**CCT College Du CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| **Assessment Title:** | CA 2 |
| **Lecturer Name:** | David McQuaid, Muhammad Iqbal, Saw Weiss, Taufique Ahmed |
| **Student Full Name:** | Lylkend Silva Albuquerque |
| **Student Number:** | 2023290 |
| **Assessment Due Date:** | 07/01/2024 |
| **Date of Submission:** | 07/01/2024 |

**Declaration**

|  |
| --- |
| **By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.** |

**MSC\_DA\_CA2**

Author: L. Albuquerque

e-mail: [2023290@student.cct.ie](mailto:2023290@student.cct.ie)

Student ID: 2023290

Dublin versus Hong Kong air traffic

Acknowledgements

I would like to thank the following people who have helped me undertake this research:

My Professor David McQuaid, for his enthusiasm in all of our classes, for his support, encouragement and patience and of course to all the others professors who strive to put us on the path of knowledge and development such as:

Sam Weiss

Dr. Muhammad Iqbal

Taufique Ahmed

The CCT college, for input throughout this MSc programme.

All my colleagues which in their way had tried to give some contributions to data collection and some expertise was given in our class group.

To the Central Statistics Office (CSO) and DATA.GOV.HK is an online web interface provided by the Office of the Government of the Hong Kong Special Administrative Region, who concisely made available all the databases in which it was possible to develop this research.

My mother Maria do Socorro who almost every day left me a message of encouragement, one of which is always in his words "There is no glory without sacrifice", special thanks.

Abstract

*This research work synthesizes data up to 2023 from the Central Statistics Office (CSO) of Ireland and data.gov.hk of the Hong Kong Government, focusing on the air transport numbers between Ireland and Hong Kong. Intriguingly, the data reveals a lack of correlation between air transport metrics and broader economic indicators in the two regions.*

*Despite the conventional belief that air transport figures are closely aligned with economic performance, the analysis of data from Ireland and Hong Kong indicates a decoupling of these variables. Passenger and cargo volumes, along with flight frequencies, exhibit trends that do not consistently mirror the economic growth patterns, trade volumes, or other financial indices in either region.*

*In Ireland, the CSO data highlights fluctuations in air transport numbers, including passenger counts and cargo loads, which do not parallel the country’s economic growth trajectory. Similarly, data from Hong Kong’s data.gov.hk shows air transport metrics evolving independently of the region's economic fluctuations.*

*This phenomenon raises questions about the factors influencing air transport trends in these regions. The abstract explores various potential influences, such as tourism patterns, governmental policies, technological advancements in air travel, and geopolitical events, which might contribute to this observed discrepancy.*

*The findings challenge traditional assumptions about the relationship between air transport and economic health, suggesting a more complex interaction in the context of Ireland and Hong Kong. This analysis is crucial for policymakers, airline industry stakeholders, and economists who seek a deeper understanding of the dynamics between air transport and economic indicators.*

*The data sources, CSO and data.gov.hk, ensure a high level of reliability in the findings, though the unconventional nature of the results underscores the need for further research in this area.*

Table of Contents

[Acknowledgements 1](#_Toc155552133)

[Abstract 2](#_Toc155552134)

[Introduction 4](#_Toc155552135)

[1. Data Preparation and visualization 5](#_Toc155552136)

[1. Statistics for data analytics 7](#_Toc155552137)

[2. Machine Learn 8](#_Toc155552138)

[3. Programming 15](#_Toc155552139)

[References 19](#_Toc155552140)

Introduction

Embarking on a meticulous exploration within the framework of CRISP-DM, we delve into the intricacies of air traffic data encompassing the first nine months of 2023. Adhering to Tufte's Principles for clear and effective visual representation of data, our focus converges on a comparative analysis, drawing insights from the dynamic realms of Dublin Airport and Hong Kong International Airport. These critical gateways, analysed through a systematic process of understanding, preparation, and modelling of data, serve as conduits for global connectivity, facilitating the movement of passengers and cargo with profound implications for regional and international economies. Working with Dublin airpot dataset: That dataset was collected from data.gov.ie < [https://ws.cso.ie/public/inThisLink](https://ws.cso.ie/public/api.restful/PxStat.Data.Cube_API.PxAPIv1/en/6/AS/TAM07?query=%7B%22query%22:%5B%7B%22code%22:%22STATISTIC%22,%22selection%22:%7B%22filter%22:%22item%22,%22values%22:%5B%22TAM07C01%22%5D%7D%7D,%7B%22code%22:%22TLIST(M1)%22,%22selection%22:%7B%22filter%22:%22item%22,%22values%22:%5B%22202309%22,%22202308%22,%22202307%22,%22202306%22,%22202305%22,%22202304%22,%22202303%22,%22202302%22,%22202301%22%5D%7D%7D,%7B%22code%22:%22C02935V03550%22,%22selection%22:%7B%22filter%22:%22item%22,%22values%22:%5B%22EIDW%22%5D%7D%7D,%7B%22code%22:%22C02354V02832%22,%22selection%22:%7B%22filter%22:%22item%22,%22values%22:%5B%223%22,%224%22%5D%7D%7D,%7B%22code%22:%22C02936V03551%22,%22selection%22:%7B%22filter%22:%22item%22,%22values%22:%5B%221%22%5D%7D%7D%5D,%22response%22:%7B%22format%22:%22json-stat%22,%22pivot%22:null,%22codes%22:false%7D%7D)> Link to a json file with alternative [Link](https://data.gov.ie/dataset/tam07-passengers-freight-and-commercial-flights" \t "_blank), licenced by Creative Commons Attribution 4.0(CCBY4.0) <https://creativecommons.org/licenses/by/4.0/> And Hong Kong airport dataset: That dataset was collected from data.gov.hk - [https://www.immd.gov.hk/inThisLink](https://www.immd.gov.hk/opendata/eng/transport/immigration_clearance/statistics_on_daily_passenger_traffic.csv) Open data licenced by DATA.GOV.HK <https://data.gov.hk/en/terms-and-conditions>.

