Sentiment Analysis of Amazon Reviews

A Minor Project II Report
Submitted in Partial fulfillment for the award of
Bachelor of Technology in CSE-AIDS

Submitted to
RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA
BHOPAL (M.P)

MINOR PROJECT II REPORT
Submitted by

Aryan Singh[0103AD211016]

Under the supervision of
Dr. Parmalik Kumar
Professor

Department of CSE-AIDS
Lakshmi Narain College of Technology, Bhopal (M.P.)
Session 2023-24

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY, BHOPAL

DEPARTMENT OF CSE-AIDS

CERTIFICATE

This is to certify that the work embodied in this project work entitled **"Sentiment Analysis of Amazon Review"** has been satisfactorily completed by the **Aryan Singh** (0103AD211016). It is a bonafide piece of work, carried out under the guidance in **Department of CSE-AIDS**, **Lakshmi Narain College of Technology, Bhopal** for the partial fulfillment of the **Bachelor of Technology** during the academic year 2023-24.

Approved By

Dr. Parmalik Kumar
(Professor)

Dr. Tripti Saxena
Prof. & Head
Department of CSE-AIDS

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY, BHOPAL

DEPARTMENT OF CSE-AIDS

ACKNOWLEDGEMENT

We express our deep sense of gratitude to Dr. Parmalik Kumar (Guide) department of CSE- AIDS L.N.C.T., Bhopal. Whose kindness valuable guidance and timely help encouraged me to complete this project.

A special thank goes to Dr. Tripti Saxena (HoD) who helped me in completing this project work. She exchanged her interesting ideas & thoughts which made this project work successful.

We would also thank our institution and all the faculty members without whom this project work would have been a distant reality.

Aryan Singh [0103AD211016]

CHAPTER 1

PROBLEM DOMAIN

In the realm of sentiment analysis applied to Amazon reviews, the central challenge revolves around extracting meaningful insights from the vast and diverse array of customer feedback available on the platform. Amazon, being one of the world's largest e-commerce platforms, hosts millions of product reviews covering a wide range of products and categories. This extensive repository of customer opinions provides a rich dataset for analysis but also presents significant challenges in terms of scale, diversity, and complexity.

Sentiment analysis, a branch of natural language processing (NLP), plays a pivotal role in addressing these challenges. The primary goal is to develop algorithms and techniques that can automatically categorize reviews as positive, negative, or neutral based on the sentiments conveyed within the text. This involves not only understanding the polarity of individual words but also considering the context in which they are used and the overall tone and sentiment expressed throughout the review.

Key aspects of the problem domain include:

1. Data Collection and Preprocessing:

- Data Collection: Gathering Amazon reviews related to specific products or categories. This can be achieved through APIs provided by Amazon or web scraping techniques. The collected data usually includes review texts, ratings, timestamps, and other metadata.
- Data Cleaning: The raw data often contains noise, such as irrelevant information, spam, or non-English text. Cleaning the data is essential to remove such noise and ensure the quality of the dataset.
- Preprocessing Steps: This includes tokenization (breaking down the text into individual words or tokens), stemming or lemmatization (reducing words to their base or root form), and removing stopwords (common words like "and," "the," which do not carry significant meaning). These steps prepare the text for analysis by normalizing and simplifying the data.

2. Sentiment Analysis Techniques:

- Lexicon-Based Methods: These methods use predefined sentiment dictionaries where words are tagged with sentiment scores. The overall sentiment of a review is determined by aggregating the sentiment scores of individual words. While straightforward, these methods may struggle with context and complex language structures.

- Machine Learning Approaches: These involve training algorithms on labeled datasets where reviews are annotated with sentiment labels (positive, negative, or neutral). Techniques include:

    1. Supervised Learning: Algorithms like Naive Bayes, Support Vector Machines (SVM), and neural networks are trained on labeled data to learn patterns and make predictions on new reviews.
    2. Deep Learning: Advanced models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, can capture complex patterns and context within the text.
    3. Transfer Learning: Leveraging pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) that have been trained on large corpora and fine-tuning them on the specific task of sentiment analysis for Amazon reviews.

3. Interpretation and Actionable Insights:

- Analyzing Results: Once sentiment analysis is performed, the results can be aggregated and visualized to uncover patterns and trends. For instance, identifying common themes in positive or negative reviews can reveal specific aspects of products that customers like or dislike.
- Business Applications: Businesses can leverage these insights in multiple ways:

1. Product Improvement: Identifying recurring issues in negative reviews can guide product development teams to address these problems, leading to better product quality.
2. Customer Service Enhancement: Understanding customer sentiment can help businesses tailor their customer service strategies to better address customer concerns and improve satisfaction.
3. Marketing Optimization: Insights from sentiment analysis can inform marketing strategies by highlighting product features that resonate with customers, enabling targeted and effective marketing campaigns.
4. Data-Driven Decision Making: Overall, sentiment analysis empowers businesses to make informed decisions based on real customer feedback, leading to enhanced customer experiences and potentially increased sales and loyalty.

- ❖ Sentiment Analysis of Amazon Product Reviews Using Machine Learning Techniques
  - ➢ Summary: This paper, published in a peer-reviewed journal of computational linguistics, delves into the application of machine learning algorithms such as Support Vector Machines (SVM) and Recurrent Neural Networks (RNNs) for sentiment analysis of Amazon product reviews. The study evaluates the performance of these techniques in classifying reviews into positive, negative, or neutral categories. The findings provide insights into the strengths and weaknesses of SVMs and RNNs, highlighting their ability to handle the complexities of natural language in varying degrees and the importance of context in sentiment classification.
- ❖ Exploring Temporal Dynamics in Sentiment Analysis of Amazon Reviews
  - ➢ Summary: Presented at a conference on data mining and sentiment analysis, this paper examines how sentiment in Amazon reviews evolves over time. It takes into account factors such as product lifecycle stages and external events, employing time-series analysis and sentiment trend modeling. The study uncovers temporal patterns in customer opinions, which can be invaluable for businesses aiming to understand how sentiment changes over time and adapt their strategies accordingly.
- ❖ Lexicon-Based Sentiment Analysis of Amazon Customer Reviews: A Comparative Study
  - ➢ Summary: Published in a journal of information retrieval and sentiment analysis, this comparative study evaluates the effectiveness of various lexicon-based sentiment analysis techniques. It compares multiple sentiment lexicons and assesses their ability to capture nuanced sentiments in diverse product reviews. The research highlights the advantages and limitations of lexicon-based approaches, emphasizing their simplicity and ease of use but also their potential struggle with context and sarcasm.
- ❖ Aspect-Based Sentiment Analysis of Amazon Electronics Reviews
  - ➢ Summary: This paper, presented at a conference on natural language processing and sentiment analysis, focuses on aspect-based sentiment analysis of Amazon electronics reviews. By identifying sentiments associated with specific product features and attributes, the study provides detailed insights into consumer preferences regarding various aspects of electronic products. This granular analysis aids

businesses in refining product development and tailoring marketing strategies to address specific consumer needs and preferences.

- ❖ Deep Learning Approaches for Sentiment Analysis of Amazon Product Reviews
    - ➢ Summary: Published in conference proceedings on machine learning and artificial intelligence, this paper explores the application of deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for sentiment analysis. The study investigates how these models can automatically learn representations of textual data and capture complex sentiment patterns. The results demonstrate the superior performance of deep learning models in handling large-scale review datasets and extracting intricate sentiment nuances.

- ❖ Sentiment Analysis of Amazon Movie Reviews: A Case Study
    - ➢ Summary: Presented at a conference on sentiment analysis and digital media, this case study focuses on analyzing sentiments in Amazon movie reviews. By applying natural language processing techniques, the research investigates the sentiments expressed by users towards movies. The insights gained can be valuable for filmmakers, distributors, and streaming platforms to understand audience preferences, optimize marketing efforts, and enhance content creation.

- ❖ Sentiment Analysis of Amazon Book Reviews: Insights into Reader Preferences
    - ➢ Summary: Published in a journal of literature and digital humanities, this paper examines sentiment analysis of Amazon book reviews. The study aims to uncover reader preferences and trends by analyzing the sentiments conveyed in book reviews. The findings provide valuable insights for authors, publishers, and marketers, helping them understand factors that influence reader satisfaction and engagement, and guiding them in making informed decisions about book publishing and promotion.

- ❖ Analyzing Product Features and Sentiment in Amazon Reviews Using Natural Language Processing
    - ➢ Summary: Presented at a conference on e-commerce and consumer behavior, this paper utilizes natural language processing techniques to analyze product features and sentiments in Amazon reviews. The research explores how consumers express sentiments towards specific product attributes such as quality, usability, and price. The actionable insights derived from this analysis can help product developers, marketers, and retailers enhance product offerings and improve customer satisfaction.

- Sentiment Analysis of Amazon Food Reviews: Understanding Consumer Preferences and Trends
    - Summary: Published in a journal of food science and consumer research, this study investigates sentiment analysis of Amazon food reviews to gain insights into consumer preferences and trends in the food industry. By analyzing the sentiments expressed in food-related reviews, the research aims to identify popular products, detect emerging food trends, and understand factors influencing consumer satisfaction and purchasing decisions.

- Cross-Domain Sentiment Analysis of Amazon Reviews: Transfer Learning Approaches
    - Summary: Presented at a conference on machine learning and transfer learning, this research explores transfer learning approaches for cross-domain sentiment analysis of Amazon reviews. It investigates the transferability of sentiment knowledge learned from one domain (e.g., electronics) to another domain (e.g., clothing). The study aims to improve sentiment classification performance across diverse product categories, demonstrating the potential of transfer learning to enhance the robustness and generalizability of sentiment analysis models.

Objective:

The primary objective of this project is to analyze sentiments expressed in Amazon reviews to extract valuable insights into consumer opinions and preferences. By applying sentiment analysis techniques to a diverse range of product reviews on the Amazon platform, the project aims to achieve the following goals:

1. Understand Consumer Sentiment:

- Comprehensive Insight: Gain a thorough understanding of consumer sentiment towards various products and categories available on Amazon.
- Pattern Identification: Identify patterns in customer feedback to discern common themes in positive and negative reviews.
- Influencing Factors: Determine the factors that influence customer satisfaction and purchase decisions, such as product quality, price, brand reputation, and specific product features.

2. Provide Actionable Insights:

- Strategic Decisions: Provide businesses, marketers, and product developers with insights that can inform strategic decision-making.
- Enhance Customer Experience: Highlight areas where products excel or fall short, helping businesses improve customer satisfaction.
- Feature Analysis: Uncover sentiments related to specific product features or attributes, aiding in targeted product improvements and marketing strategies.

3. Explore Sentiment Trends:

- Temporal Dynamics: Analyse how sentiment evolves over time, examining temporal dynamics across different product categories and periods.
- Seasonal Variations: Identify seasonal variations and trends, helping businesses understand peak times for positive or negative feedback.

- Emerging Trends: Detect shifts in consumer preferences and emerging trends that can inform future product development and marketing campaigns.

4. Evaluate Sentiment Analysis Techniques:

- Performance Comparison: Evaluate the performance of various sentiment analysis techniques, including lexicon-based methods, machine learning algorithms, and hybrid approaches.
- Effectiveness and Limitations: Identify the strengths and weaknesses of different techniques to determine the most suitable approach for analyzing Amazon reviews.
- Best Practices: Establish best practices for sentiment analysis in the context of e-commerce reviews, contributing to the broader field of natural language processing.

Scope:

The scope of the project encompasses the following key aspects:

1. Data Collection:

- Diverse Dataset: Collect a sizable dataset of Amazon reviews covering a wide range of products and categories.
- Metadata: Include associated metadata such as product ratings, review dates, and product identifiers to provide context for the analysis.
- Source Diversity: Ensure the dataset includes reviews from different regions and languages (primarily focusing on English for simplicity and consistency).

2. Data Preprocessing:

- Text Normalization: Standardize text data to a consistent format, handling issues like capitalization and punctuation.
- Tokenization: Break down review text into individual words or tokens to facilitate analysis.
- Stopword Removal: Remove common words that do not carry significant meaning, such as "and," "the," "is".
- Noise Handling: Address inconsistencies and noise in the review text, such as typos, slang, and abbreviations.
- Data Cleaning: Handle missing data and remove irrelevant or spam content to ensure the dataset's quality and reliability.

3. Sentiment Analysis Techniques:

- Lexicon-Based Methods: Utilize predefined sentiment dictionaries to score and categorize sentiments based on word usage.
- Machine Learning Algorithms: Apply algorithms like Support Vector Machines (SVM), Naive Bayes, and deep learning models (e.g., CNNs, LSTMs) to classify sentiments.

- Hybrid Approaches: Explore the combination of lexicon-based and machine learning methods to leverage the strengths of both approaches.

4. Analysis and Interpretation:

- Sentiment Distribution: Analyze the overall sentiment distribution across the dataset, identifying the proportion of positive, negative, and neutral reviews.
- Consumer Preferences: Uncover insights into consumer preferences and satisfaction by examining sentiment trends and patterns.
- Product and Feature Analysis: Determine which products and features elicit strong sentiments and why, providing a detailed understanding of consumer feedback.
- Trend Analysis: Explore sentiment trends over time, identifying how opinions change across different seasons, product lifecycle stages, and in response to external events.

5. Presentation of Findings:

- Comprehensive Report: Summarize the results of the sentiment analysis in a detailed report.
- Key Insights: Highlight key insights and trends discovered during the analysis.
- Actionable Recommendations: Provide actionable recommendations for businesses, marketers, and product developers based on the findings.
- Visualizations: Include visualizations such as charts and graphs to effectively communicate trends and patterns to stakeholders.
- Stakeholder Engagement: Present findings to stakeholders through presentations or interactive dashboards, ensuring the insights are accessible and understandable.

CHAPTER 4

PROBLEM ANALYSIS AND REQUIREMENT SPECIFICATION

Problem Analysis:

1. Identification of Problem Domain:

- Sentiment Analysis: The primary focus is on analyzing sentiments expressed in Amazon reviews to extract insights into consumer opinions and preferences. This involves using natural language processing (NLP) techniques to classify the sentiment of reviews as positive, negative, or neutral.
- Challenges: Key challenges include handling noisy text data (e.g., typos, slang), addressing complex linguistic features like sarcasm and irony, and ensuring the scalability and efficiency of sentiment analysis techniques for large-scale datasets.
- Data Variety: Amazon reviews encompass a diverse range of product categories, each with unique linguistic characteristics and sentiment expressions, complicating the analysis.

