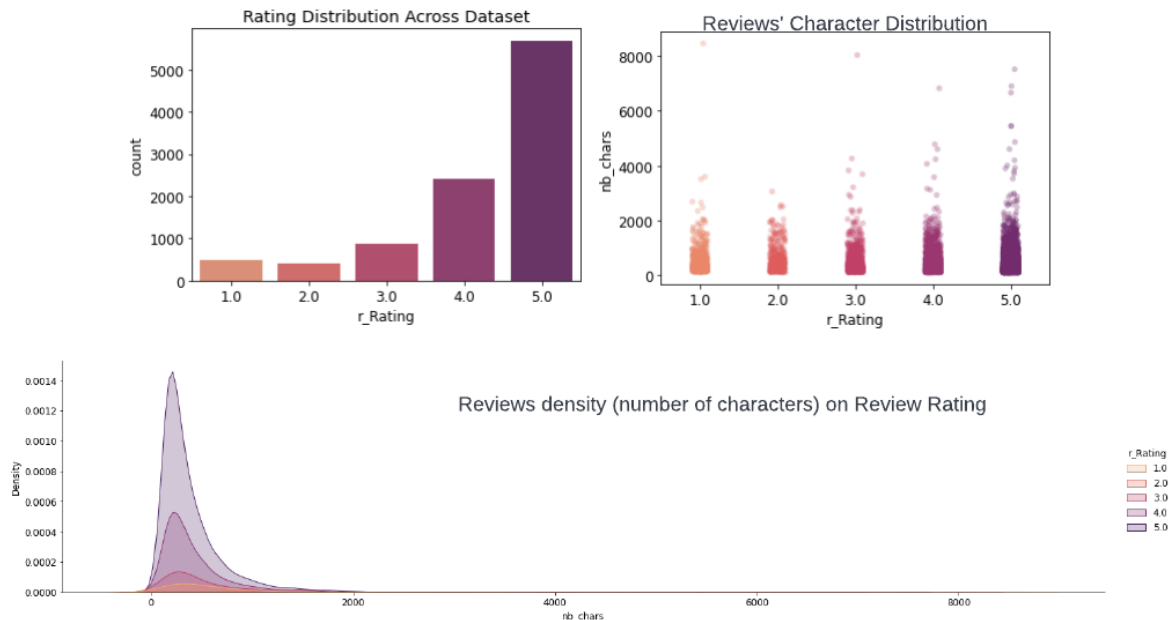


Figure 16. roBERTa visual evaluation, Own source

Por último, con en el Notebook ATT_Visual_Sentiment.ipynb, dedicado específicamente a la impresión de visualizaciones pertinentes a la investigación, pudimos interpretar variables como la de *tipo de viaje-u_Trip* (sólo, acompañado y de qué manera, o desconocido) así como los porcentajes de usuarios por grupo de edad. Se destaca que, de los usuarios que hay información, la mayoría viajan en parejas o en grupo de amigos, es un pequeño porcentaje el que viaja sólo (y realiza reviews en Tripadvisor). También se despliega un gráfico de barras en el que se diferencian los ítems con más reviews (Appendix 3): beaches, Landmarks, Museums, Malls and Parks lideran la lista.