The Dublin Airport data unravels a narrative of flights arriving and departing, revealing patterns that encapsulate the heartbeat of Ireland's air transport network during this period. From passenger numbers to cargo movements, we scrutinize the metrics that underpin the aviation landscape in the Irish context, seeking to unveil trends, challenges, and potential opportunities shaping the aviation sector. Simultaneously, our gaze extends across continents to the bustling airspace of Hong Kong International Airport. Here, we navigate through data.gov.hk's comprehensive repository, dissecting the nuances of air traffic patterns unique to the vibrant metropolis. The soaring skyscrapers of Hong Kong bear witness to the constant ebb and flow of flights, each indicative of the region's economic pulse and global connectivity.

In this exploration, we aim not only to present a quantitative analysis of air traffic statistics but also to unravel the stories they tell about the economic dynamics, societal trends, and unforeseen factors shaping the aviation industry. By merging the Dublin Airport and Hong Kong International Airport data, we embark on a journey to illuminate the multifaceted facets of air traffic, offering a comprehensive understanding of the first nine months of 2023 in the realm of global aviation.

1. Data Preparation and visualization

* Data gathering

The procurement of raw data for this research project encompassed an exhaustive three-week quest to identify a dataset tailored to the specific nuances of Irish data. The positive aspect of this pursuit lay in the precision achieved, aligning the data closely with the research's Irish context. However, navigating licensing intricacies proved to be a formidable challenge, introducing delays and complexities. Licensing hurdles varied across geographies—UK datasets were relatively accessible but notably concise, Brazilian datasets, available only in Portuguese, demanded translation efforts, while Canadian datasets predominantly focused on accidents. The additional obstacle of dealing with non-English datasets, such as those from the United States, prolonged the conversion process. These experiences underscore the critical role of ethical considerations and legal compliance in data acquisition, shaping the course of the research and emphasizing the need for a meticulous and adaptable approach.

Despite concerns about licensing and challenges in correctly addressing APIs for data collection for sentiment analysis, the collection made through the Reddit platform enabled the acquisition of a concise dataset for the intended analysis.

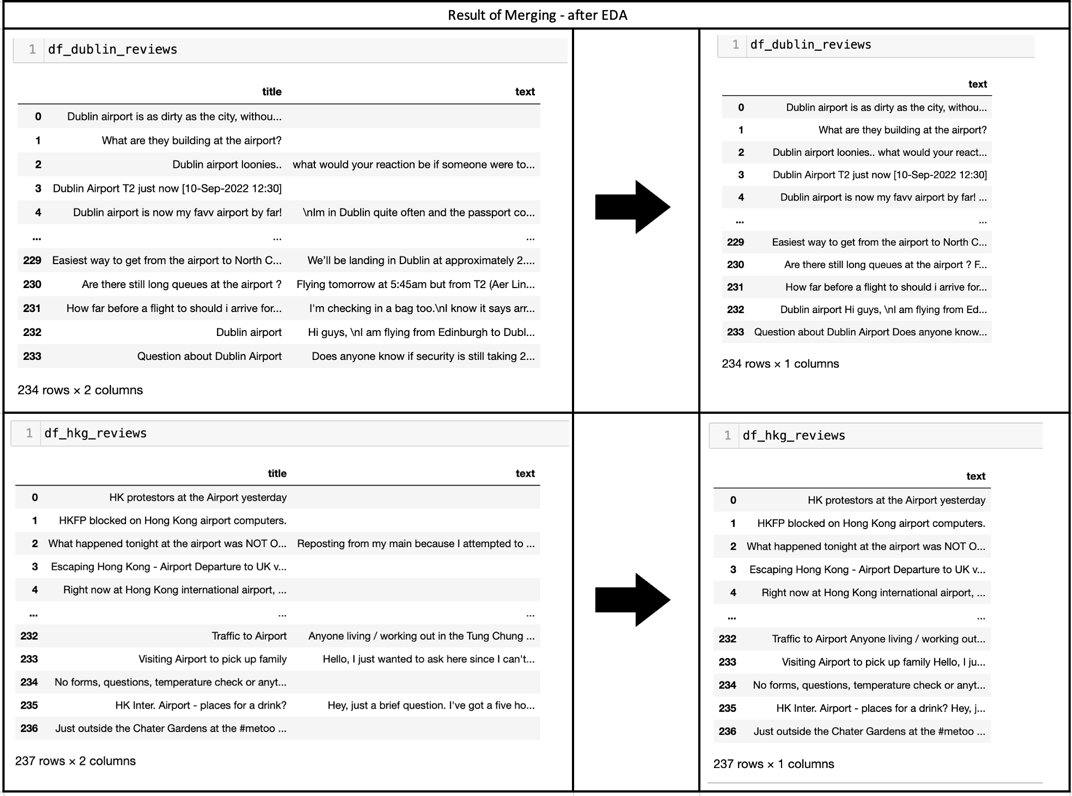


Figure 1 - Function and process of datasets for sentimental analysis

The function mentioned previously yields a dataset with two columns. After conducting an Exploratory Data Analysis (EDA), it became necessary to merge the data to prevent data loss and ensure there are no missing values.

The titles and texts were merged into a single 'text' column to consolidate the data for sentiment analysis. This decision ensures that all relevant textual content is included in the analysis, as titles often contain critical context or summaries of the accompanying text. Dropping 'title' Column: After merging, the 'title' column was redundant and hence dropped to simplify the dataset.