2. Understanding Stakeholder Needs:

- Stakeholders:
  - Businesses: Need actionable insights to improve products, enhance customer service, and refine marketing strategies.
  - Marketers: Require detailed sentiment trends to tailor marketing campaigns and understand consumer preferences.
  - Product Developers: Seek specific feedback on product features to guide design and development.
  - Consumers: Benefit indirectly through improved products and services based on their feedback.
- Pain Points:
  - Businesses and Marketers: Lack of detailed insights into consumer sentiment and preferences.
  - Product Developers: Need precise feedback on product features and attributes.
  - Consumers: Desire products that better meet their expectations and preferences.

3. Data Exploration and Analysis:

- Dataset Characteristics:

- ➢ Review Distribution: Analyze the distribution of reviews across different product categories.
- ➢ Review Lengths: Examine the length of reviews to understand the variability in detail and verbosity.
- ➢ Sentiment Distribution: Assess the overall sentiment distribution to identify any biases or trends in the data.

- Preprocessing Needs: Understand the necessary preprocessing steps to clean and normalize the data for effective analysis.

4. Identification of Key Objectives:

- Consumer Sentiment Trends: Identify and understand trends in consumer sentiment across various product categories and time periods.
- Influencing Factors: Determine the factors that most significantly influence consumer sentiment, such as product features, pricing, and brand reputation.
- Actionable Insights: Provide businesses with clear, actionable insights to improve products, services, and marketing strategies based on sentiment analysis findings.

Requirement Specification:

1. Data Requirements:

- Dataset Selection:
  - ➢ Size and Diversity: Collect a sizable dataset of Amazon reviews that covers a wide range of product categories to ensure comprehensive analysis.
  - ➢ Metadata: Ensure the availability of essential metadata such as product ratings, review dates, and product identifiers.

- Data Preprocessing:
  - ➢ Text Normalization: Standardize text data to a uniform format, addressing capitalization, punctuation, and spelling variations.
  - ➢ Tokenization: Break down the text into individual words or tokens.
  - ➢ Stopword Removal: Remove common words that do not contribute significant meaning.
  - ➢ Missing Data Handling: Address missing or incomplete data appropriately.
  - ➢ Noise Reduction: Eliminate irrelevant information and spam from the dataset.

2. Functional Requirements:

- Sentiment Analysis Techniques:
  - Lexicon-Based Methods: Use predefined dictionaries to score and categorize sentiment.
  - Machine Learning Algorithms: Implement algorithms such as Support Vector Machines (SVM), Naive Bayes, and deep learning models (e.g., CNNs, LSTMs).
  - Hybrid Approaches: Combine lexicon-based and machine learning methods to leverage the strengths of both.

- Performance Metrics:
  - Accuracy: Measure the percentage of correctly classified sentiments.
  - Precision and Recall: Evaluate the relevance and completeness of sentiment classification.
  - F1-Score: Provide a balanced measure of precision and recall.

- Temporal Analysis:
  - Time Intervals: Select appropriate time intervals (e.g., months, quarters) for analyzing sentiment trends.
  - Trend Visualization: Develop methods for visualizing sentiment trends over time.

3. Non-Functional Requirements:

- Scalability: Ensure that the sentiment analysis techniques can efficiently handle large-scale datasets.
- Accuracy: Aim for high accuracy and reliability in sentiment classification to ensure meaningful insights.
- Computational Resources:
  - Hardware: Define memory and processing power requirements.
  - Software: Specify necessary software dependencies and environments for implementing sentiment analysis models.

4. Output Requirements:

- Reporting:
  - Comprehensive Report: Include sections for data exploration, methodology, results, analysis, and recommendations.
  - Detailed Insights: Summarize key findings and actionable recommendations for stakeholders.

- Visualization:
  - Charts and Graphs: Use visual aids to present sentiment analysis results effectively.
  - Heatmaps: Visualize the intensity and distribution of sentiments across different products and time periods.

5. Usability and Accessibility Requirements:

- User Interface:
  - Ease of Use: Design an intuitive interface for interacting with sentiment analysis models, if applicable.
  - Navigation: Ensure easy navigation through different functionalities.
  - Accessibility: Incorporate features to make the interface accessible to users with disabilities.

- Documentation:
  - Comprehensive Documentation: Provide detailed documentation to facilitate understanding and usage of the sentiment analysis techniques.
  - User Guides: Create user guides and tutorials for stakeholders to effectively utilize the insights generated by the project.

By addressing these detailed aspects, the project aims to deliver a robust and comprehensive sentiment analysis solution for Amazon reviews, providing valuable insights that can drive business improvements and enhance customer satisfaction.

CHAPTER 5

DETAILED DESIGN (MODELING AND ERD/DFD)

Detailed Design Phase: Sentiment Analysis of Amazon Reviews

The detailed design phase involves creating a comprehensive architecture for the sentiment analysis system, represented through System Architecture Diagrams, Entity-Relationship Diagrams (ERD), and Data Flow Diagrams (DFD). This ensures a structured, efficient, and scalable solution.

1. Modeling

System Architecture Design

1) Components:
   - Data Acquisition: Module responsible for collecting reviews from Amazon. This involves APIs or web scraping tools.
   - Preprocessing: Cleans and prepares data for analysis. This includes text normalization, tokenization, and stopword removal.
   - Sentiment Analysis Algorithms: Implements chosen algorithms like lexicon-based methods, SVM, and LSTM models.
   - Result Visualization: Presents analysis results using various visualization techniques.

2) Interaction:
   - Data Flow: Data flows from the Data Acquisition module to the Preprocessing module, then to the Sentiment Analysis module, and finally to the Result Visualization module.
   - Feedback Loops: Results from the analysis may feedback into the preprocessing phase for iterative improvement.

Sentiment Analysis Model Design

1) Algorithm Selection:
   - Lexicon-Based Methods: Use predefined sentiment dictionaries.
   - Machine Learning Models: Implement and train models like SVM and deep learning models (e.g., LSTM).
   - Hybrid Approaches: Combine both lexicon-based and machine learning methods to leverage the strengths of each.

2) Model Training:
   - Training Data: Use a labeled dataset from Amazon reviews.
   - Process: Split data into training and testing sets, preprocess the data, and train models using cross-validation techniques.

3) Model Evaluation:
   - Metrics: Use accuracy, precision, recall, and F1-score to evaluate model performance.
   - Techniques: Employ confusion matrices and cross-validation to ensure robustness and generalizability of the models.

Data Visualization Design

1) Visualization Techniques:
   - Bar Charts and Pie Charts: Display sentiment distributions across different products or categories.
   - Heatmaps: Show intensity and distribution of sentiments over time or across categories.

2) User Interaction:
   - Filtering: Allow users to filter results by product category, time period, or sentiment score.
   - Interactivity: Implement features like hover-over details, clickable elements to drill down into specific data points, and adjustable parameters for dynamic analysis.

2. Entity-Relationship Diagram (ERD)

1. Entities:
   - Review: Attributes include review ID, review text, rating, and review date.
   - Product: Attributes include product ID, name, category, and brand.
   - User: Attributes include user ID, username, and demographic information.

2. Relationships:
   - Posted By: Each review is posted by a user, linking Review and User entities.
   - About: Each review is about a product, linking Review and Product entities.

3. Attributes:
   - Review: review_id (PK), review_text, rating, review_date, product_id (FK), user_id (FK).
   - Product: product_id (PK), product_name, category, brand.
   - User: user_id (PK), username, demographic_info.

3. Data Flow Diagram (DFD)

Level 0 DFD

1. Processes:
   - Data Acquisition: Collect reviews from Amazon.
   - Preprocessing: Clean and prepare data.
   - Sentiment Analysis: Perform sentiment classification.
   - Result Visualization: Display analysis results.

2. Data Stores:
   - Review Dataset: Stores raw and processed reviews.
   - Sentiment Results: Stores results of the sentiment analysis.

3. External Entities:
   - Amazon Review Platform: Source of the review data.
   - User Interface: Destination for visualization and interaction with analysis results.

Level 1 DFD

1. Detailed Processes:
   - Text Tokenization: Break down reviews into tokens.
   - Sentiment Score Calculation: Calculate sentiment scores using selected algorithms.
   - Visualization Rendering: Render visualizations for user interaction.

2. Data Flows:
    - Data from Amazon Review Platform: Flows into Data Acquisition.
    - Preprocessed Data: Flows into Sentiment Analysis from Preprocessing.
    - Analysis Results: Flows into Result Visualization from Sentiment Analysis.
    - User Queries: Flows from User Interface to Data Stores and Visualization processes.

By thoroughly designing the system architecture, ERD, and DFD, the project team can create a well-structured and efficient sentiment analysis system for Amazon reviews. This structured approach ensures clarity in implementation, scalability, and effectiveness in processing and analyzing large volumes of review data.

Minimum Hardware Requirements:

1. Processor:
   - Minimum: Intel Core i3 or equivalent
   - Recommended: Intel Core i5 or i7, or equivalent AMD processor for better performance

2. Memory (RAM):
   - Minimum: 4 GB
   - Recommended: 8 GB or more, especially for handling larger datasets and more complex calculations

3. Storage:
   - Minimum: 10 GB of free disk space
   - Recommended: SSD (Solid State Drive) for faster read/write speeds

4. Display:
   - Resolution: At least 1366x768
   - Recommended: Full HD (1920x1080) or higher for better readability and workspace

5. Network:
   - Internet: Reliable internet connection for installing packages and accessing online resources

Minimum Software Requirements:

1. Operating System:
   - Windows: Windows 7 or later
   - MacOS: MacOS 10.12 (Sierra) or later

- Linux: Any modern distribution (e.g., Ubuntu 16.04+)

2. Python:
    - Python Version: Python 3.6 or later

3. Jupyter Notebook Installation:
    - Anaconda Distribution(Recommended): Anaconda includes Python, Jupyter Notebook, and many common data science libraries. It simplifies the setup process.
    - Alternatively, you can install Jupyter Notebook using pip:

```bash
pip install notebook
```

4. Browser:
    - Supported Browsers: Chrome, Firefox, Safari, or Edge

Installation Steps:

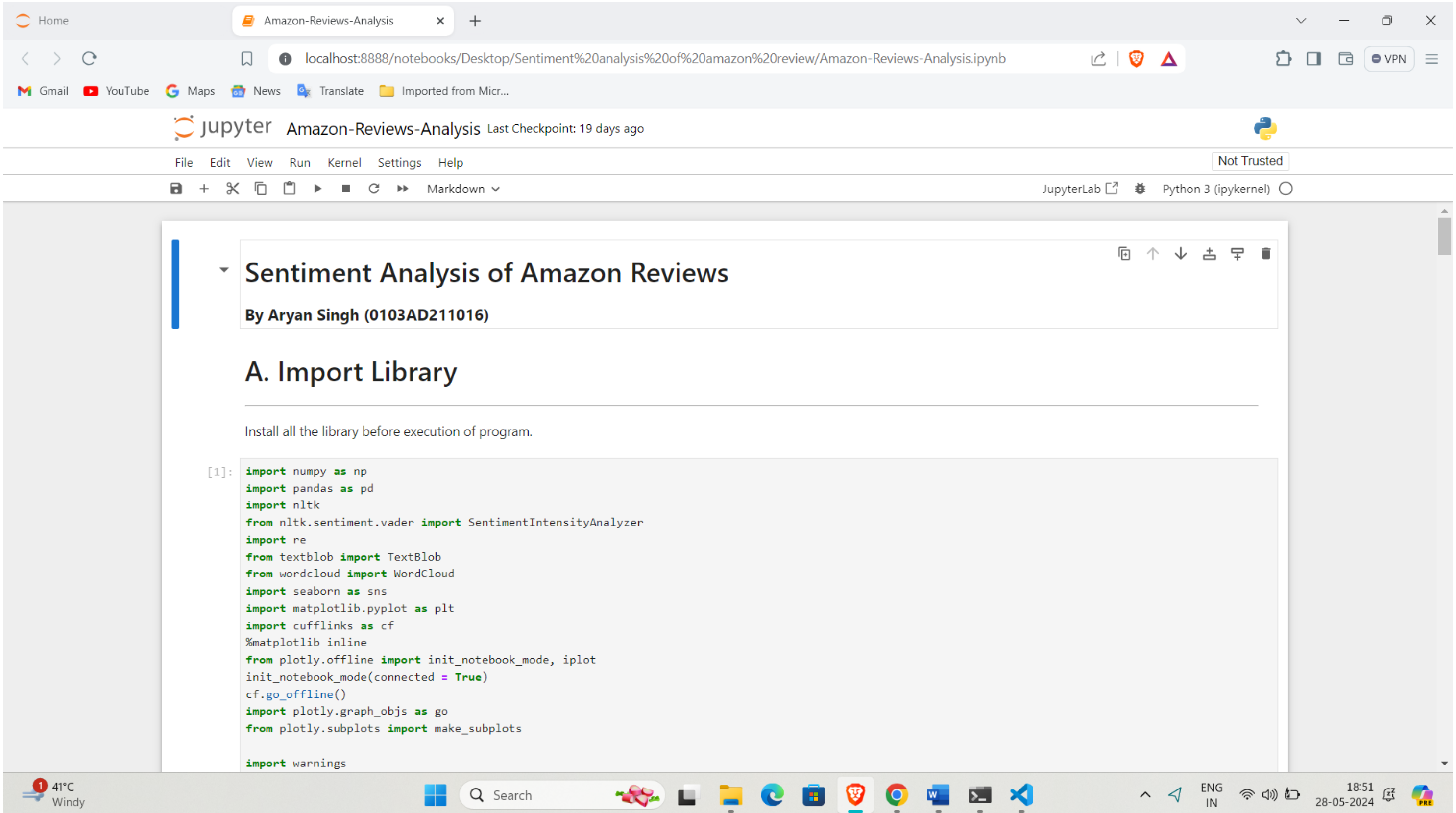
1. Install Anaconda (Recommended):
    - Download the Anaconda distribution from the [official website] (https://www.anaconda.com/products/individual).
    - Follow the installation instructions for your operating system.