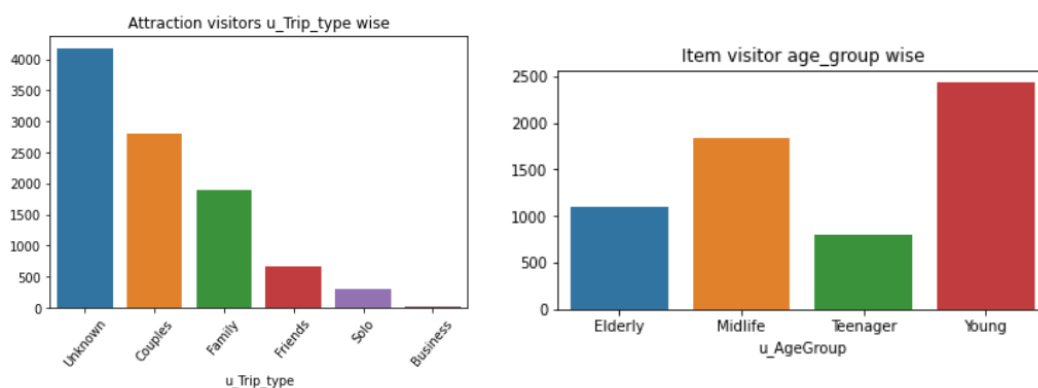


Figure 17. Visual_Sentiment, Own source

6. Resultados

En este capítulo se analizarán los resultados obtenidos con las diferentes técnicas utilizadas. Se pretende explicar la consecución o no de los objetivos y los descubrimientos del trabajo. El resultado de todas las exploraciones realizadas se despliegan en: 2 JSON con los datos originales obtenidos mediante el scraping con Octoparse, 6 notebooks con el código en Python, y 3 archivos csv con los Dataframes limpios y listos para ser procesados. Esta investigación ha publicado el código procesado en el portal Github [13], para que otros se aprovechen de los datos recolectados y les técnicas utilizadas, y así aportar a la comunidad de Científicos de Datos y el tratamiento del Big Data.

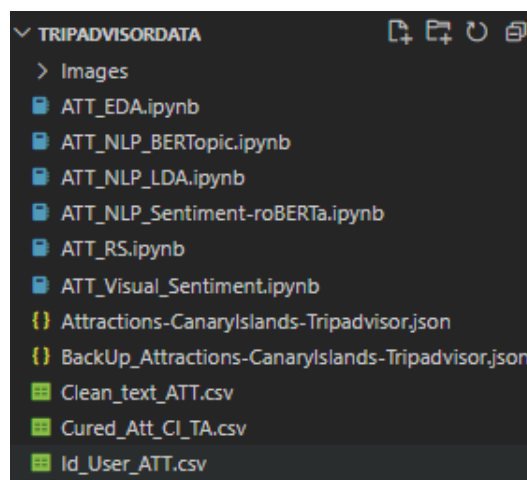


Figure 18. Investigation files, Own source

Esta investigación partió con un objetivo académico orientado hacia la ingestión de datos con herramientas open source y su tratamiento con NLP, y con otro business-oriented basado en la búsqueda de oportunidades mediante la creación de RSs y la interpretación del comportamiento de los datos desde el comienzo de la pandemia Covid-19 hasta el día de hoy. Los siguientes apartados responden a las cuatro preguntas que esta investigación se planteó dentro de esos objetivos.

- 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?
- Q3. Can we recommend new content/Attractions based on the data that we have available on TripAdvisor.com?
- 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 presentó severas dificultades para la obtención de datos con herramientas open-source. Aunque con la librería *scrapy* se pueden conseguir precisos y cómodos resultados en la mayor parte de websites, en este caso no fue posible obtener los datos de esta manera. Tampoco con *beautifulsup* y *selenium*. Lo que sí fue efectivo y con una curva de aprendizaje corta, fue la herramienta Octoparse. Con esta aplicación, el usuario puede enseñar al spider dónde hacer click y dónde recoger datos sin tener que interpretar html en gran parte del proceso. La recolección de los datos se puede hacer en una máquina local (siempre y cuando sean menos de 10.000 registros), o en la propia nube de Octoparse con una cuenta Premium (registros ilimitados).

La recolección de datos se paró manualmente tras 13 horas, al haber sido procesadas al rededor de las 1.650 Atracciones (que contaran con más de 25 reviews, por motivos de fiabilidad). Así, se recogieron 12460 registros sobre 768 items/Attractions. Cada registro contó con información de Usuarios (u_), Reviews (r_) y Items (i_). Se define este proceso como ELT (Extract, Load, Transform), pues no hy ingreso de datos en streaming (añadiendo registros a diario), por lo que no hubo necesidad de creación de una base de datos. Fueron Loades en Local y luego Transformados en los Notebooks con Python.

Como se puede ver en la Figura 19, había imperfecciones en los datos, con varias columnas mostrando valores NaN, y otras mostrando la información en la columna equivocada. Esto, a simple vista, mediante el Exploratory Data Analysis (EDA) se encontraron otras variables que necesitaron múltiples correcciones.

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	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

Tras una limpieza y reducción del dataframe un 21% en el archivo ATT_EDA.ipynb se creó una copia del dataframe en el archivo Curet_ATT_CI_TA.csv. Finalmente, el resultado del dataframe se visualiza en la Figura 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)

Respondiendo a la Q2, se puede asegurar que el preprocesamiento de los datos fue adecuado para su posterior uso con state-of-the-art techniques. Este estudio procesó reviews de TripAdvisor con LDA y con BERTopic (en cuanto a Topic Modelling), así como con roBERTa (en cuanto a SA) de forma satisfactoria. Los resultados del Topic Modelling servirán para el reporte que responda la 14, también las visualizaciones del SA. Una muestra de los resultados de SA se puede ver a continuación (Figura 21), donde se puede diferenciar a grandes rasgos que los resultados son adecuados.

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

Estas aplicaciones se pueden consultar en los notebooks: ATT_NLP_LDA.ipynb, ATT_NLP_BERTopic.ipynb y ATT_NLP_Sentiment-roBERTa.ipynb.

6.3 Exploring Collaborative Filtering (Q3)

La tercera pregunta de investigación, corresponde al lado business. Con esto la investigación pretende, desarrollar herramientas y exponer informaciones que puedan ser útiles para OTAs, DMOs y para Travel Sustainability. Por ello, la creación de RSs, como un nuevo feature para las Atracciones en Canary Islands se presentó como una solución.

La solución propuesta de desarrollar dos RS de tipo Collaborative Filtering Model-based, se realizó sin inconvenientes, los datos pudieron procesarse por lo que tuvieron un preprocesamiento adecuado. La investigación consiguió desarrollar un RS con Neural Networks y otro de Decition Trees (XGBoost) que tieron resultados adecuados:

	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 Masp...
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

Por otro lado, se encontraron dos inconvenientes: en primer lugar, la curva de aprendizaje del modelo nn_recomender muestra niveles elevados y, por tanto, insuficientes. Por otr lado, el rmse no se pudo computar para la comprobación del modelo rmse. Se puede afirmar pues, que la investigación pudo desarrollar RS adecuados con los datos que cuenta, pero debe mejorar su nivel de fiabilidad, mediante la consecución de una mayor cantidad de datos, un mejor procisamiento de los modelos u otros modelos que devuelvan mejores resultados. En segundo lugar, se presentaba como una solución más potente y recurrente la creación de un modelo híbrido, que uniera los dos RS desarrollados. Esta solución no ha podido desarrollarse satisfactoriamente, por lo que se confirma la necesidad de mejoras en este punto del proyecto.

6.4 Topic Modelling and Data Visualisation report (Q4)

El último resultado de esta investigación tiene que ver con dar al contexto la importancia que tiene, en tener en cuenta la posición socioeconómica que rodea al territorio al que aluye esta investigación. Así mismo, para este reporte, se interpretan los datos de TripAdvisor como válidos para hacer juicios, en términos generales, sobre el turismo en Canarias. Pues como se destaca en diferentes investigaciones, el e-Travel es una forma de viaje que gana importancia año a año por su repercusión en los negocios y los viajeros que lo utilizan. Por lo que podemos diferenciar el viajero que utiliza estas plataformas como un tipo de turista deseado, bien sea por la posibilidad de analizar su opinión, o por que se le caracteriza como un consumidor de Atracciones y un turista activo.