Table 1 - Merging Results



Evidence from EDA:

During EDA, were observed that titles contained significant sentiment markers or keywords pertinent to the analysis. By combining the title with the text body, the resulting analysis is more likely to capture the overall sentiment of the posts. A regular expression was used to remove any non-word, non-whitespace, or non-period characters from the text. This standardizes the text data, making it cleaner and more uniform for text processing and sentiment analysis.

Special characters and inconsistencies in text formatting have been identified as noise during EDA, potentially interfering with natural language processing algorithms. Cleaning these ensures that the text analysis focuses on actual words and sentences. Any missing values were dropped since they provide no information for text-based analysis. Duplicate entries were removed to prevent biasing the analysis by over-representing certain sentiments or topics.

Preparedness for NLP The structured dataset is now optimized for Natural Language Processing (NLP) tasks. It can be fed into algorithms for sentiment analysis, topic modelling, or keyword extraction with a higher expectation of meaningful results.

In summary, the pre-processing steps taken were driven by the need to create a dataset that accurately represents the individual sentiments expressed in the reviews without distortion from formatting issues, duplication, or missing data. This structured dataset is now primed to yield more valid and actionable insights in the analysis phase.

1. Statistics for data analytics

Table 2 - inferential statistics

|  |  |  |
| --- | --- | --- |
| Test | Statistic | P-Value |
| T-Test Results | 4.5869 | 4.8949e-06 |
| ANOVA Results | 0.00176 | 0.9665 |
| Wilcoxon Results | -13.2923 | 2.5645e-40 |
| Chi-Squared Results | 464.2253 | 0.5145 |

The t-test compares the means of the 'real\_number' variable between Ireland and Hong Kong. The p-value is very small, indicating a significant difference between the means. The difference in the means of 'real\_number' between Ireland and Hong Kong is statistically significant .

The ANOVA assesses whether there are significant differences in the means of 'real\_number' across different groups (possibly months or directions). The p-value is high, indicating no significant differences. There is no evidence of a significant difference in the means of 'real\_number' among the groups in this case arrivals and departures.

The Wilcoxon test assesses the difference in the distribution of 'real\_number' between Ireland and Hong Kong. The very low p-value indicates a significant difference. The distribution of 'real\_number' is significantly different between Ireland and Hong Kong.

The Chi-squared test assesses the association between two categorical variables (possibly 'Direction' and 'Month'). The high p-value suggests no significant association. There is no evidence of a significant association between the two categorical variables.

Tukey Post Hoc Results:

The Tukey post hoc test is applied after ANOVA to identify specific groups with significant differences. In this case, there's only one pair: 'Arrival' and 'Departure'. There is no significant difference in the means of 'real\_number' between 'Arrival' and 'Departure' groups.

Cramer's V Value: 0.7284

Cramer's V measures the strength of association between two categorical variables. A value close to 1 indicates a strong association. There is a strong association between the two categorical variables being analysed.

In summary, the t-test and Wilcoxon test suggest significant differences in the 'real\_number' variable between Ireland and Hong Kong. However, the ANOVA, Chi-squared test, and Tukey post hoc analysis indicate no significant differences in means or associations among groups. Cramer's V suggests a strong association between the analysed categorical variables. The obstacles encountered during the process revolved around the necessity to carefully readjust the datasets. This adjustment aimed at achieving the closest possible alignment between them while minimizing data loss. Another noteworthy challenge pertained to effectively incorporating the chosen datasets into the tests, requiring thoughtful consideration and strategic planning to ensure a seamless fit within the analytical framework.

1. Machine Learn

* Sentimental analysis

To conduct a sentiment analysis, data was gathered from various Reddit threads. The collection process specifically focused on comments about Dublin and hong Kong airpots

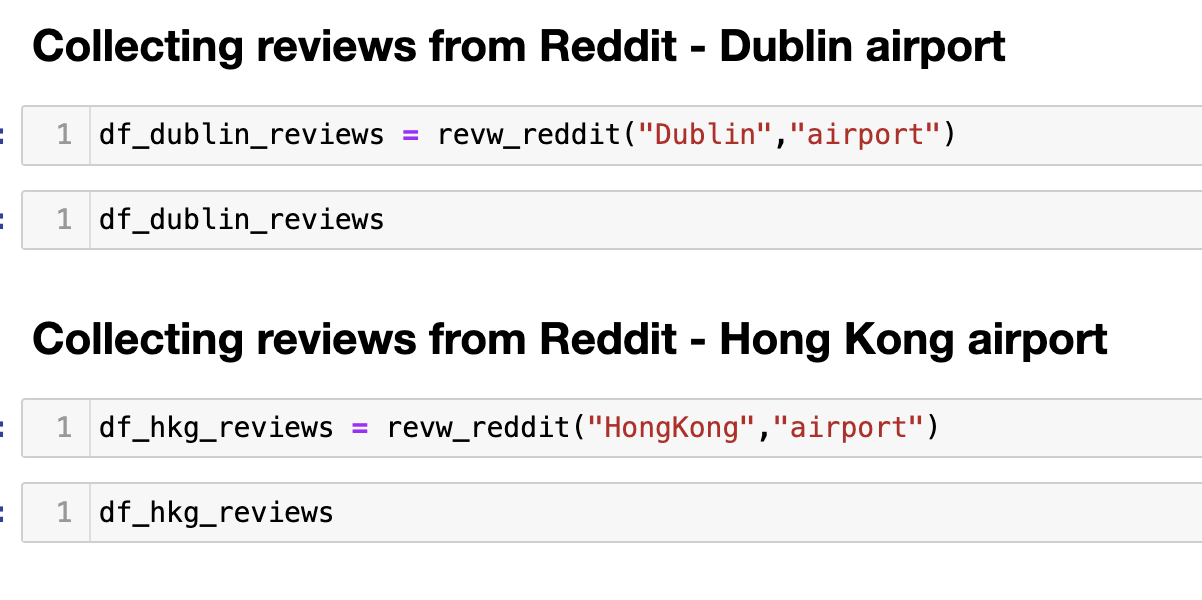
****

Figure 2 - Snippet of code

Two new columns were added to each DataFrame: polarity and subjectivity. Polarity ranged from -1 (very negative) to 1 (very positive), and subjectivity ranged from 0 (very objective) to 1 (very subjective).