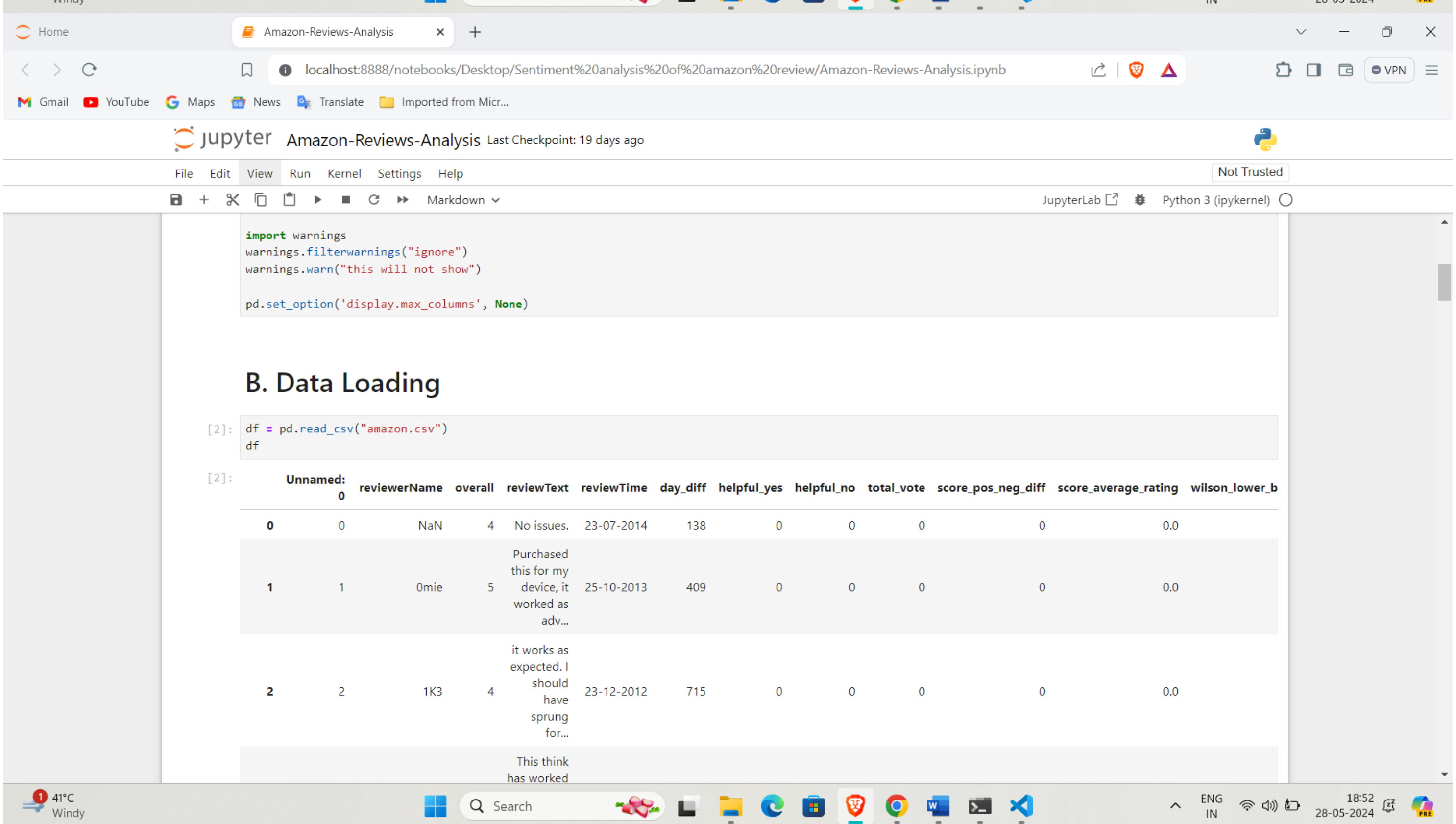
2. Launch Jupyter Notebook:
    - Open Anaconda Navigator and click on the Jupyter Notebook icon.
    - Alternatively, open a terminal (or Anaconda Prompt on Windows) and run:

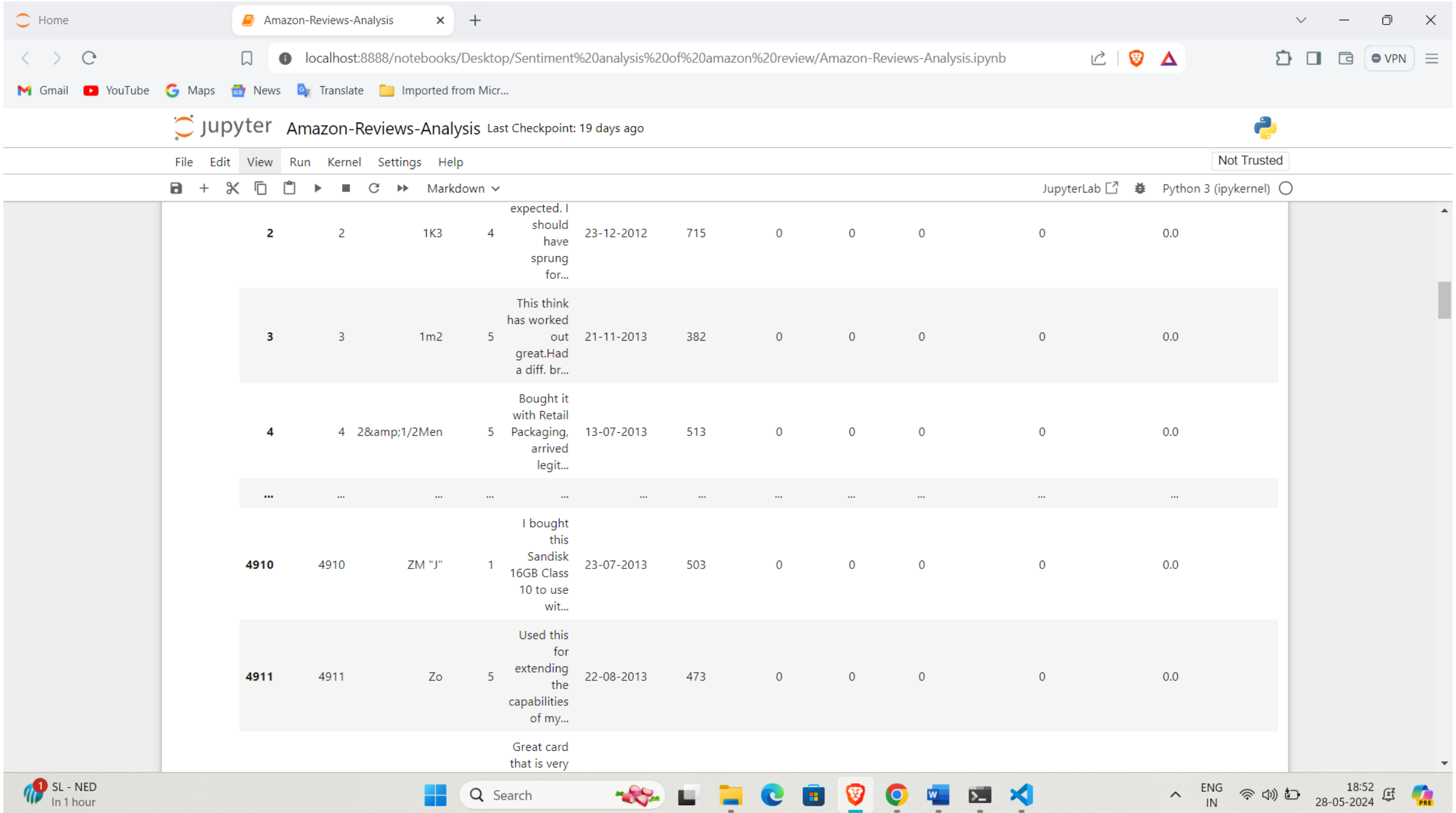
```bash
jupyter notebook
```

3. Access Jupyter Notebook:
    - Jupyter Notebook will open in your default web browser, displaying the notebook dashboard.

SNAPSHOTS OF INPUT AND OUTPUT

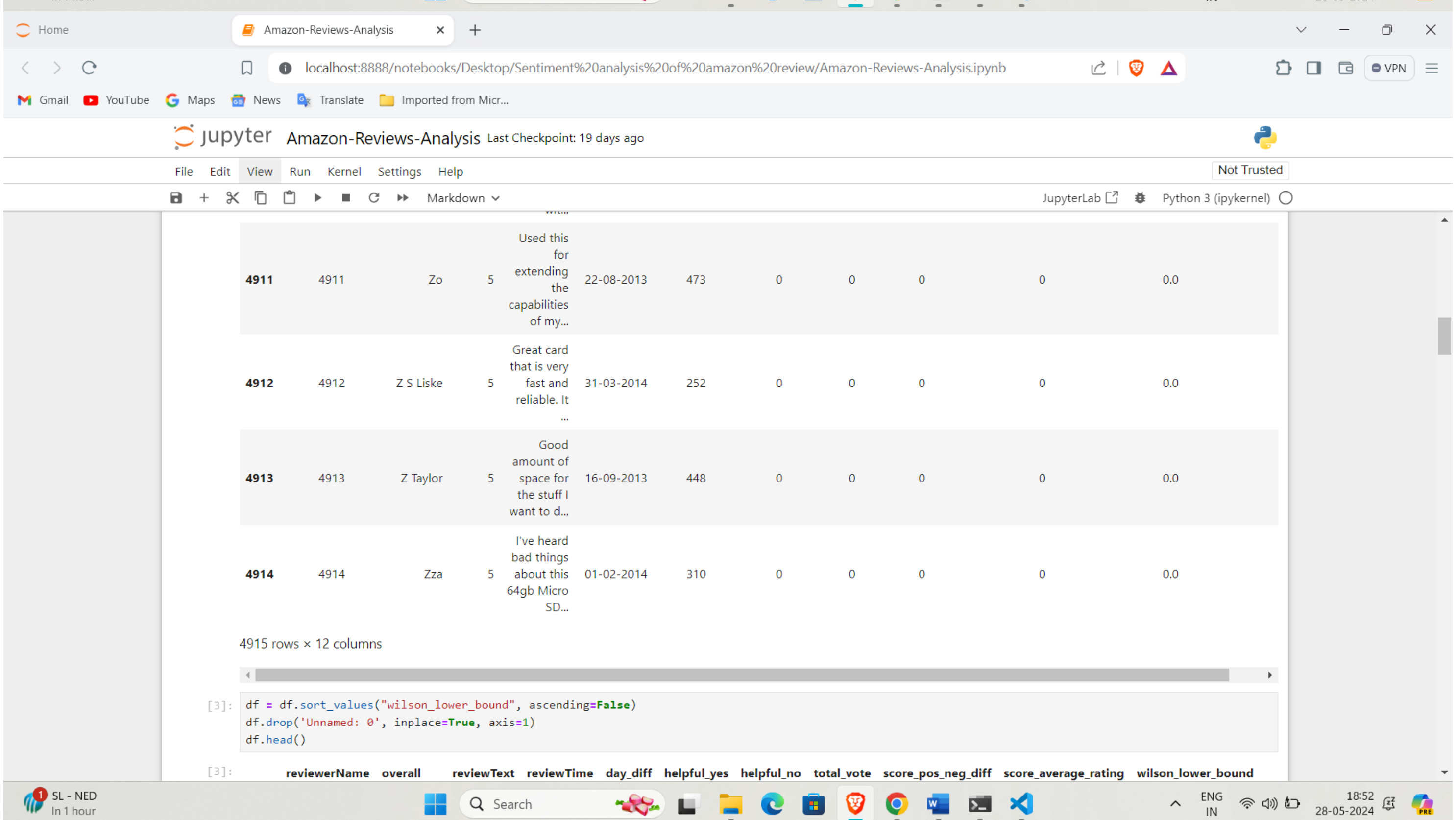






Amazon-Reviews-Analysis Last Checkpoint: 19 days ago

2	2	1K3	4	expected. I should have sprung for...	23-12-2012	715	0	0	0	0	0.0
3	3	1m2	5	This think has worked out great.Had a diff. br...	21-11-2013	382	0	0	0	0	0.0
4	4	2&1/2Men	5	Bought it with Retail Packaging, arrived legit...	13-07-2013	513	0	0	0	0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
4910	4910	ZM "J"	1	I bought this Sandisk 16GB Class 10 to use wit...	23-07-2013	503	0	0	0	0	0.0
4911	4911	Zo	5	Used this for extending the capabilities of my...	22-08-2013	473	0	0	0	0	0.0
				Great card that is very							



Amazon-Reviews-Analysis Last Checkpoint: 19 days ago

4911	4911	Zo	5	Used this for extending the capabilities of my...	22-08-2013	473	0	0	0	0	0.0
4912	4912	Z S Liske	5	Great card that is very fast and reliable. It ...	31-03-2014	252	0	0	0	0	0.0
4913	4913	Z Taylor	5	Good amount of space for the stuff I want to d...	16-09-2013	448	0	0	0	0	0.0
4914	4914	Zza	5	I've heard bad things about this 64gb Micro SD...	01-02-2014	310	0	0	0	0	0.0

4915 rows × 12 columns

```
[3]: df = df.sort_values("wilson_lower_bound", ascending=False)
     df.drop('Unnamed: 0', inplace=True, axis=1)
     df.head()
```

[3]: reviewerName overall reviewText reviewTime day_diff helpful_yes helpful_no total_vote score_pos_neg_diff score_average_rating wilson_lower_bound

Jupyter Amazon-Reviews-Analysis Last Checkpoint: 19 days ago

[3]:
```python
df = df.sort_values("wilson_lower_bound", ascending=False)
df.drop('Unnamed: 0', inplace=True, axis=1)
df.head()
```

[3]:

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no	total_vote	score_pos_neg_diff	score_average_rating	wilson_lower_bound
2031	Hyoun Kim "Faluzure"	5	[[UPDATE - 6/19/2014]]So my lovely wife boug...	05-01-2013	702	1952	68	2020	1884	0.966337	0.957544
3449	NLee the Engineer	5	I have tested dozens of SDHC and micro-SDHC ca...	26-09-2012	803	1428	77	1505	1351	0.948837	0.936519
4212	SkincareCEO	1	NOTE: please read the last update (scroll to ...	08-05-2013	579	1568	126	1694	1442	0.925620	0.912139
317	Amazon Customer "Kelly"	1	If your card gets hot enough to be painful, it...	09-02-2012	1033	422	73	495	349	0.852525	0.818577
4672	Twister	5	Sandisk announcement of the first 128GB micro ...	03-07-2014	158	45	4	49	41	0.918367	0.808109