Así pues, como se comenta en el punto 2.1.1 *La Industria Turística en Canarias*, la cantidad de empleo y de PIB que genera esta industria se vio disminuída al rededor de un 65%. Estos datos, se ven reflejados también en la cantidad de reviews que se han realizado. En la Figura 23 se pueden apreciar dos recesos, uno a mediados de marzo de 2021 (comienzo de las restricciones y cierre de las fronteras) y otro a mediados de enero (segundo paquete de medidas altamente restrictivas hacia las libertades de los ciudadanos), coincidiendo con las políticas de los gobiernos para reducir los niveles de contagios del virus Covid-19. También se diferencia una mejora progresiva a partir del verano de 2021 (medidas de reducción de libertades se empezaron a reducir), y que a fechas de Agosto de 2022, la cantidad de reviews en las Islas Canarias en TripAdvisor significa un 40% de lo que era antes de la pandemia. Estos datos demuestran el acierto en la contextualización de la problemática.

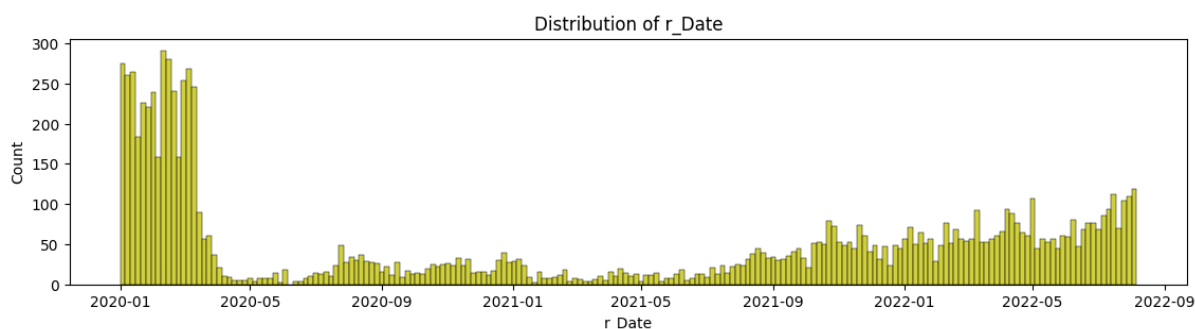


Figure 22. Number of reviews per time, Own source

Adentrándonos más en la forma en que estas reviews están distribuidas por perfil de viaje, ya vimos como el tipo de viaje que se realiza con más frecuencia es en familia, luego en grupo de amigos, luego en pareja y, por último, sólo (Figura 17). Si tomáramos estos datos como un censo de viaje, se podría confirmar que la forma de viaje preferida a las Islas

Canarias es en grupo, y que las Atracciones de libre entrada o de interés cultural que se presentan y se pueden realizar son más aprovechadas en familia, que en pareja o sólo. También, desde una perspectiva de viaje sostenible, hay cifras abrumadoras en cuanto a la distribución del turismo en Canarias. Es sorprendente la desigual distribución de registros en las diferentes islas:

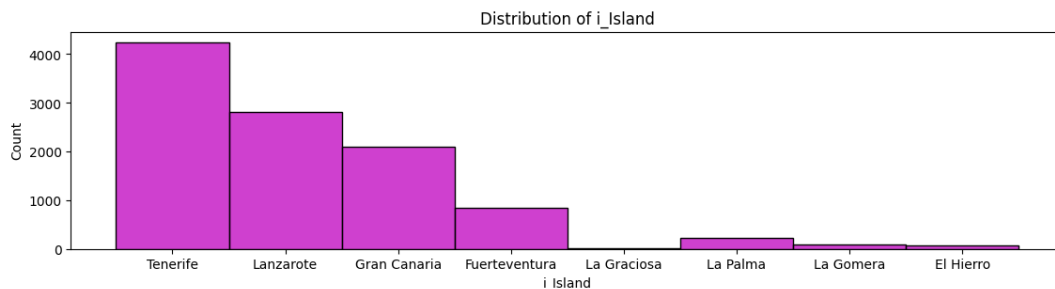


Figure 23. Distribution of reviews per island, Own source

En cuanto a los conocimientos extraídos del Topic Modelling, las soluciones propuestas muestran resultados interesting findings. Como veíamos en la figura 11 sobre LDA, podemos destacar difernetes tópicos y su relación con los demás, donde destaca una fuerte relación de las palabras Park y Beach en todos los tópicos. En el caso de BERTopic, además de las relaciones que se visualizan en la similarity matrix (Appendix 2), el Topic Word Scores destaca Parks, Plants, Museum, Market, Dunes, Church, Golf and Marina:



Figure 24. Own source

Además, otra potente feature de BERTopic es el Hierarchical Clustering, donde se pueden ver las siguientes relaciones dentro de los mismos tres grandes 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. Conclusión y trabajos futuros

La investigación se ha enfrentado con éxito a los diferentes retos planteados en sus cuatro fases. En primer lugar, la consecución de datos de TripAdvisor, plataforma que cuenta con una API de pago para la liberación de información, a la vez que con limitantes a la hora de realizar scraping. Seguidamente, el uso de técnicas de NLP con librerías como *nltk*, *pyLDAvis*, *gensim* y transformers como *roBERTa* y *BERTopic*. En tercer lugar, la construcción de RSs de tipo *Collaborative Filtering model-based* con TensorFlow. Y, por último, la realización de un reporte con visualizaciones para la comprensión de los datos, a lo que se suman las librerías *seaborn* y *matplotlib*.

Con esto tenido en cuenta, de la primera fase se concluye que la herramienta que ha dado mejores resultados para la obtención de datos de TripAdvisor es Octoparse, debido a sus facilidades y capacidades. Facilitando el scraping en comparación con herramientas como Scrapy o BeautifulSoup o Selenium. La segunda conclusión es que los Topics que destacan por su uso y relación entre la totalidad de las reviews son Beach and Parks, según LDA y Parks, Plants, Museum, Market, Dunes y Church según BERTopic, dando ambos resultados útiles para la interpretación de los datos. También destaca la pertinencia mostrada por el Hierarchical Clustering de BERTopic. Así mismo, se concluye que la utilización de RSs de Collaborative Filtering model-based para con los datos de TripAdvisor es adecuada, pero que es necesaria una revisión con mayor cantidad de datos o mejoras en los modelos para conseguir mejores resultados, así como la creación de un Hybrid RS.

Por último, los resultados de todas las técnicas aplicadas, la visualización de resultados y el reporte realizado, invita a la aceptación de los datos de TripAdvisor como válidos y representativos. Con sus reviews los usuarios aportan al crecimiento del e-Traveling, y fomentan el turismo activo y consciente. Del mismo modo, se evidencia una posibilidad de redistribución de la cantidad de turismo por cada isla, debido a la disparidad entre las mismas. También se concluye que a la fecha de Agosto de 2022, las aportaciones de usuarios de TripAdvisor en Attractions de las Islas Canarias se aproxima al 40% de lo que era antes de la pandemia Covid-19. Evidenciando un gran espacio de recuperación turística, y con ello, de oportunidad de reconstrucción

Como trabajo futuro, se plantea la exploración de nuevos modelos de RSs junto con la inyección de datos en forma de streaming con su correspondiente pipeline o sistema de flujo de datos, para el desarrollo de un RS comerciable con fines de Travel Sustainability.

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., Soulaïmane, 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., Subramaniaswamy, 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:



Q Search reviews...