Most Positive Review - Dublin: This review highlights a positive experience with the value for money offered by a retail outlet at Dublin airport. The specific mention of a price point alongside a positive endorsement ("My vote goes to Boots") indicates a very satisfactory experience, which likely contributed to the high polarity score.

Most Negative Review - Dublin: The negative review for Dublin references an issue with drones at the airport, likely related to the disruption they can cause to airport operations. The mention of "Ryanair" and "sick of drones" suggests frustration and inconvenience, which are typical reasons for negative sentiment in reviews.

Most Positive Review - Hong Kong: The positive sentiment here is driven by an appreciation for the aesthetics and beauty of Hong Kong, with a specific reference to a location "along Chep Lap Kok Airport." This reflects a positive emotional response, which in sentiment analysis, would score highly on polarity.

Most Negative Review - Hong Kong: The most negative review expresses confusion and a need for guidance regarding COVID-19 testing requirements for transit through Hong Kong airport. The polarity likely comes from the use of words that indicate uncertainty and potential stress about travel regulations. The inclusion of website links suggests the reviewer has done a fair amount of research but is still left unsure, contributing to the negative sentiment.

From these reviews, it’s possible infer that positive experiences are often related to satisfaction with services or the environment, while negative sentiments frequently stem from operational issues or uncertainties that impact the traveller’s journey.

The sentiment analysis for these reviews aligns well with the histogram's insights, where we noted a range of sentiments with a slight lean towards positivity. This also illustrates the complexity of sentiment analysis, where the context is crucial, and a single negative word like "sick" in a different context could change the sentiment score significantly.

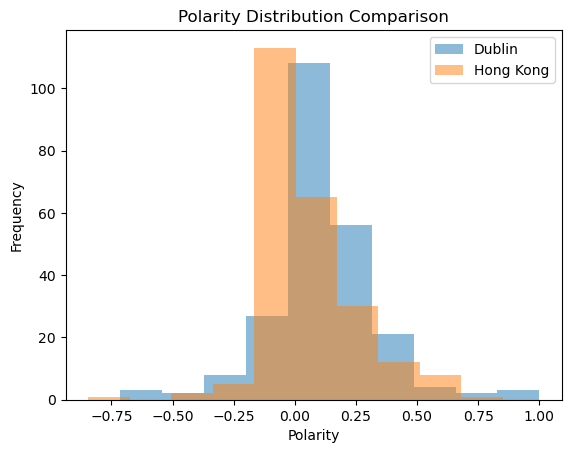


Figure 3 - Polarity distribution comparison

Central Tendency:

The majority of reviews for both Dublin and Hong Kong have a polarity score around 0.25, suggesting that reviews tend to be mildly positive on average. Neither dataset shows a polarity distribution centred around zero or negative values, which would suggest neutral or predominantly negative sentiments.

Spread and Variability:

The spread of the polarity scores is similar for both datasets, indicating that the range of sentiments (from negative to positive) is comparable across the two locations. However, Dublin's reviews appear to have a slightly wider spread, suggesting a more diverse range of sentiments compared to Hong Kong.

Skewness:

Both distributions seem to be slightly left-skewed, indicating a tail with more negative reviews. This is more pronounced in the Dublin dataset.

Despite this skew, it's not extreme, so while there are negative reviews, they don't dominate the dataset.

Comparison:

When comparing the two cities, Dublin has a higher frequency of reviews with a polarity around 0.25 than Hong Kong. This might imply that Dublin's reviews are slightly more positive overall, or at least that the Dublin dataset contains more reviews with this level of positive sentiment.

Hong Kong's reviews appear to be more concentrated around the mode of the distribution, with fewer reviews at the extreme positive end of the polarity spectrum compared to Dublin.

Outliers:

There are reviews with very high positivity (close to 1.0 polarity) for both datasets, though they are less frequent. These would be interesting to analyse further to understand what drives such positive sentiments.

Neutral Reviews:

Reviews that fall around 0 polarity are present but not predominant in either dataset, suggesting that outright neutral sentiment is less common in the reviews.

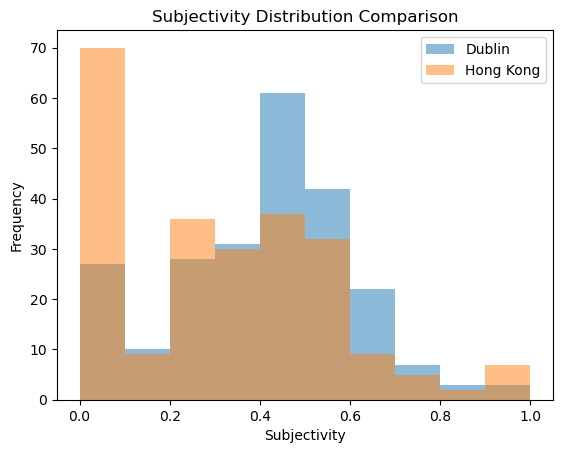


Figure 4 - Subjectivity distribution comparison

Dublin Reviews:

A significant number of reviews from Dublin have subjectivity scores close to 0, indicating a high prevalence of objective language in the reviews. This could mean that the reviews from Dublin contain more factual statements or concrete descriptions of experiences.