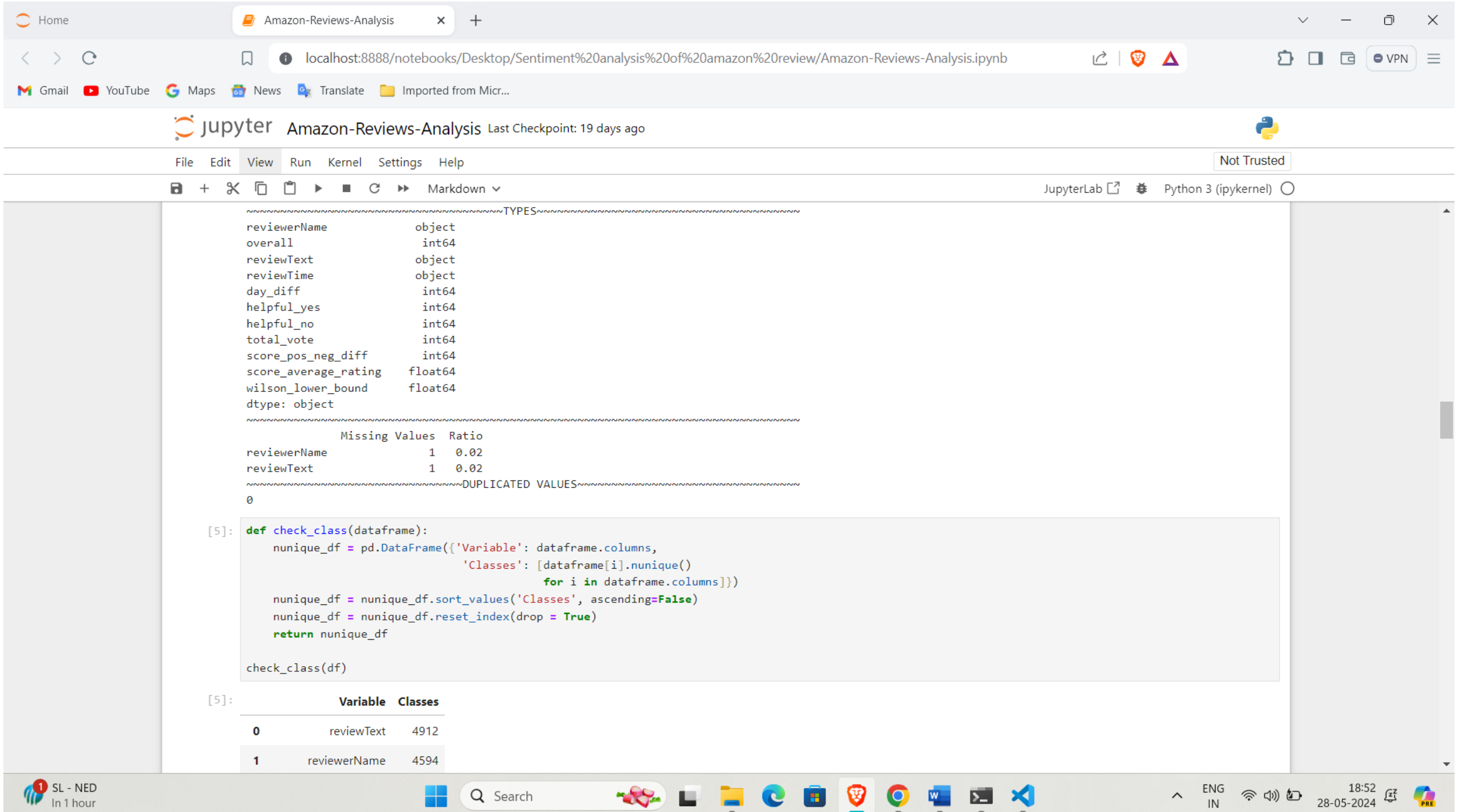
Jupyter Amazon-Reviews-Analysis Last Checkpoint: 19 days ago

[4]:
```python
def missing_values_analysis(df):
    na_columns_ = [col for col in df.columns if df[col].isnull().sum() > 0]
    n_miss = df[na_columns_].isnull().sum().sort_values(ascending=True)
    ratio_ = (df[na_columns_].isnull().sum() / df.shape[0]*100).sort_values(ascending=True)
    missing_df = pd.concat([n_miss, np.around(ratio_, 2)], axis=1, keys=['Missing Values', 'Ratio'])
    missing_df = pd.DataFrame(missing_df)
    return missing_df

def check_dataframe(df, head=5, tail=5):
    print("SHAPE".center(82, '~'))
    print('Rows: {}'.format(df.shape[0]))
    print('Colums: {}'.format(df.shape[1]))
    print("TYPES".center(82,'~'))
    print(df.dtypes)
    print("".center(82, '~'))
    print(missing_values_analysis(df))
    print('DUPLICATED VALUES'.center(82, '~'))
    print(df.duplicated().sum())

check_dataframe(df)
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~SHAPE~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rows: 4915
Colums: 11
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TYPES~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
reviewerName             object
overall                   int64
reviewText               object
reviewTime               object
day_diff                  int64
helpful_yes               int64
helpful_no                int64
total_vote                int64
```



```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TYPES~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
reviewerName             object
overall                   int64
reviewText               object
reviewTime               object
day_diff                  int64
helpful_yes               int64
helpful_no                int64
total_vote                int64
score_pos_neg_diff        int64
score_average_rating    float64
wilson_lower_bound      float64
dtype: object
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
              Missing Values  Ratio
reviewerName               1   0.02
reviewText                 1   0.02
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~DUPLICATED VALUES~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
0
```

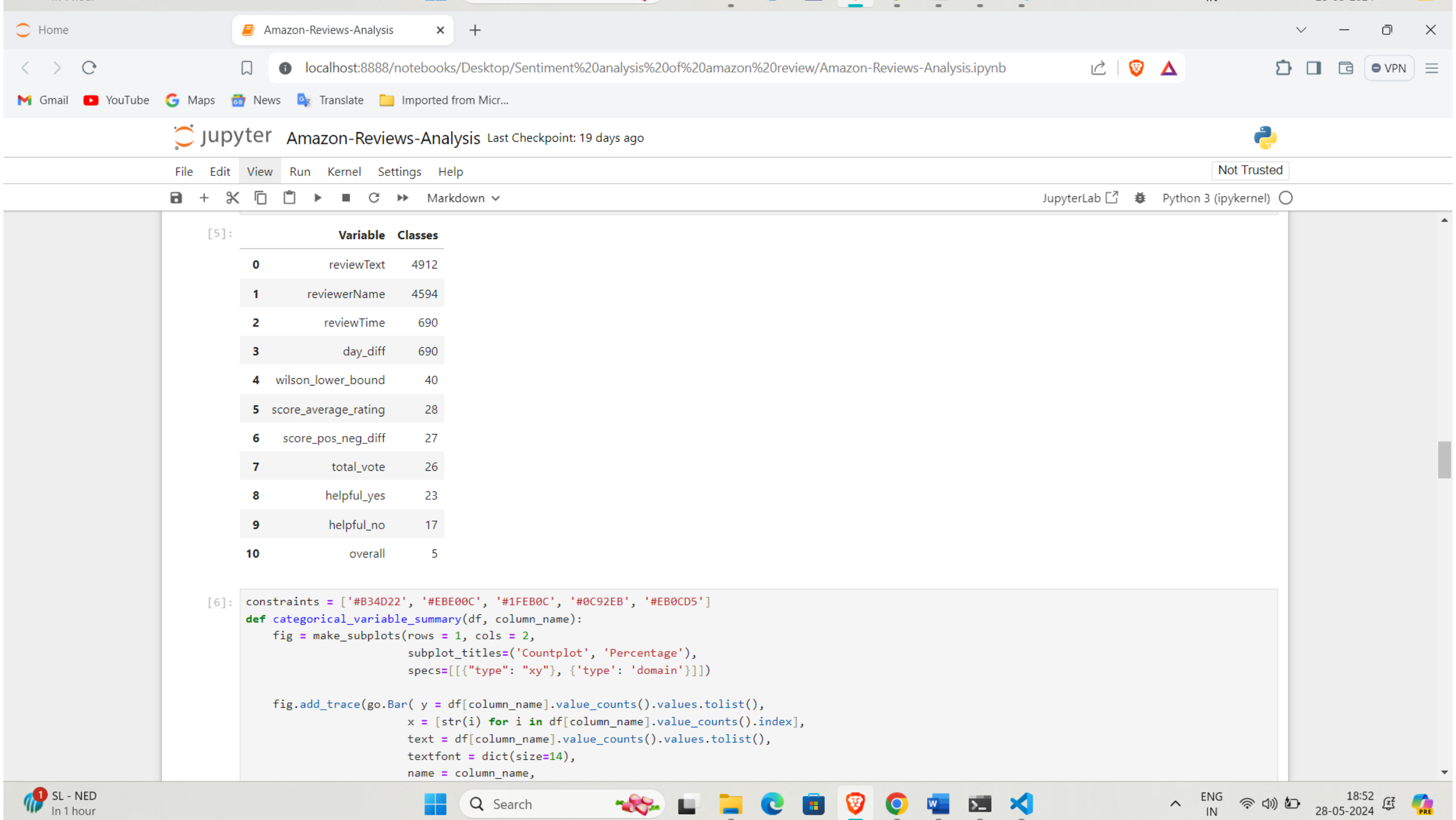
[5]:

```python
def check_class(dataframe):
    nunique_df = pd.DataFrame({'Variable': dataframe.columns,
                               'Classes': [dataframe[i].nunique()
                                           for i in dataframe.columns]})
    nunique_df = nunique_df.sort_values('Classes', ascending=False)
    nunique_df = nunique_df.reset_index(drop = True)
    return nunique_df

check_class(df)
```

[5]:

	Variable	Classes
0	reviewText	4912
1	reviewerName	4594



[5]:

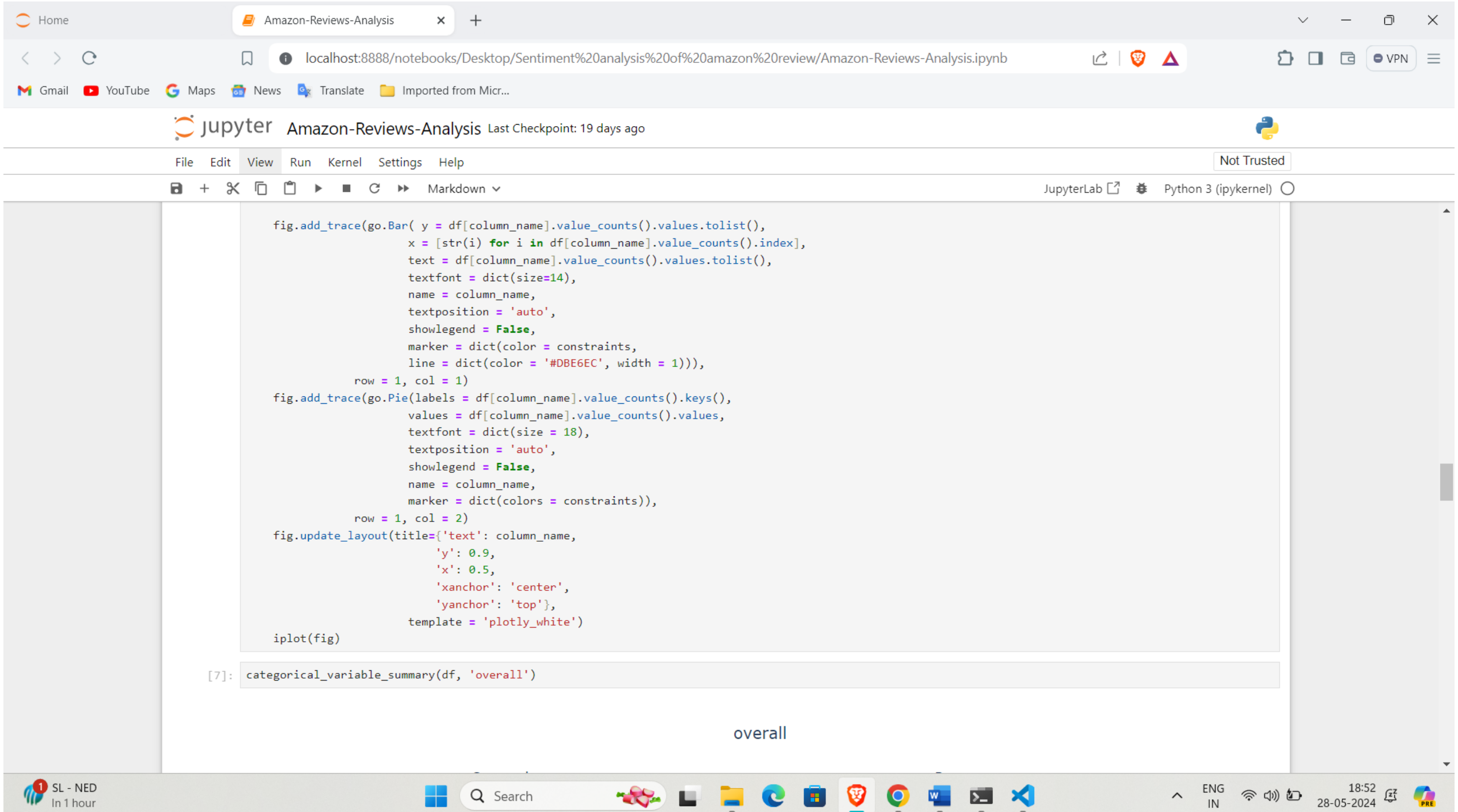
	Variable	Classes
0	reviewText	4912
1	reviewerName	4594

	Variable	Classes
2	reviewTime	690
3	day_diff	690
4	wilson_lower_bound	40
5	score_average_rating	28
6	score_pos_neg_diff	27
7	total_vote	26
8	helpful_yes	23
9	helpful_no	17
10	overall	5

[6]:

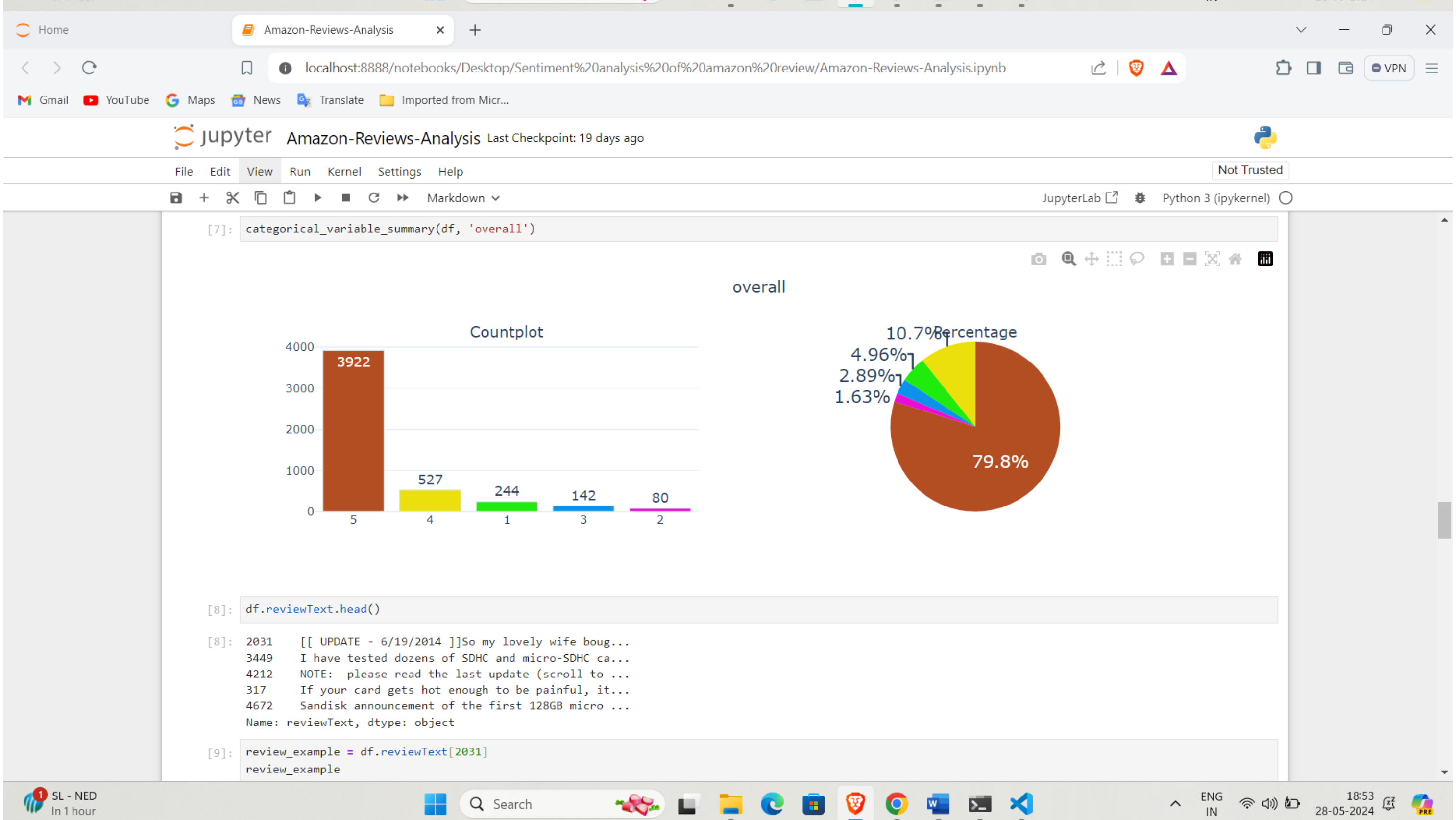
```python
constraints = ['#B34D22', '#EBE00C', '#1FEB0C', '#0C92EB', '#EB0CD5']
def categorical_variable_summary(df, column_name):
    fig = make_subplots(rows = 1, cols = 2,
                        subplot_titles=('Countplot', 'Percentage'),
                        specs=[[{"type": "xy"}, {'type': 'domain'}]])

    fig.add_trace(go.Bar( y = df[column_name].value_counts().values.tolist(),
                        x = [str(i) for i in df[column_name].value_counts().index],
                        text = df[column_name].value_counts().values.tolist(),
                        textfont = dict(size=14),
                        name = column_name,
```



```
    fig.add_trace(go.Bar( y = df[column_name].value_counts().values.tolist(),
                          x = [str(i) for i in df[column_name].value_counts().index],
                          text = df[column_name].value_counts().values.tolist(),
                          textfont = dict(size=14),
                          name = column_name,
                          textposition = 'auto',
                          showlegend = False,
                          marker = dict(color = constraints,
                          line = dict(color = '#DBE6EC', width = 1))),
                  row = 1, col = 1)
    fig.add_trace(go.Pie(labels = df[column_name].value_counts().keys(),
                          values = df[column_name].value_counts().values,
                          textfont = dict(size = 18),
                          textposition = 'auto',
                          showlegend = False,
                          name = column_name,
                          marker = dict(colors = constraints)),
                  row = 1, col = 2)
    fig.update_layout(title={'text': column_name,
                              'y': 0.9,
                              'x': 0.5,
                              'xanchor': 'center',
                              'yanchor': 'top'},
                          template = 'plotly_white')
    iplot(fig)
```

```
[7]: categorical_variable_summary(df, 'overall')
```



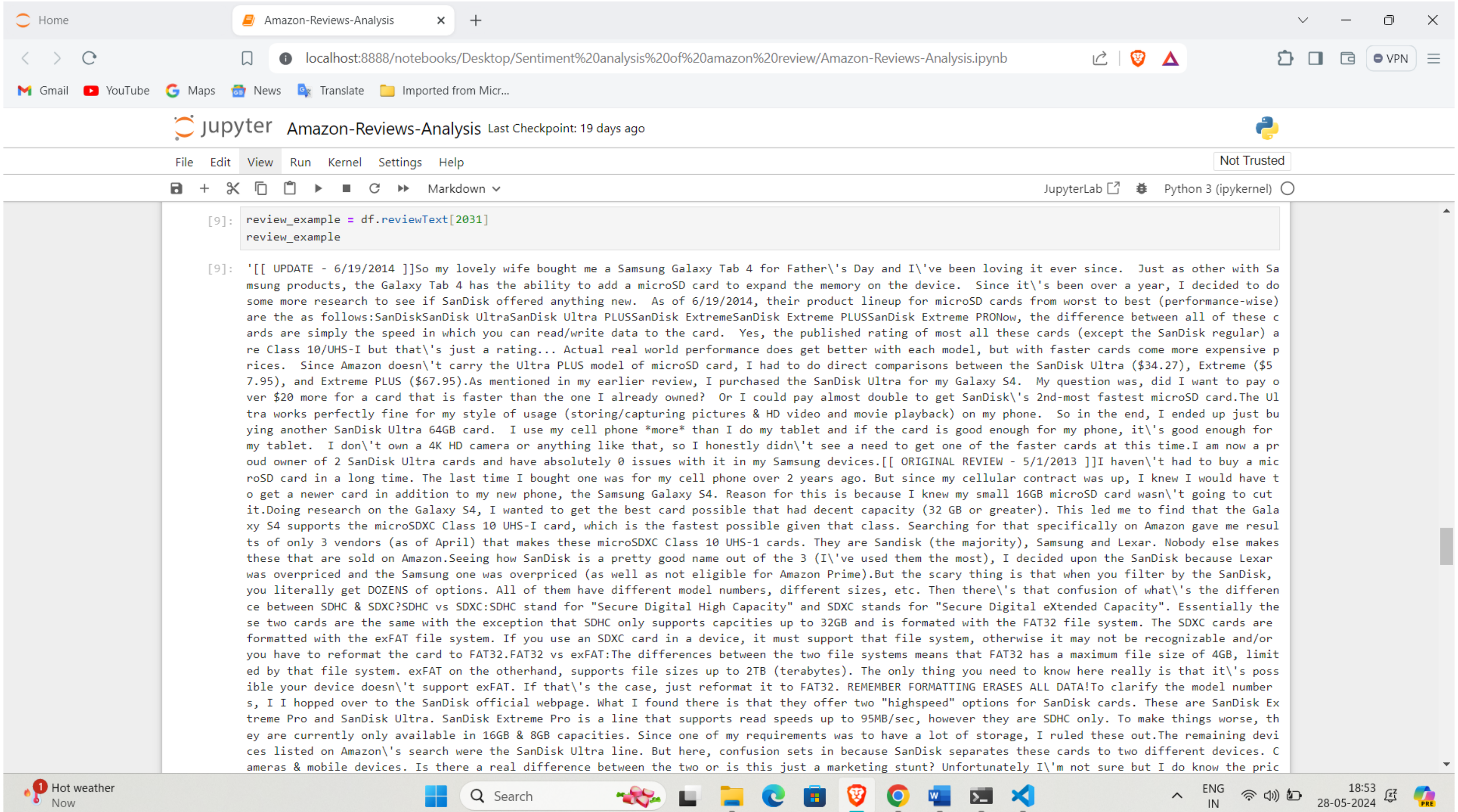
```
[7]: categorical_variable_summary(df, 'overall')
```