Filters English Most Recent ¹¹

Popular mentions

cable car national park warm clothes highest point half day trip
views are spectacular rock formations worth the trip out of this world
visiting tenerife summit permit clouds volcano landscape teide

Inspire786516
2 contributions

0

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

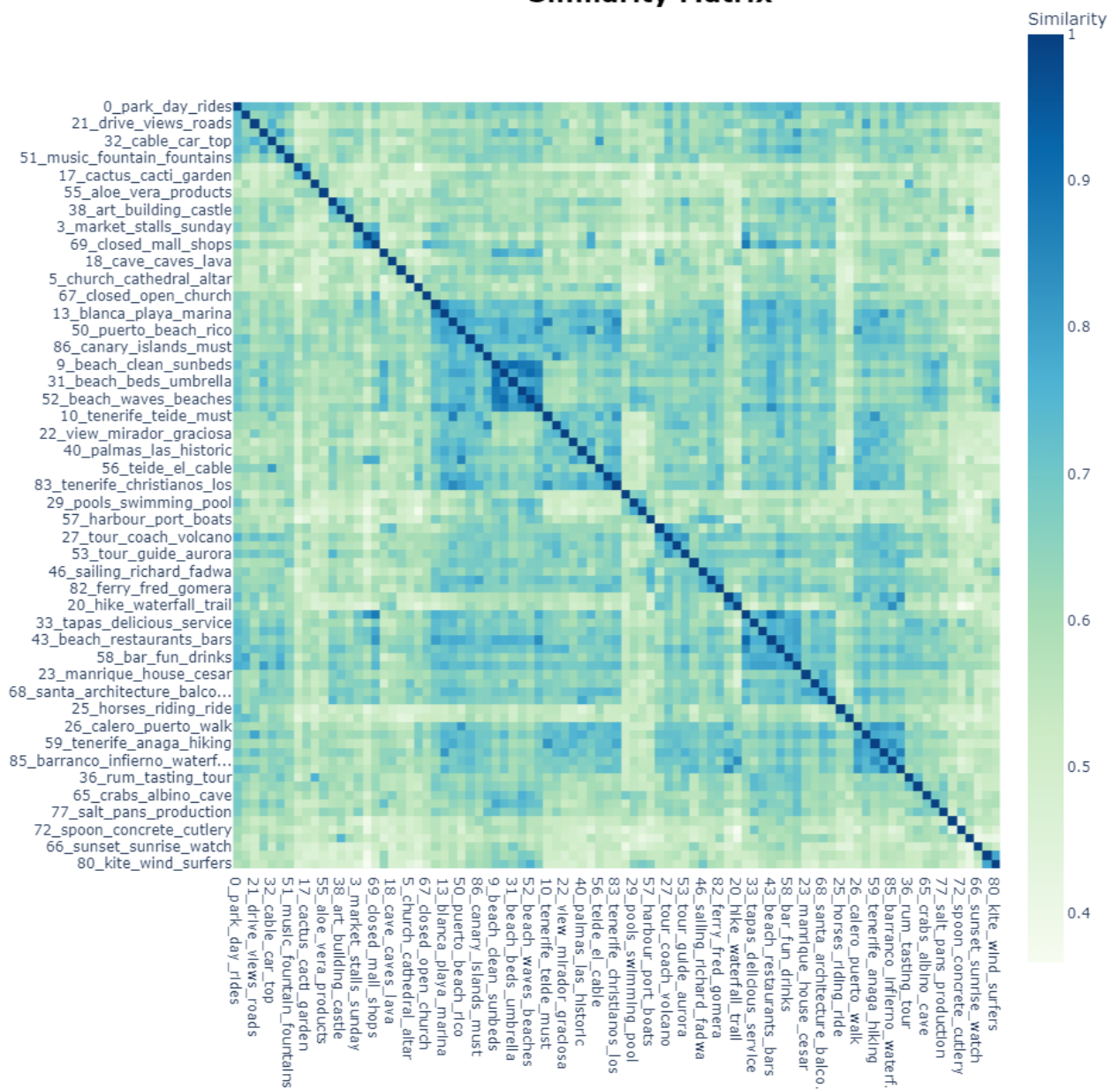
I went up to the Teide National Park as part of a half day coach tour. This took us to the base of the volcano but as the cable car was closed due to a technical issue I could not get near the top. Still amazing views in the area of the remains of previous volcano eruptions. I would suggest that if you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