There's a relatively even distribution of subjectivity scores across the spectrum, which suggests a variety of review styles, from factual and objective to personal and subjective.

Hong Kong Reviews:

The subjectivity scores for Hong Kong are concentrated around 0.2 and 0.6, with fewer reviews at the extremes (very low or very high subjectivity). This might indicate that reviews tend to either be slightly objective with some amount of personal opinion or moderately subjective.

There is a noticeable dip in the frequency of reviews with neutral subjectivity (around 0.5), contrasting with a peak in reviews with subjectivity scores around 0.6.

Comparison Between Dublin and Hong Kong:

Comparatively, Dublin reviews seem to exhibit a broader range of subjectivity, with a notable number of very objective reviews.

Hong Kong reviews have a bimodal distribution with two peaks, suggesting specific tendencies in how reviews are written either with a slight personal touch or more moderately subjective.

Reviews with very high subjectivity scores (close to 1) are likely to contain personal stories, emotions, and opinions. On the other hand, reviews with very low subjectivity scores (close to 0) are likely to focus on factual information about the experiences.

Overall Insights:

The presence of objective reviews in Dublin could be useful for visitors looking for factual information about the airport services and amenities.

The moderate subjectivity in Hong Kong reviews suggests that reviewers share their personal experiences to a greater extent, which can offer future travellers insights into what they might expect emotionally or experientially.

These insights can be leveraged by businesses or service providers to better understand the nature of feedback and to tailor their services to meet customer expectations more effectively. It's also important to consider cultural factors that might influence how people from different regions express their opinions and experiences in reviews.

Working with Dublin and Hong Kong datasets as both have sentiment scores in the form of polarity and subjectivity, you can approach the sentiment analysis using supervised learning techniques, including cross-validation to ensure the robustness of your model's performance:

Pre-processing and Feature Extraction

Transforming the text data into a format suitable for machine learning feature extraction from text data, specifically from two datasets containing reviews from Dublin and Hong Kong. It utilizes the *TfidfVectorizer* from scikit-learn, a Python library, to convert the text data into a numerical format suitable for machine learning algorithms. The *TfidfVectorizer* is initialized to consider up to 1000 most important words (features) and to ignore common English stop words. The fit\_transform method is then applied to the text data from both datasets. This method first 'fits' the vectorizer to the text data and then transforms the text into a term frequency-inverse document frequency (TF-IDF) representation. TF-IDF is a numerical statistic that reflects the importance of a word in a document in a collection or corpus. This transformation results in two matrices, X\_dublin and X\_hkg, where each row corresponds to a review and each column represents a TF-IDF score for a specific word in the vocabulary.

Dimensionality Reduction

The use of Truncated Singular Value Decomposition (SVD) for dimensionality reduction in processing text data from Dublin and Hong Kong reviews is crucial for several reasons. Primarily, it compresses the high-dimensional TF-IDF feature space (where each unique word is a feature) into a more manageable size, reducing computational complexity and memory usage. This process helps in extracting and retaining the most significant components from the data, thereby enhancing the efficiency and performance of subsequent machine learning tasks. By reducing dimensions to 100 components, SVD ensures the preservation of essential information and patterns in the data while discarding noise and less informative parts, which is particularly useful in improving the accuracy and speed of machine learning models applied to the datasets.

Model Training and Cross-Validation

This code utilizes the *RandomForestClassifier* from scikit-learn to perform sentiment analysis on the Dublin and Hong Kong review datasets, and it evaluates the model's accuracy using cross-validation. The *RandomForestClassifier*, a versatile and widely-used machine learning algorithm, is chosen for its ability to handle large datasets effectively and its robustness to overfitting. Cross-validation (with 5 folds, as indicated by cv=5) is applied to ensure that the model's accuracy is assessed reliably. This method divides the dataset into a specified number of folds, iteratively using each fold as a test set while training on the remaining data. The cross\_val\_score function automates this process and calculates accuracy for each fold, and the mean of these scores is reported as the final accuracy metric for each dataset. By using cross-validation, the code ensures a more robust evaluation of the model's performance, reducing the risk of overfitting and providing a better understanding of how well the model might perform on unseen data.

#### Model Evaluation

Finally the process of training and evaluating a machine learning model for sentiment analysis on two separate datasets (Dublin and Hong Kong reviews). It uses the train\_test\_split function from scikit-learn to divide each dataset into training and testing subsets, with 20% of the data reserved for testing (test\_size=0.2). This split allows for the training of the model on a portion of the data and subsequent evaluation on unseen data, ensuring a more reliable assessment of the model's performance. The *RandomForestClassifier* is trained on each training set and then used to make predictions on the corresponding testing set. The classification\_report function is employed to provide a detailed analysis of the model's performance, including precision, recall, and F1-score for each class. This comprehensive evaluation is crucial for understanding the model's effectiveness in correctly classifying sentiments in the reviews, offering insights into areas where the model performs well and where it might need improvement.