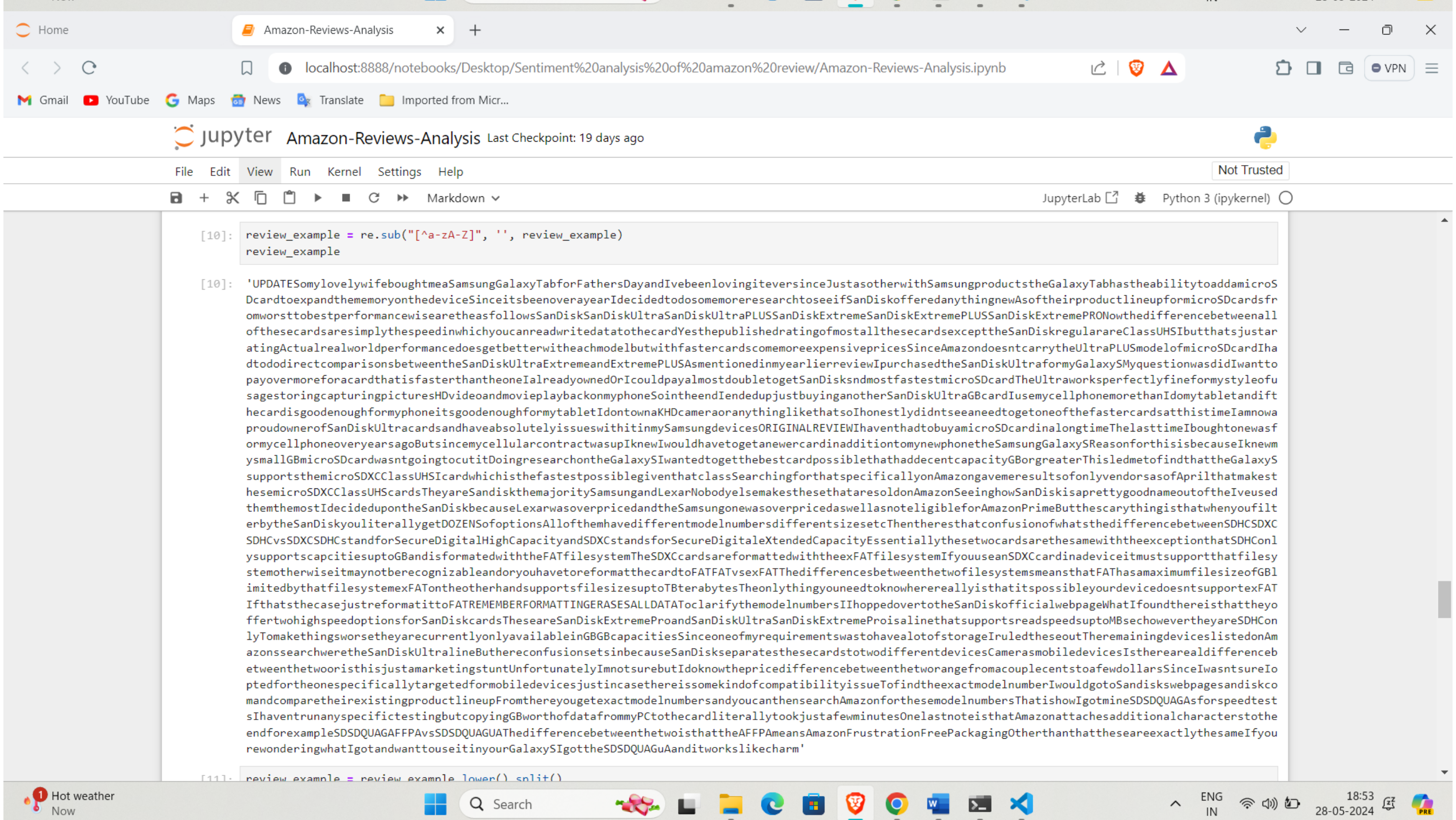


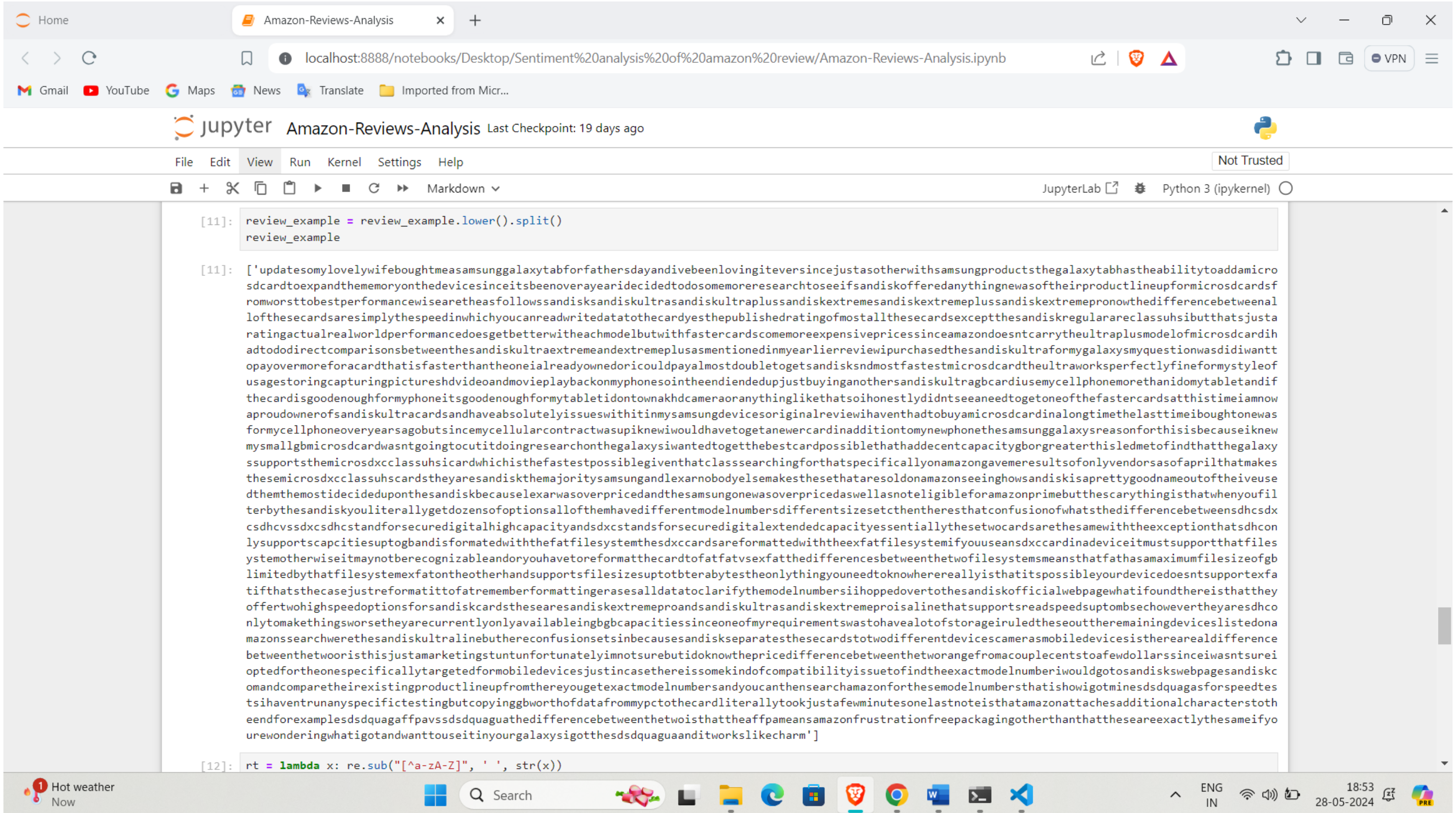
```
[8]: df.reviewText.head()
```

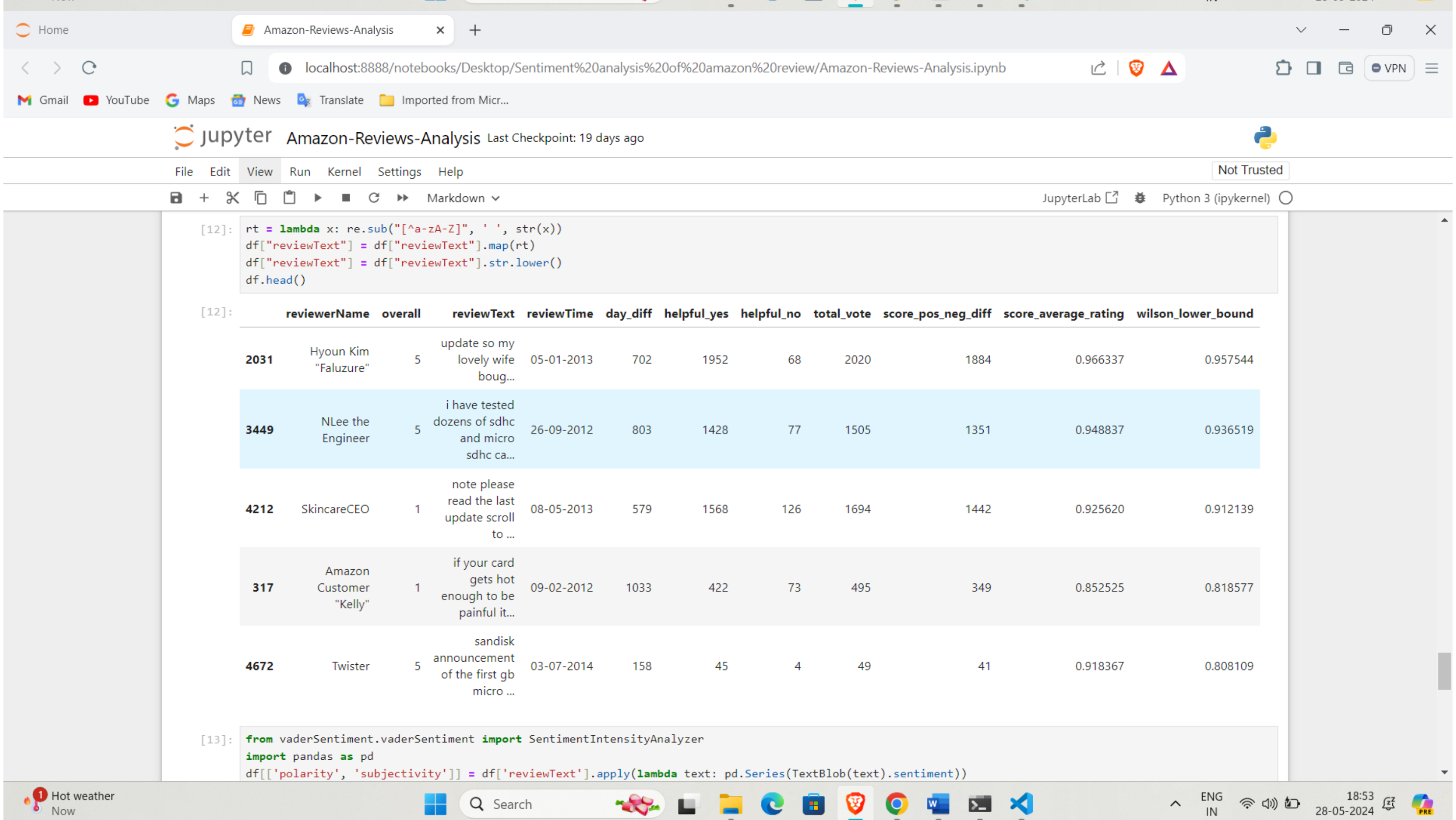
```
[8]: 2031    [[ UPDATE - 6/19/2014 ]]So my lovely wife boug...
     3449    I have tested dozens of SDHC and micro-SDHC ca...
     4212    NOTE:  please read the last update (scroll to ...
     317     If your card gets hot enough to be painful, it...
     4672    Sandisk announcement of the first 128GB micro ...
     Name: reviewText, dtype: object
```

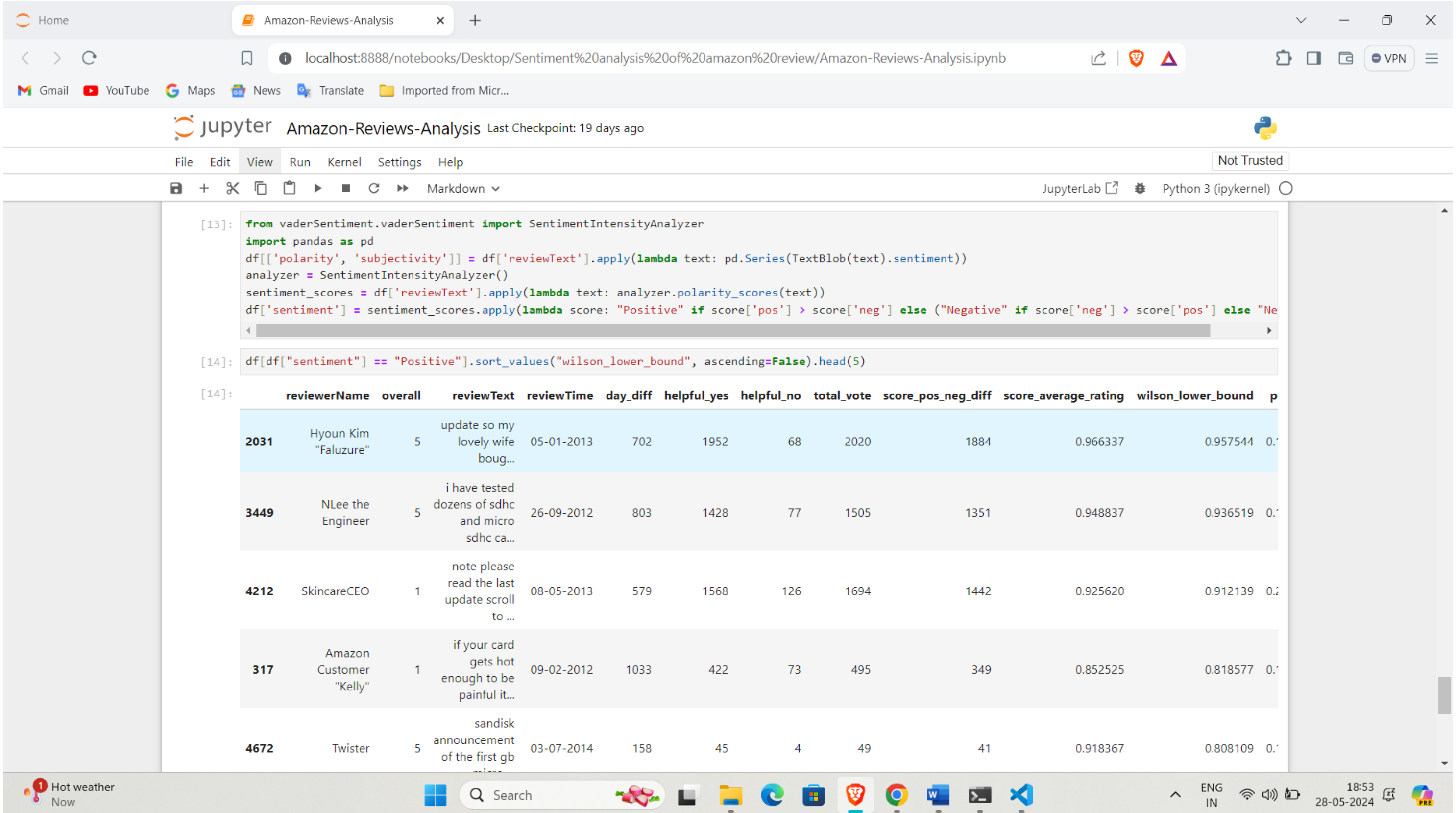
```
[9]: review_example = df.reviewText[2031]
     review_example
```


Home
Amazon-Reviews-Analysis
localhost:8888/notebooks/Desktop/Sentiment%20analysis%20of%20amazon%20review/Amazon-Reviews-Analysis.ipynb
Gmail YouTube Maps News Translate Imported from Micr...
jupyter Amazon-Reviews-Analysis Last Checkpoint: 19 days ago
File Edit View Run Kernel Settings Help
Not Trusted
Markdown
JupyterLab
Python 3 (ipykernel)
[9]: review_example = df.reviewText[2031]
review_example
[9]: '[[UPDATE - 6/19/2014]]So my lovely wife bought me a Samsung Galaxy Tab 4 for Father\'s Day and I\'ve been loving it ever since. Just as other with Samsung products, the Galaxy Tab 4 has the ability to add a microSD card to expand the memory on the device. Since it\'s been over a year, I decided to do some more research to see if SanDisk offered anything new. As of 6/19/2014, their product lineup for microSD cards from worst to best (performance-wise) are the as follows:SanDiskSanDisk UltraSanDisk Ultra PLUSSanDisk ExtremeSanDisk Extreme PLUSSanDisk Extreme PRONow, the difference between all of these cards are simply the speed in which you can read/write data to the card. Yes, the published rating of most all these cards (except the SanDisk regular) are Class 10/UHS-I but that\'s just a rating... Actual real world performance does get better with each model, but with faster cards come more expensive prices. Since Amazon doesn\'t carry the Ultra PLUS model of microSD card, I had to do direct comparisons between the SanDisk Ultra ($34.27), Extreme ($57.95), and Extreme PLUS ($67.95).As mentioned in my earlier review, I purchased the SanDisk Ultra for my Galaxy S4. My question was, did I want to pay over $20 more for a card that is faster than the one I already owned? Or I could pay almost double to get SanDisk\'s 2nd-most fastest microSD card.The Ultra works perfectly fine for my style of usage (storing/capturing pictures & HD video and movie playback) on my phone. So in the end, I ended up just buying another SanDisk Ultra 64GB card. I use my cell phone *more* than I do my tablet and if the card is good enough for my phone, it\'s good enough for my tablet. I don\'t own a 4K HD camera or anything like that, so I honestly didn\'t see a need to get one of the faster cards at this time.I am now a proud owner of 2 SanDisk Ultra cards and have absolutely 0 issues with it in my Samsung devices.[[ORIGINAL REVIEW - 5/1/2013]]I haven\'t had to buy a microSD card in a long time. The last time I bought one was for my cell phone over 2 years ago. But since my cellular contract was up, I knew I would have to get a newer card in addition to my new phone, the Samsung Galaxy S4. Reason for this is because I knew my small 16GB microSD card wasn\'t going to cut it.Doing research on the Galaxy S4, I wanted to get the best card possible that had decent capacity (32 GB or greater). This led me to find that the Galaxy S4 supports the microSDXC Class 10 UHS-I card, which is the fastest possible given that class. Searching for that specifically on Amazon gave me results of only 3 vendors (as of April) that makes these microSDXC Class 10 UHS-1 cards. They are Sandisk (the majority), Samsung and Lexar. Nobody else makes these that are sold on Amazon.Seeing how SanDisk is a pretty good name out of the 3 (I\'ve used them the most), I decided upon the SanDisk because Lexar was overpriced and the Samsung one was overpriced (as well as not eligible for Amazon Prime).But the scary thing is that when you filter by the SanDisk, you literally get DOZENS of options. All of them have different model numbers, different sizes, etc. Then there\'s that confusion of what\'s the difference between SDHC & SDXC?SDHC vs SDXC:SDHC stand for "Secure Digital High Capacity" and SDXC stands for "Secure Digital eXtended Capacity". Essentially these two cards are the same with the exception that SDHC only supports capcities up to 32GB and is formated with the FAT32 file system. The SDXC cards are formatted with the exFAT file system. If you use an SDXC card in a device, it must support that file system, otherwise it may not be recognizable and/or you have to reformat the card to FAT32.FAT32 vs exFAT:The differences between the two file systems means that FAT32 has a maximum file size of 4GB, limited by that file system. exFAT on the otherhand, supports file sizes up to 2TB (terabytes). The only thing you need to know here really is that it\'s possible your device doesn\'t support exFAT. If that\'s the case, just reformat it to FAT32. REMEMBER FORMATTING ERASES ALL DATA!To clarify the model numbers, I I hopped over to the SanDisk official webpage. What I found there is that they offer two "highspeed" options for SanDisk cards. These are SanDisk Extreme Pro and SanDisk Ultra. SanDisk Extreme Pro is a line that supports read speeds up to 95MB/sec, however they are SDHC only. To make things worse, they are currently only available in 16GB & 8GB capacities. Since one of my requirements was to have a lot of storage, I ruled these out.The remaining devices listed on Amazon\'s search were the SanDisk Ultra line. But here, confusion sets in because SanDisk separates these cards to two different devices. Cameras & mobile devices. Is there a real difference between the two or is this just a marketing stunt? Unfortunately I\'m not sure but I do know the pric
Hot weather
Now
Search
ENG IN
18:53
28-05-2024