Written September 16, 2022

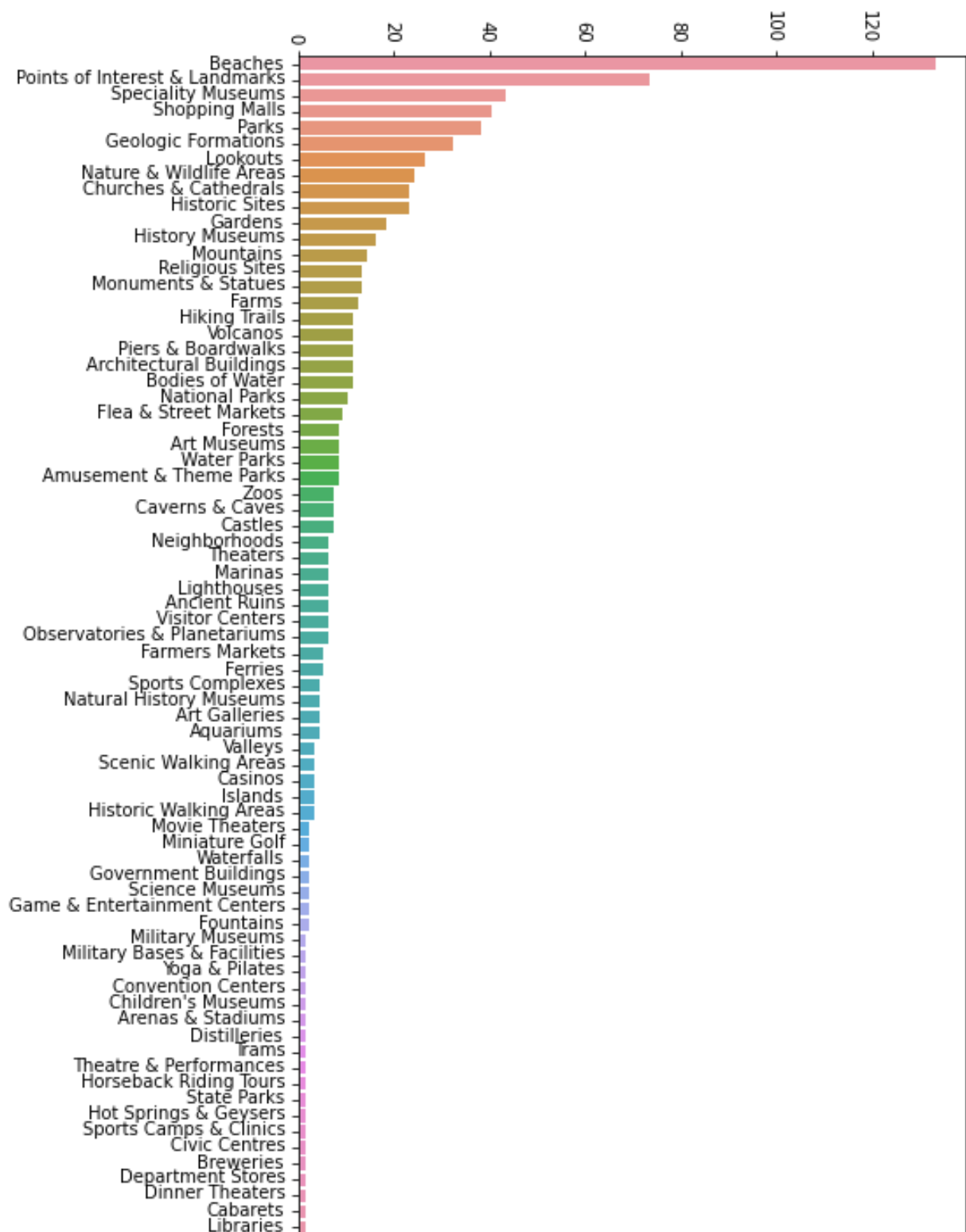
This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews.

Appendix 2:

Similarity Matrix



Appendix 3:



Appendix 4:

Hierarchical Clustering