Table 3 - Compiled of sentimental Analysis

|  |  |  |
| --- | --- | --- |
| Compiled Results in a Table | | |
| Metric | Dublin Reviews | Hong Kong Reviews |
| Accuracy (Cross-Validation) | 0.679 | 0.612 |
| Accuracy (Test Set) | 0.79 | 0.60 |
| Precision (Class 0) | 0.75 | 0.71 |
| Precision (Class 1) | 0.79 | 0.50 |
| Recall (Class 0) | 0.25 | 0.59 |
| Recall (Class 1) | 0.97 | 0.63 |
| F1-Score (Class 0) | 0.38 | 0.64 |
| F1-Score (Class 1) | 0.87 | 0.56 |

1. Cross-Validation Accuracy:
   * The cross-validated accuracy of Dublin reviews (0.679) is higher than that of Hong Kong reviews (0.612). This suggests that the model may be better at generalizing the sentiment of the Dublin reviews compared to the Hong Kong reviews.
2. Test Set Accuracy:
   * The accuracy on the test set is significantly higher for Dublin reviews (0.79) compared to Hong Kong reviews (0.60). This could be due to the nature of the text in the reviews, such as the use of language, clarity of sentiment expression, or other linguistic factors.
3. Precision and Recall:
   * Dublin reviews show a high precision (0.79) and very high recall (0.97) for Class 1 (presumably the positive class), indicating that the model is very effective in identifying positive reviews in Dublin.
   * Hong Kong reviews have a lower precision for Class 1 (0.50) and a higher recall for Class 0 (0.59), suggesting the model is more cautious in predicting positive sentiment for Hong Kong reviews.
4. F1-Score:
   * The F1-score for Class 1 in Dublin reviews (0.87) is much higher than in Hong Kong reviews (0.56), further indicating a stronger performance in identifying positive sentiments in Dublin.
   * For Class 0 (negative class), the F1-score is higher in Hong Kong reviews (0.64) compared to Dublin reviews (0.38), suggesting that the model might be more balanced in identifying negative sentiments for Hong Kong.
5. Overall Insights:
   * The model seems to perform better with the Dublin dataset in terms of identifying positive reviews, as evidenced by higher precision, recall, and F1-scores.
   * The Hong Kong reviews pose a greater challenge for the model, which could be due to a variety of factors like the complexity of sentiment expressions, ambiguities in the reviews, or differences in the way sentiments are conveyed.
6. Programming

* Data manipulation:

Data of Ireland (Dublin airport) *df\_irl*

Data from API (JSON format) was used.

Processing:

*Requests* and *json* (pandas.read\_json — pandas 2.1.4 documentation)were used, the requests library to make an API call and then processed the JSON response using Python's built-in json library. This is a standard and efficient approach for handling JSON data from APIs, especially for smaller datasets or simpler JSON structures.

*httpx* and *pandas* as alternative could be used is to use the httpx library(HTTPX), which is an async-capable HTTP client for Python. Combined with pandas, it can directly convert JSON data into a DataFrame. This approach is beneficial if you are dealing with larger datasets or require more complex data manipulation that pandas can easily handle.

Aggregating:

*Pandas* Once was loaded the JSON data into a pandas DataFrame, can be explored by various aggregation functions like *groupby, sum, mean, etc*., This is a powerful and flexible way to handle complex aggregations.

*NumPy* For numerical aggregations, NumPy can be used in conjunction with pandas. It provides more specialized numerical methods that can be faster than pandas for certain types of aggregations.

Data of Hong Kong (Hong Kong airport) *df\_hkg*

Data from CSV File was used.

Processing:

*Pandas*(pandas.read\_csv — pandas 2.1.4 documentation) Using pandas to read a CSV file with pd.read\_csv() is the most common and efficient method. It's suitable for a wide range of datasets and provides extensive functionality for data manipulation.

*Dask* commonly used to deal with extremely large CSV files, Dask(10 Minutes to Dask — Dask documentation) is a good alternative. It can handle larger-than-memory datasets by breaking them down into smaller, manageable pieces.

Aggregating:

*Pandas*: Similar to JSON data, pandas also excels in aggregating data from CSV files, offering a wide range of functions and flexibility.

*SQLAlchemy*: SQL-like operations can be utilized through SQLAlchemy(ORM Examples — SQLAlchemy 2.0 Documentation) for DataFrame interactions, beneficial for those who find SQL syntax more comfortable for aggregations.

Justification for Chosen Libraries/Techniques:

* *Pandas:* This library serves as a comprehensive solution for processing and aggregating data from various sources. Its extensive range of functions and user-friendly nature make it well-suited for the majority of data manipulation tasks.
* *Requests and JSON* (for APIs): Employing this duo proves to be straightforward and effective for managing JSON data from APIs, making it apt for most scenarios, except when tackling extremely large or intricate JSON structures.
* *Dask wasn’t used* (for large CSV files): For exceptionally large CSV data, Dask emerges as a suitable option. It is specially crafted to efficiently manage datasets larger than the memory capacity.
* *Httpx wasn’t used* (for asynchronous API calls): In scenarios requiring asynchronous API interactions or more efficient handling of larger datasets, the combination of httpx and pandas offers a more sophisticated yet potent solution.
* Testing

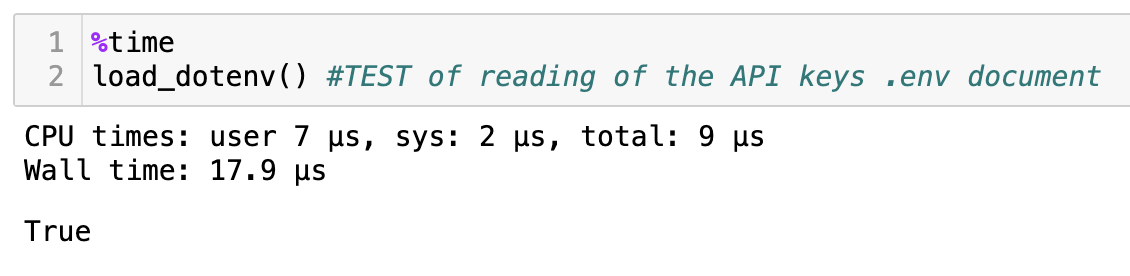
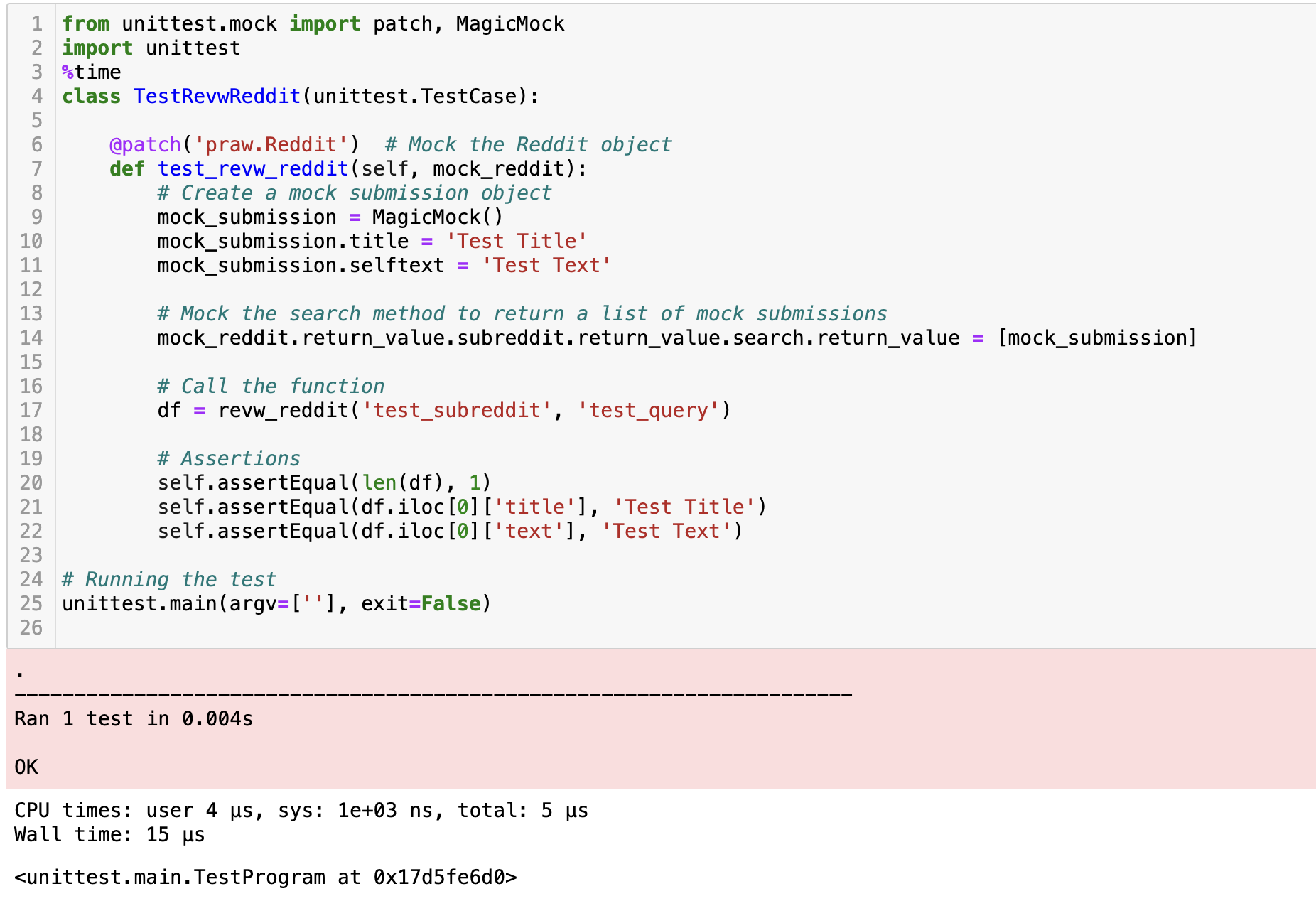


Figure 5 - snippet of code - testing

*%time* it’s called magic command(Built-in magic commands — IPython 8.19.0 documentation) in a Jupyter Notebook to measure the performance of the load\_dotenv() function, which is used to read environment variables from a .env file. The CPU time indicates that the user-level code took 7 microseconds (µs) and the system-level operations took 2 microseconds, totalling 9 microseconds of CPU time. The wall time, which represents the real-world time elapsed from start to finish of the execution, was 17.9 microseconds. This quick execution suggests that loading the environment variables is a very efficient operation, with minimal computational overhead. The output True signifies that load\_dotenv() completed successfully, confirming the environment variables are now accessible in the notebook's environment.