Home
Amazon-Reviews-Analysis
localhost:8888/notebooks/Desktop/Sentiment%20analysis%20of%20amazon%20review/Amazon-Reviews-Analysis.ipynb
Gmail YouTube Maps News Translate Imported from Micr...
jupyter Amazon-Reviews-Analysis Last Checkpoint: 19 days ago
File Edit View Run Kernel Settings Help
Not Trusted
Markdown
JupyterLab
Python 3 (ipykernel)
[10]: review_example = re.sub("[^a-zA-Z]", '', review_example)
review_example
[10]: 'UPDATESomylovelywifeboughtmeaSamsungGalaxyTabforFathersDayandIvebeenlovingiteversinceJustasotherwithSamsungproductstheGalaxyTabhastheabilitytoaddamicroSDcardtoexpandthememoryonthedeviceSinceitsbeenoverayearIdecidedtodosomemoreresearchtoseeifSanDiskofferedanythingnewAsoftheirproductlineupformicroSDcardsfromworsttobestperformancewisearetheasfollowsSanDiskSanDiskUltraSanDiskUltraPLUSSanDiskExtremeSanDiskExtremePLUSSanDiskExtremePRONowthedifferencebetweenallofthesecardsaresimplythespeedinwhichyoucanreadwritedatatothecardYesthepublishedratingofmostallthesecardsexcepttheSanDiskregulararareClassUHSIbutthatsjustaratingActualrealworldperformancedoesgetbetterwitheachmodelbutwithfastercardscomemoreexpensivepricesSinceAmazondoesntcarrytheUltraPLUSmodelofmicroSDcardIhadtododirectcomparisonsbetweentheSanDiskUltraExtremeandExtremePLUSAsmentionedinmyearlierreviewIpurchasedtheSanDiskUltraformyGalaxySMyquestionwasdidIwanttopayovermoreforacardthatisfasterthantheoneIalreadyownedOrIcouldpayalmostdoubletogetSanDisksndmostfastestmicroSDcardTheUltraworksperfectlyfineformystyleofusagestoringcapturingpicturesHDvideoandmovieplaybackonmyphoneSointheendIendedupjustbuyinganotherSanDiskUltraGBcardIusemycellphonemorethanIdomytabletandifthecardisgoodenoughformyphoneitsgoodenoughformytabletIdontownaKHDcameraoranythinglikethatsoIhonestlydidntseeaneedtogetoneofthefastercardsatthistimeIamnowaproudownerofSanDiskUltracardsandhaveabsolutelyissueswithitinmySamsungdevicesORIGINALREVIEWIhaventhadtobuyamicroSDcardinalongtimeThelasttimeIboughtonewasformycellphoneoveryearsagoButsincemycellularcontractwasupIknewIwouldhavetogetanewercardinadditiontomynewphonetheSamsungGalaxySReasonforthisisbecauseIknewmysmallGBmicroSDcardwasntgoingtocutitDoingresearchontheGalaxySIwantedtogetthebestcardpossiblethathaddecentcapacityGBorgreaterThisledmetofindthattheGalaxySsupportsthemicroSDXCClassUHSIcardwhichisthefastestpossiblegiventhatclassSearchingforthatspecificallyonAmazongavemeresultsofonlyvendorsasofAprilthatmakesthesemicroSDXCClassUHScardsTheyareSandiskthemajoritySamsungandLexarNobodyelsemakesthesethataresoldonAmazonSeeinghowSanDiskisaprettygoodnameoutoftheIveusedthemthemostIdecideduponthe SanDiskbecauseLexarwasoverpricedandtheSamsungonewasoverpricedaswellasnoteligibleforAmazonPrimeButthescarythingisthatwhenyoufilterbytheSanDiskyouliterallygetDOZENSofoptionsAllofthemhavedifferentmodelnumbersdifferentsizesetcThentheresthatconfusionofwhatsthedifferencebetweenSDHCSDXCSDHCvsSDXCSDHCstandforSecureDigitalHighCapacityandSDXCstandsforSecureDigitaleXtendedCapacityEssentiallythesetwocardsarethesamewiththeexceptionthatSDHConlysupportscapcitiesuptoGBandisformatedwiththeFATfilesystemTheSDXCcardsareformattedwiththeexFATfilesystemIfyouuseanSDXCcardinadeviceitmustsupportthatfilesystemotherwiseitmaynotberecognizableandoryouhavetoreformatthecardtoFATFATvsexFATThedifferencesbetweenthetwofilesystemsmeansthatFAThasamaximumfilesizeofGBlimitedbythatfilesystemexFATontheotherhandsupportsfilesizesuptoTBterabytesTheonlythingyouneedtoknowhereallyisthatitspossibleyourdevicedoesntsupportexFATIfthatsthecasejustreformatittoFATREMEMBERFORMATTINGERASESALLDATAToclarifythemodelnumbersIIhoppedovertotheSanDiskofficialwebpageWhatIfoundthereisthattheyoffertwohighspeedoptionsforSanDiskcardsTheseareSanDiskExtremeProandSanDiskUltraSanDiskExtremeProisalinethatsupportsreadspeedsuptoMBsechowevertheyareSDHConlyTomakethingsworsetheyarecurrentlyonlyavailableinGBGBcapacitiesSinceoneofmyrequirementswastohavealotofstorageIruledtheseoutTheremainingdeviceslistedonAmazonssearchweretheSanDiskUltralineButhereconfusionsetsinbecauseSanDiskseparatesthesecardstotwodifferentdevicesCamerasmobiledevicesIstherearealdifferencebetweenthetwooristhisjustamarketingstuntUnfortunatelyImnotsurebutIdoknowthepricedifferencebetweenthetworangefromacouplecentstoafewdollarsSinceIwasntsureIoptedfortheonespecificallytargetedformobiledevicesjustincasethereissomekindofcompatibilityissueTofindtheexactmodelnumberIwouldgotoSandiskswebpagesandiskcomandcomparetheirexistingproductlineupFromthereyougetexactmodelnumbersandyoucanthensearchAmazonforthesemodelnumbersThatishowIgotmineSDSDQUAGAsforspeedtestsIhaventrunanyspecifictestingbutcopyingGBworthofdatafrommyPCtothecardliterallytookjustafewminutesOnelastnoteisthatAmazonattachesadditionalcharacterstotheendforexampleSDSDQUAGAFFPAvsSDSDQUAGUAThedifferencebetweenthetwoisthattheAFFPAmeansAmazonFrustrationFreePackagingOtherthanthattheseareexactlythesameIfyouwonderingwhatIgotandwanttouseitinyourGalaxySIgottheSDSDQUAGuAanditworkslikecharm'
[11]: review_example = review_example.lower().split()
Hot weather
Now
Search
ENG IN
18:53
28-05-2024


Home
Amazon-Reviews-Analysis
localhost:8888/notebooks/Desktop/Sentiment%20analysis%20of%20amazon%20review/Amazon-Reviews-Analysis.ipynb
Gmail
YouTube
Maps
News
Translate
Imported from Micr...
jupyter
Amazon-Reviews-Analysis
Last Checkpoint: 19 days ago
File Edit View Run Kernel Settings Help
Not Trusted
Markdown
JupyterLab
Python 3 (ipykernel)
[11]: review_example = review_example.lower().split()
review_example
[11]: ['updatesomylovelywifeboughtmeasamsunggalaxytabforfathersdayandivebeenlovingiteversincejustasotherwithsamsungproductsthegalaxytabhastheabilitytoaddamicro
sdcardtoexpandthememoryonthedevicesinceitsbeenoverayearidecidedtodosomemoreresearchtoseeifsandiskofferedanythingnewasoftheirproductlineupformicrosdcardsf
romworsttobestperformancewisearetheasfollowssandisksandiskultrasandiskultraplussandiskextremesandiskextremeplussandiskextremepronowthedifferencebetweenal
lofthesecardsaresimplythespeedinwhichyoucanreadwritedatatothecardyesthepublishedratingofmostallthesecardsexceptthesandiskregularareclassuhsibutthatsjusta
ratingactualrealworldperformancedoesgetbetterwitheachmodelbutwithfastercardscomemoreexpensivepricessinceamazondoesntcarrytheultraplusmodelofmicrosdcardih
adtododirectcomparisonsbetweenthesandiskultraextremeandextremeplusasmentionedinmyearlierreviewipurchasedthesandiskultraformygalaxysmyquestionwasdidiwantt
opayovermoreforacardthatisfasterthantheoneialreadyownedoricouldpayalmostdoubletogetsandisksndmostfastestmicrosdcardtheultraworksperfectlyfineformystyleof
usagestoringcapturingpictureshdvideoandmovieplaybackonmyphonesointheendiendedupjustbuyinganothersandiskultragbcardiusemycellphonemorethanidomytabletandif
thecardisgoodenoughformyphoneitsgoodenoughformytabletidontownakhdcameraoranythinglikethatsoihonestlydidntseeaneedtogetoneofthefastercardsatthistimeiamnow
aproudownerofsandiskultracardsandhaveabsolutelyissueswithitinmysamsungdevicesoriginalreviewihaventhadtobuyamicrosdcardinalongtimethelasttimeiboughtonewas
formycellphoneoveryearsagobutsincemycellularcontractwasupiknewiwouldhavetogetanewercardinadditiontomynewphonethesamsunggalaxysreasonforthisisbecauseiknew
mysmallgbmicrosdcardwasntgoingtocutitdoingresearchonthegalaxysiwantedtogetthebestcardpossiblethathaddecentcapacitygborgreaterthisledmetofindthatthegalaxy
ssupportsthemicrosdxcclassuhsicardwhichisthefastestpossiblegiventhatclasssearchingforthatspecificallyonamazongavemeresultsofonlyvendorsasofaprilthatmakes
thesemicrosdxcclassuhscardstheyaresandiskthemajoritysamsungandlexarnobodyelsemakesthesethataresoldonamazonseeinghowsandiskisaprettygoodnameoutoftheiveuse
dthemthemostidecideduponthesandiskbecauselexarwasoverpricedandthesamsungonewasoverpricedaswellasnoteligibleforamazonprimebutthescarythingisthatwhenyoufil
terbythesandiskyouliterallygetdozensofoptionsallofthemhavedifferentmodelnumbersdifferentsizesetcthenthe resthatconfusionofwhatsthedifferencebetweensdhcsdx
csdhcvssdxcsdhcstandforsecuredigitalhighcapacityandsdxcstandsforsecuredigitalextendedcapacityessentiallythesetwocardsarethesamewiththeexceptionthatsdhcon
lysupportscapcitiesuptogbandisformatedwiththefatfilesystemthesdxccardsareformattedwiththeexfatfilesystemifyouuseansdxccardinadeviceitmustsupportthatfiles
ystemotherwiseitmaynotberecognizableandoryouhavetoreformatthecardtofatfatvsexfatthedifferencesbetweenthetwofilesystemsmeansthatfathasamaximumfilesizeofgb
limitedbythatfilesystemexfatontheotherhandsupportsfilesizesuptotbterabytestheonlythingyouneedtoknowherereallyisthatitspossibleyourdevicedoesntsupportexfa
tifthatsthecasejustreformatittofatrememberformattingerasesalldatatoclarifythemodelnumbersiihoppedovertothesandiskofficialwebpagewhatifoundthereisthatthey
offertwohighspeedoptionsforsandiskcardsthesearesandiskextremeproandsandiskultrasandiskextremeproisalinethatsupportsreadspeedsuptombsechowevertheyaresdhco
nlytomakethingsworsetheyarecurrentlyonlyavailableingbgbcapacitiessinceoneofmyrequirementswastohavealotofstorageiruledtheseouttheremainingdeviceslistedona
mazonssearchwerethesandiskultralinebuthereconfusionsetsinbecausesandiskseparatesthesecardstotwodifferentdevicescamerasmobiledevicesistherearealdifference
betweenthetwooristhisjustamarketingstuntunfortunatelyimnotsurebutidoknowthepricedifferencebetweenthetworangefromacouplecentstoafewdollarssinceiwasntsurei
optedfortheonespecificallytargetedformobiledevicesjustincasethereissomekindofcompatibilityissuetofindtheexactmodelnumberiwouldgotosandiskswebpagesandiskc
omandcomparetheirexistingproductlineupfromthereyougetexactmodelnumbersandyoucanthensearchamazonforthesemodelnumbersthatishowigotminesdsdquagasforspeedtes
tsihaventrunanyspecifictestingbutcopyinggbworthofdatafrommypctothecardliterallytookjustafewminutesonelastnoteisthatamazonattachesadditionalcharacterstoth
eendforexamplesdsdquagaffpavssdsdquaguathedifferencebetweenthetwoisthattheaffpameansamazonfrustrationfreepackagingotherthanthattheseareexactlythesameifyo
urewonderingwhatigotandwanttouseitinyourgalaxysigotthesdsdquaguaanditworkslikecharm']
[12]: rt = lambda x: re.sub("[^a-zA-Z]", ' ', str(x))
Hot weather
Now
Search
ENG
IN
18:53
28-05-2024


Home
Amazon-Reviews-Analysis
localhost:8888/notebooks/Desktop/Sentiment%20analysis%20of%20amazon%20review/Amazon-Reviews-Analysis.ipynb
Gmail
YouTube
Maps
News
Translate
Imported from Micr...
jupyter
Amazon-Reviews-Analysis
Last Checkpoint: 19 days ago
File Edit View Run Kernel Settings Help
Not Trusted
Markdown
JupyterLab
Python 3 (ipykernel)
[12]: rt = lambda x: re.sub("[^a-zA-Z]", ' ', str(x))
df["reviewText"] = df["reviewText"].map(rt)
df["reviewText"] = df["reviewText"].str.lower()
df.head()
[12]:
reviewerName overall reviewText reviewTime day_diff helpful_yes helpful_no total_vote score_pos_neg_diff score_average_rating wilson_lower_bound
2031 Hyoun Kim "Faluzure" 5 update so my lovely wife boug... 05-01-2013 702 1952 68 2020 1884 0.966337 0.957544
3449 NLee the Engineer 5 i have tested dozens of sdhc and micro sdhc ca... 26-09-2012 803 1428 77 1505 1351 0.948837 0.936519
4212 SkincareCEO 1 note please read the last update scroll to ... 08-05-2013 579 1568 126 1694 1442 0.925620 0.912139
317 Amazon Customer "Kelly" 1 if your card gets hot enough to be painful it... 09-02-2012 1033 422 73 495 349 0.852525 0.818577
4672 Twister 5 sandisk announcement of the first gb micro ... 03-07-2014 158 45 4 49 41 0.918367 0.808109
[13]: from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import pandas as pd
df[['polarity', 'subjectivity']] = df['reviewText'].apply(lambda text: pd.Series(TextBlob(text).sentiment))
Hot weather
Now
Search
ENG
IN
18:53
28-05-2024

