A picture containing logo

Description automatically generated

**Team 6 Final Capstone Project – European Vacation Hotel Recommender**

Presented by: Addis Ashenafi, Carlos Capado, Neil Ralston

**Introduction:**

After 26 weeks of Bootcamp, we are ready to travel abroad for some rest and relaxation! A weekend trip to Europe sounds refreshing – but selecting a hotel seems daunting as there are so many to choose from!

Never fear, your cohorts on Team 6 have developed an application to help simplify the hotel selection process. Using a [Kaggle dataset](https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe) containing over 500,000 customer reviews of nearly 1,500 luxury hotels across Europe, our application processes customer review comments to recommend other hotels with similar reviews.

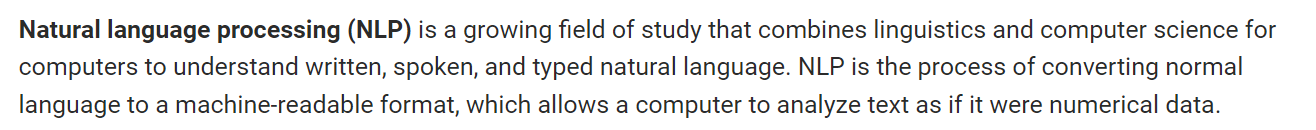
Our inspiration for this project came from two primary sources: the [Goodreads book recommender](https://github.com/alvessab/project_4_book_rec) application developed by a predecessor bootcamp team and an example hotel recommendation system article provided on the [Clever Programmer](https://thecleverprogrammer.com/2021/02/13/hotel-recommendation-system-with-machine-learning/) website.

The expected behavior from our application is that with a user-defined input – a target hotel name – similar hotels based on customer review comments will be returned.

**Methodology:**

Our application uses a Natural Language Processing (NLP) machine learning model to process customer review comments and determine similarity patterns within a content-based model.

According to our Bootcamp course module, NLP is defined as:



The process breaking NLP down into a series of smaller, less-complex tasks is referred to as the “NLP Pipeline”. The NLP Pipeline from the course module is shown below:

Diagram

Description automatically generated

**Raw Text:**

Due to the large file size of the raw dataset (over 230mb), it was uploaded to an Amazon S3 bucket.

The raw dataset included the following columns:

Table

Description automatically generated

At the outset of the project, the following adjustments were made to clean the raw dataset:

* Converted Review\_Date to a date/time format
* Dropped rows missing latitude (lat) and longitude (lng) coordinate data
  + Fewer than one percent of rows were dropped
* Replaced the country name “United Kingdom” with “UK”
* Added a new column containing the hotel country only (as extracted from the hotel address)
* Added a new column with overall rating sentiment based on binned review scores
* Replaced “No Negative” and “No Positive” review comments with “None”

The cleaned dataset advanced into the next phase of the NLP Pipeline included the following columns:

Text

Description automatically generated with medium confidence

**Tokenization and Stop Word Filtering:**

The following adjustments were made to the cleaned dataset before proceeding with tokenizing and filtering:

* Positive and negative review comments were combined into one column named “All\_Reviews”.
* Only the first review comment for each hotel was kept; all duplicate instances of a hotel name were dropped. This greatly simplified the dataset for use in downstream NLP analysis tasks.

The next step was to extract keywords from the combined “All\_Reviews” column using the Rapid Automatic Keyword Extraction (RAKE) from the NLTK library. This process removed stopwords (words that have little or no linguistic value), white space, and punctuation from the text in the “All\_Reviews” column – leaving a clean set of keywords.

To prepare for final processing, the keywords were further refined as follows:

* Tokenizing: separating the words from paragraphs or sentences, into individual words.
* Stemming: a technique used to extract the base form of the words by removing affixes from them.
  + Example: the stem of the words eating, eats, and eaten is eat.
* Lemmatizing: the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.
  + Example: runs, running, ran are all forms of the word run, therefore run is the lemma of all these words

A sample of the key word development process is shown below:

Table

Description automatically generated

**Term Frequency/Inverse Document Frequency (TF-IDF) Processing:**

TF-IDF is a statistical weight showing the importance of a word in a document. The TF-IDF algorithm uses a relatively simple but intuitive approach to weighting words, making it a great starting point for our application.

**Term frequency (TF)** measures the frequency of a word occurring in a document, while **inverse document frequency (IDF)** measures the significance of a word across a set of documents.

Multiplying these two numbers determines the TF-IDF score. The following explanation is provided in the module:

Text

Description automatically generated

Next, the data needs to be converted into a numerical vector. TF-IDF vectorization is accomplished by determining the TF-IDF score for each word in a dataset, and then transforming the data into a vector. With these vectors, similarity between datasets can be determined by comparing TF-IDF vectors using cosine similarity.

In our application, cosine similarity is used to measure the similarity of the TF-IDF scores of keywords derived from the hotel review comments.

The resulting cosine similarity matrix is shown below:

A picture containing table

Description automatically generated

The following link provides access to an excellent article that explains further details about TF-IDF:

<https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>

**Machine Learning:**

Using the following code, our content filtering recommender model uses the TF-IDF scores and resulting cosine similarity calculations to determine which hotels have comparable review comments to a target selection.

Text

Description automatically generated

For example, Senator Hotel Vienna is the selected target, the hotel with the highest review similarity score (with a minimum user rating of 8.0) is the Guitart Grand Passage.

Graphical user interface, text, application

Description automatically generated

Although relatively straightforward and simple to use, our TF-IDF model has some inherent drawbacks – an inability to consider the semantic context of words, for example.

**Website Development:**

The cosine similarity matrix was saved as a NumPy array (.npy) in an AWS S3 bucket and accessed via s3fs, which provides a direct mapping of S3 to a local filesystem.

**Visualization Development:**

One of the most important aspects of data analysis is being able to tell a story through data visualizations. As such, we created several dashboards which assist us in telling a story for users trying to find the perfect hotel to stay at in Europe. In the first illustration, we can see all the hotels in Europe which have been rated. Through the filters, it is easy to parse out hotels by rating by average score.

Map

Description automatically generated

In the following dashboards we visualized top rated hotels in individual Countries throughout Europe which greatly help narrow down the search. Additionally, the heat density maps are equipped with filters to be able to search by average score as well as a more detailed individual score. There were of 6 individual heat density maps created for this project. An example heat density map is shown below.

Map

Description automatically generated

In the last dashboard, we visualized the data via bar graph to display the top-rated country for hotel stays based on the average score; additionally, we also visualized the top hotel overall in Europe based on average reviews.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated