**NLP M1 Report**

**Introduction and Motivation**

In the realm of Natural Language Processing (NLP), the ability to classify text accurately and effectively has become a fundamental task with widespread applications. From sentiment analysis to information retrieval, the capacity to decipher textual data plays a pivotal role in numerous domains. Our project endeavors to delve into this realm by addressing the task of question classification and answering system, utilizing real-world datasets sourced from two prominent StackExchange forums: Android Enthusiasts and Ask Different (Apple).

Choice of Dataset:

Our selection of the "Android vs. iOS" dataset stemmed from a strategic consideration of relevance, accessibility, and potential for exploration. The dataset offers a rich repository of questions posted by users on both Android and iOS platforms, enabling us to conduct comprehensive analyses and develop sophisticated models. By leveraging this dataset, we aim to tackle a significant challenge in NLP—differentiating between questions pertaining to Android and iOS platforms and providing accurate responses.

Relevance and Significance:

The ubiquity of mobile technology, particularly the Android and iOS ecosystems, underscores the importance of understanding user queries and preferences within these domains. As mobile devices continue to shape various aspects of modern life, the ability to categorize and address user inquiries accurately holds immense practical value. Our project seeks to contribute to this endeavor by employing advanced NLP techniques to classify questions and build an effective question-answering system.

Challenges and Opportunities:

The task of question classification and answering system presents several challenges, including handling unstructured text data, mitigating class imbalances, and ensuring model robustness across diverse query types. However, these challenges also represent opportunities for innovation and learning. By navigating through these complexities, we aim to gain insights into the intricacies of text classification and contribute to the development of practical solutions that enhance user experiences in the mobile technology domain.

Project Objectives:

In the subsequent sections of this report, we will delve into a comprehensive literature review, exploring recent research relevant to our problem domain. Additionally, we will conduct a thorough analysis of the dataset at hand, uncovering insights and potential challenges that will inform our approach in the subsequent milestones.

**Literature Review**

Recent advancements in question answering (QA) systems have showcased the potential of natural language processing (NLP) techniques in automating information retrieval and knowledge dissemination. (Lavanya, 2021). Traditional approaches, primarily relying on keyword matching and syntactic search algorithms, have paved the way for more sophisticated methodologies that prioritize semantic understanding and context-based responses ( Beta et. al, 2023 ) . However, challenges persist, particularly in low-resource languages and specialized domains such as mobile technology platforms like Android and iOS.

Addressing Challenges in Low-Resource Languages

Das and Saha (2022) address the challenge of building QA systems in low-resource languages like Bengali by employing supervised learning algorithms and leveraging machine-readable dictionaries such as WordNet. Their system achieves high accuracy in question classification and answer retrieval, demonstrating the feasibility of building comprehensive QA systems in languages with limited linguistic resources. The adoption of supervised learning methods and the use of a text corpus as the system's repository underscore the importance of utilizing available resources effectively in overcoming language-specific challenges.

Harnessing Deep Learning and NLP Techniques

Tzu-Hsuan Lin et al. (2022) leverage deep learning and NLP techniques, specifically the Bidirectional Encoder Representations from Transformers (BERT) model, to develop an intelligent question and answer system tailored to the construction industry's needs. By integrating BERT with a mobile chatbot interface, their system enables conversational machine understanding and facilitates user-friendly information searches in the context of building information modeling and artificial intelligence of things (BIM-AIOT). The utilization of machine learning models and NLP techniques enables accurate prediction and efficient information retrieval, empowering professionals in the Architecture, Engineering, and Construction (AEC) domain to make informed decisions swiftly.