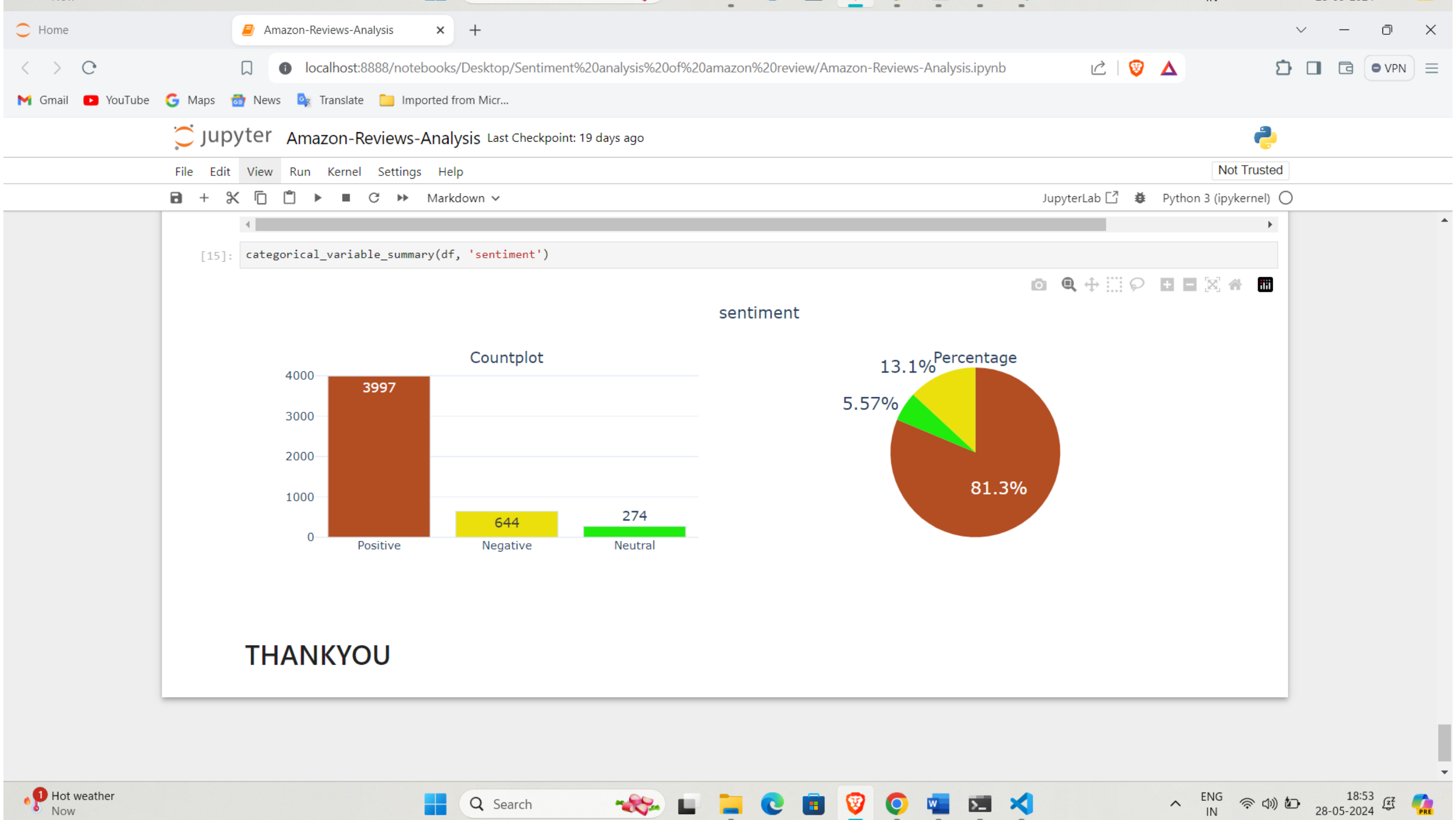


```
[13]: from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import pandas as pd
df[['polarity', 'subjectivity']] = df['reviewText'].apply(lambda text: pd.Series(TextBlob(text).sentiment))
analyzer = SentimentIntensityAnalyzer()
sentiment_scores = df['reviewText'].apply(lambda text: analyzer.polarity_scores(text))
df['sentiment'] = sentiment_scores.apply(lambda score: "Positive" if score['pos'] > score['neg'] else ("Negative" if score['neg'] > score['pos'] else "Ne
```

```
[14]: df[df["sentiment"] == "Positive"].sort_values("wilson_lower_bound", ascending=False).head(5)
```

[14]:

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no	total_vote	score_pos_neg_diff	score_average_rating	wilson_lower_bound	p
2031	Hyoun Kim "Faluzure"	5	update so my lovely wife boug...	05-01-2013	702	1952	68	2020	1884	0.966337	0.957544	0.
3449	NLee the Engineer	5	i have tested dozens of sdhc and micro sdhc ca...	26-09-2012	803	1428	77	1505	1351	0.948837	0.936519	0.
4212	SkincareCEO	1	note please read the last update scroll to ...	08-05-2013	579	1568	126	1694	1442	0.925620	0.912139	0.
317	Amazon Customer "Kelly"	1	if your card gets hot enough to be painful it...	09-02-2012	1033	422	73	495	349	0.852525	0.818577	0.
4672	Twister	5	sandisk announcement of the first gb	03-07-2014	158	45	4	49	41	0.918367	0.808109	0.



```
[15]: categorical_variable_summary(df, 'sentiment')
```


THANKYOU

CODING

```
import numpy as np
import pandas as pd
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import re
from textblob import TextBlob
from wordcloud import WordCloud
import seaborn as sns
import matplotlib.pyplot as plt
import cufflinks as cf
%matplotlib inline
from plotly.offline import init_notebook_mode, iplot
init_notebook_mode(connected = True)
cf.go_offline()
import plotly.graph_objs as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings("ignore")
warnings.warn("this will not show")
pd.set_option('display.max_columns', None)

df = pd.read_csv("amazon.csv")
df

df = df.sort_values("wilson_lower_bound", ascending=False)
df.drop('Unnamed: 0', inplace=True, axis=1)
df.head()

def missing_values_analysis(df):
```

```
    na_columns_ = [col for col in df.columns if df[col].isnull().sum() > 0]
    n_miss = df[na_columns_].isnull().sum().sort_values(ascending=True)



    ratio_ = (df[na_columns_].isnull().sum() / df.shape[0]*100).sort_values(ascending=True)
    missing_df = pd.concat([n_miss, np.around(ratio_, 2)], axis=1, keys=['Missing Values', 'Ratio'])
    missing_df = pd.DataFrame(missing_df)
    return missing_df
def check_dataframe(df, head=5, tail=5):
    print("SHAPE".center(82, '~'))
    print('Rows: {}'.format(df.shape[0]))
    print('Colums: {}'.format(df.shape[1]))
    print("TYPES".center(82,'~'))
    print(df.dtypes)
    print("".center(82, '~'))
    print(missing_values_analysis(df))
    print('DUPLICATED VALUES'.center(82, '~'))
    print(df.duplicated().sum())
check_dataframe(df)

def check_class(dataframe):
    nunique_df = pd.DataFrame({'Variable': dataframe.columns,
                               'Classes': [dataframe[i].nunique()
                                           for i in dataframe.columns]})
    nunique_df = nunique_df.sort_values('Classes', ascending=False)
    nunique_df = nunique_df.reset_index(drop = True)
    return nunique_df
check_class(df)

constraints = ['#B34D22', '#EBE00C', '#1FEB0C', '#0C92EB', '#EB0CD5']
def categorical_variable_summary(df, column_name):
    fig = make_subplots(rows = 1, cols = 2,
                        subplot_titles=('Countplot', 'Percentage'),
                        specs=[[{"type": "xy"}, {'type': 'domain'}]])
```

```
    fig.add_trace(go.Bar( y = df[column_name].value_counts().values.tolist(),
                      x = [str(i) for i in df[column_name].value_counts().index],
                      text = df[column_name].value_counts().values.tolist(),
                      textfont = dict(size=14),
                      name = column_name,
                      textposition = 'auto',
                      showlegend = False,
                      marker = dict(color = constraints,
                      line = dict(color = '#DBE6EC', width = 1))),
              row = 1, col = 1)
    fig.add_trace(go.Pie(labels = df[column_name].value_counts().keys(),
                      values = df[column_name].value_counts().values,
                      textfont = dict(size = 18),
                      textposition = 'auto',
                      showlegend = False,
                      name = column_name,
                      marker = dict(colors = constraints)),
              row = 1, col = 2)
    fig.update_layout(title={'text': column_name,
                          'y': 0.9,
                          'x': 0.5,
                          'xanchor': 'center',
                          'yanchor': 'top'},
                      template = 'plotly_white')
    iplot(fig)

categorical_variable_summary(df, 'overall')


df.reviewText.head()


review_example = df.reviewText[2031]
review_example
```

```
review_example = re.sub("[^a-zA-Z]", '', review_example)
review_example

review_example = review_example.lower().split()
review_example




rt = lambda x: re.sub("[^a-zA-Z]", ' ', str(x))
df["reviewText"] = df["reviewText"].map(rt)
df["reviewText"] = df["reviewText"].str.lower()
df.head()

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import pandas as pd
df[['polarity', 'subjectivity']] = df['reviewText'].apply(lambda text: pd.Series(TextBlob(text).sentiment))
analyzer = SentimentIntensityAnalyzer()
sentiment_scores = df['reviewText'].apply(lambda text: analyzer.polarity_scores(text))
df['sentiment'] = sentiment_scores.apply(lambda score: "Positive" if score['pos'] > score['neg'] else ("Negative" if score['neg'] > score['pos'] else "Neutral"))

df[df["sentiment"] == "Positive"].sort_values("wilson_lower_bound", ascending=False).head(5)

categorical_variable_summary(df, 'sentiment')
```

Project Limitations

1. Data Quality:
   - Impact on Accuracy: Sentiment analysis performance is significantly influenced by data quality. Reviews may contain typos, slang, or irrelevant information (e.g., user discussions not related to the product). These issues can degrade the accuracy of sentiment models.
   - Mitigation: Implementing robust preprocessing techniques can help clean the data. Techniques like spell correction, removal of non-informative text, and normalization are crucial. However, complete elimination of noise is challenging.

2. Language Variations:
   - Challenges: Amazon hosts reviews in multiple languages and dialects, which can be problematic for sentiment models predominantly trained on English data. This can result in inaccurate sentiment detection for non-English reviews.
   - Mitigation: Developing multilingual models or employing translation services before sentiment analysis can help address this issue. Nevertheless, nuances in different languages may still pose challenges.

3. Sarcasm and Irony:
   - Detection Difficulty: Sarcasm and irony are often expressed through context and tone, which are difficult for algorithms to detect. Misinterpretation can lead to incorrect sentiment classification.
   - Mitigation: Advanced techniques like context-aware models, sarcasm detection algorithms, or sentiment analysis with additional context (e.g., user history) can improve detection, but these are complex and computationally intensive.

4. Scalability:
   - Resource Intensive: Processing large-scale datasets requires significant computational resources. Scalability issues can arise in terms of both hardware (e.g., CPU/GPU power, memory) and time required for data processing and model training.

- Mitigation: Utilizing cloud computing services, distributed processing frameworks like Apache Spark, and efficient algorithms can help manage scalability. However, costs and implementation complexity increase with scalability needs.

5. Generalization:
   - Domain Specificity: Models trained on specific datasets might not generalize well to other product categories. Differences in language use and sentiment expression across domains necessitate retraining or fine-tuning of models.
   - Mitigation: Using transfer learning and domain adaptation techniques can help improve generalization. However, these approaches may still require domain-specific data for optimal performance.

Future Scope

1. Multimodal Sentiment Analysis:
   - Expansion: Incorporating text, images, and videos from reviews can provide a more comprehensive sentiment analysis. For instance, analyzing images posted by users along with text reviews can offer deeper insights into product satisfaction.
   - Implementation: Requires developing models that can handle and integrate multiple data types. Techniques like Convolutional Neural Networks (CNNs) for image analysis and combined text-image models will be necessary.

2. Aspect-Based Sentiment Analysis:
   - Granular Insights: By focusing on specific product features (e.g., battery life, camera quality for electronics), aspect-based analysis can provide detailed insights into which aspects are viewed positively or negatively.
   - Implementation: Requires identifying and extracting specific aspects from reviews and then analyzing sentiments related to those aspects. This can be achieved through Natural Language Processing (NLP) techniques like Named Entity Recognition (NER) and topic modeling.

3. Cross-Domain Analysis:
   - Transfer Learning: Enhancing sentiment models to work across different product categories by leveraging knowledge from related domains can improve model applicability.

- Implementation: Techniques like transfer learning and domain adaptation can be employed to adapt models trained in one domain (e.g., electronics) for use in another (e.g., clothing).

4. Real-Time Analysis:

   - Immediate Insights: Developing capabilities for real-time analysis of reviews as they are posted can help businesses respond quickly to customer feedback and emerging trends.
   - Implementation: Requires setting up real-time data pipelines and employing streaming data processing frameworks. Latency and throughput will be critical factors to manage.

5. Integration with Recommendation Systems:

   - Personalization: Using sentiment analysis to enhance recommendation systems can provide more personalized product suggestions based on user sentiments.
   - Implementation: Integrating sentiment scores into recommendation algorithms, possibly using collaborative filtering or content-based filtering techniques, can enhance recommendation relevance.

6. Sentiment Trend Forecasting:

   - Predictive Insights: Using historical sentiment data to forecast future trends can help businesses anticipate market changes and adjust strategies accordingly.
   - Implementation: Time-series analysis and predictive modeling techniques, such as ARIMA (AutoRegressive Integrated Moving Average) or machine learning models like LSTM for sequential data, can be used for trend forecasting.

CHAPTER 10

REFRENCENCES

1. Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the 2008 International Conference on Web Search and Data Mining, 219-230.*
2. Zhang, Y., & Chen, Q. (2015). A novel hybrid method for aspect-based sentiment analysis on product reviews. *Information Sciences, 293, 310-322.*
3. Akhtar, M. S., & Binti Mohamed Arshad, F. H. (2018). Aspect-based sentiment analysis of movie reviews on Twitter using deep learning approaches. *Expert Systems with Applications, 102, 205-213.*
4. Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st International Conference on World Wide Web, 191-200.*
5. Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 497-501.*
6. Xu, S., & Tan, C. L. (2014). Analyzing and predicting question quality in community question-answering services. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 175-184.*
7. Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 804-814.*
8. Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics, 37*(1), 9-27.*
9. Choi, Y., & Cardie, C. (2010). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 793-803.*
10. Liu, K., & Xu, G. (2017). Aspect-level sentiment classification with deep memory network. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 214-224.*