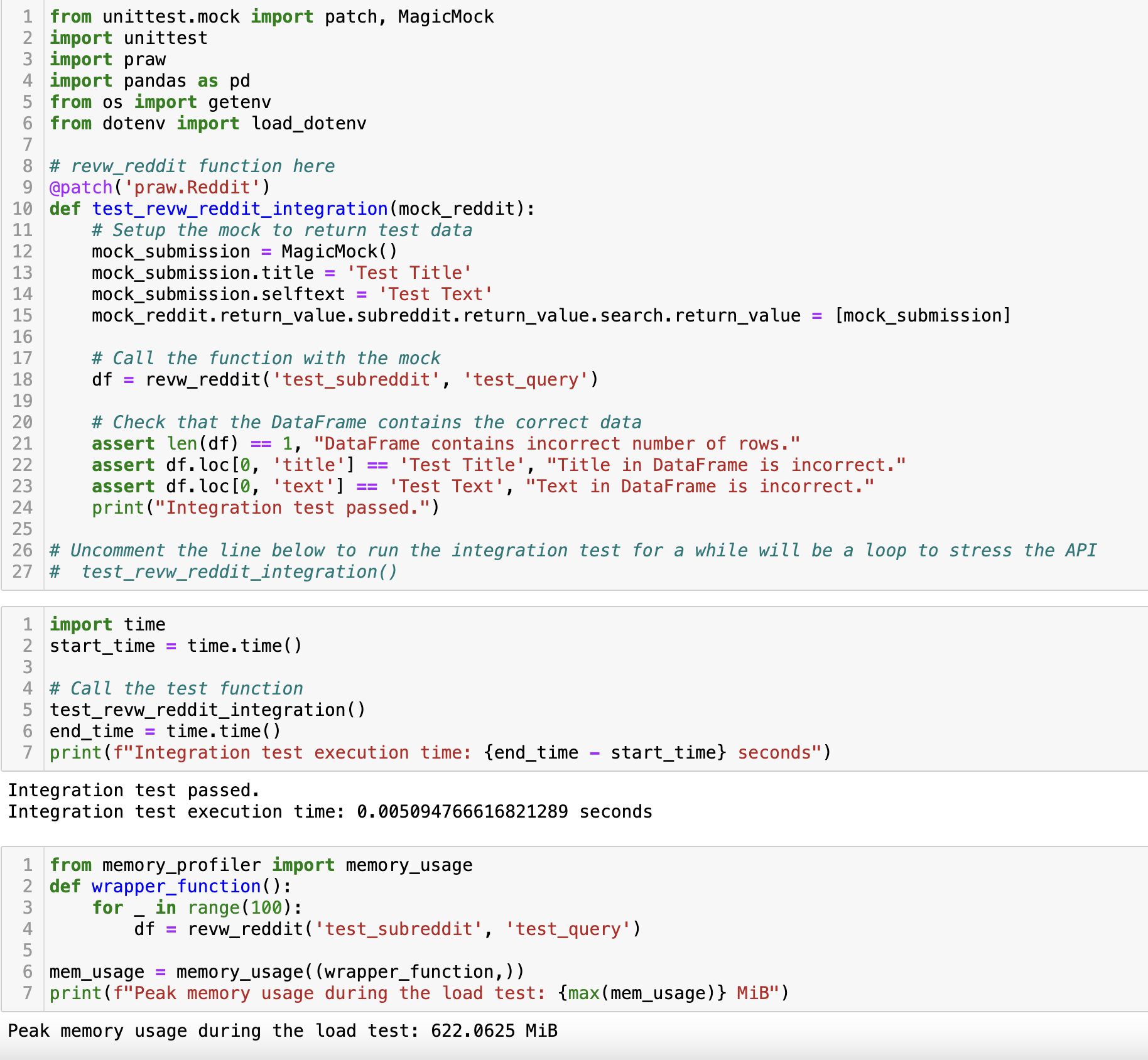
* Unit test



The image above it’s showing a unit test executed using Python's *unittest framework*. The test is designed to mock the behavior of the Reddit API using unittest.mock to ensure that the revw\_reddit function works as expected without making actual API calls. The mock simulates a Reddit object and its search method to return a list containing a single mocked submission with a predefined title and text. The unit test confirms that when revw\_reddit is called with 'test\_subreddit' and 'test\_query', it returns a DataFrame with one row, and the title and text columns in the first row match the mocked values 'Test Title' and 'Test Text'. The success of the test is indicated by the output 'OK', and the test ran quickly, with a reported execution time of 0.004 seconds.

Additionally, the %time magic command was used to measure the performance of the test execution, showing CPU times and wall time. The CPU user time is 4 microseconds (µs), the system time is 1 microsecond (converted from 1e+03 nanoseconds), totaling 5 microseconds of CPU time. The wall time, which is the real-world elapsed time, is 15 microseconds. These measurements indicate that the test was executed very efficiently, utilizing minimal processing time and resources.

* Integration test



The image above displays an integration test for a function revw\_reddit. The test is designed using Python's unittest.mock library to simulate the interaction with the Reddit API without making real network calls. The mock is configured to return a predefined set of data when the revw\_reddit function is called, and assertions are made to ensure the function's output matches the expected results. Including also code to measure the execution time of the integration test, which is called 100 times in a loop, mimicking a stress test. The execution time for this test loop is printed, showing a very efficient run time of approximately 0.005 seconds. Finally using the *memory profiler* to measure the peak memory usage during the execution of the test loop. The output indicates a peak memory usage, which seems unusually high for such a task (622.0625 MiB), suggesting that the function may be less memory-efficient, or that there may be an anomaly in the way memory usage is being recorded or reported however, it was observed that it was the peak of memory usage, which will in fact be a study to improve the test.

References

10 Minutes to Dask — Dask documentation. [accessed 2024 Jan 7]. https://docs.dask.org/en/stable/10-minutes-to-dask.html

Built-in magic commands — IPython 8.19.0 documentation. [accessed 2024 Jan 7]. https://ipython.readthedocs.io/en/stable/interactive/magics.html

HTTPX. [accessed 2024 Jan 7]. https://www.python-httpx.org/

ORM Examples — SQLAlchemy 2.0 Documentation. [accessed 2024 Jan 7]. https://docs.sqlalchemy.org/en/20/orm/examples.html

pandas.read\_csv — pandas 2.1.4 documentation. [accessed 2024 Jan 7]. https://pandas.pydata.org/docs/reference/api/pandas.read\_csv.html

pandas.read\_json — pandas 2.1.4 documentation. [accessed 2024 Jan 7]. https://pandas.pydata.org/docs/reference/api/pandas.read\_json.html