Implications for Android vs. iOS Question Classification

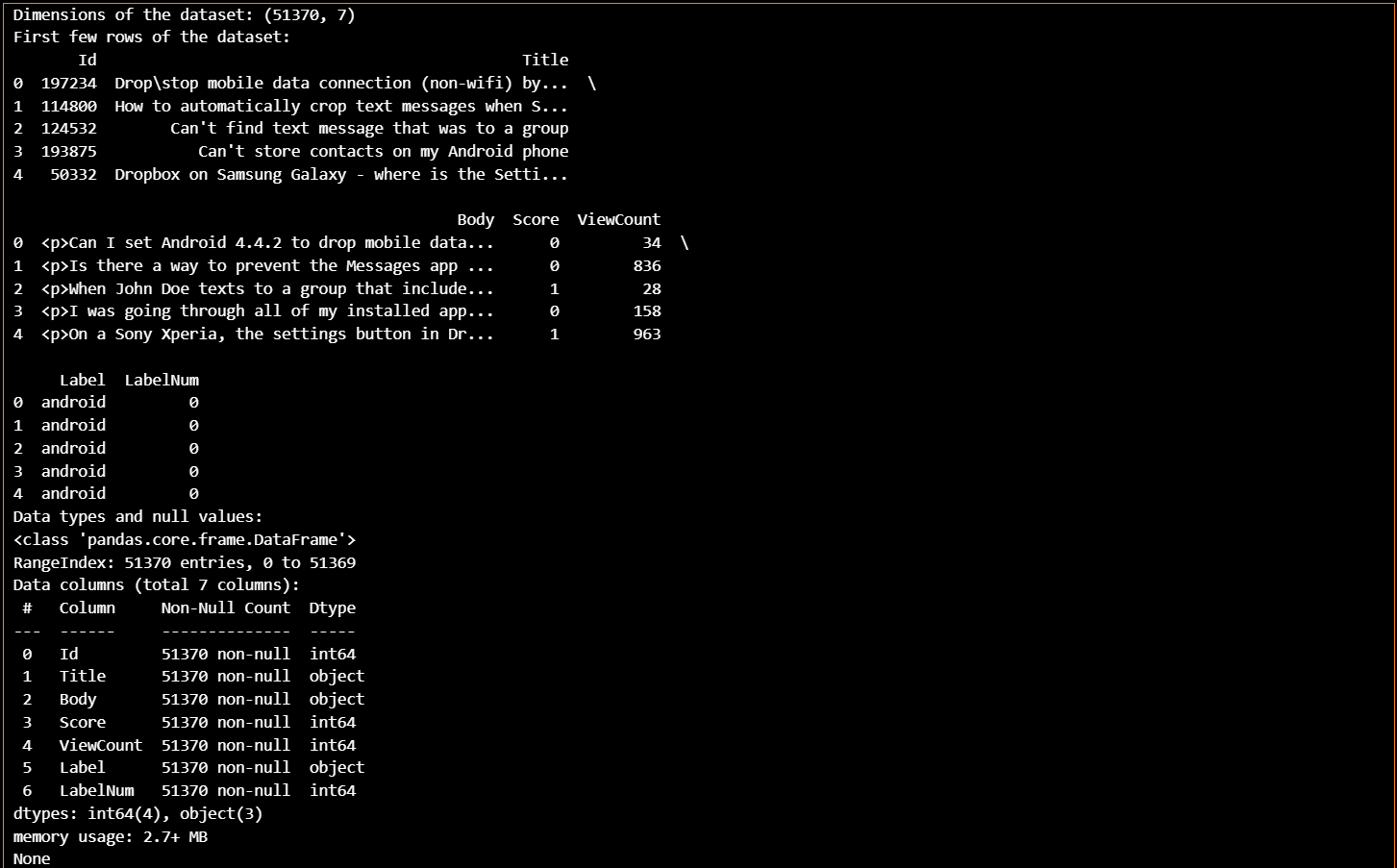
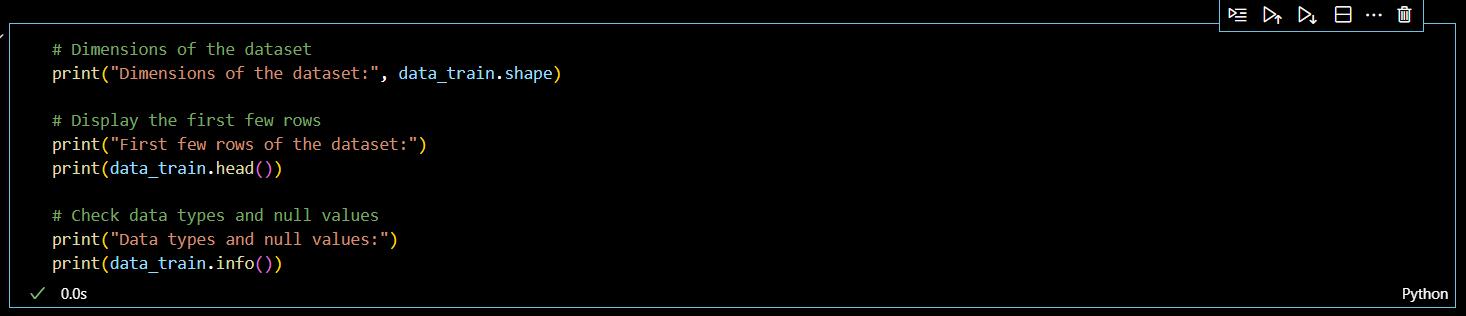
While the aforementioned studies focus on diverse domains such as language-specific QA systems and BIM-AIOT integration, their methodologies and insights offer valuable lessons for the task of classifying questions related to Android and iOS support platforms. By adopting supervised learning algorithms, deep learning models, and NLP techniques, researchers can enhance the accuracy and robustness of question classification systems, enabling more effective differentiation between user queries in the mobile technology domain. Furthermore, the utilization of large text corpora and domain-specific knowledge repositories can facilitate the development of comprehensive QA systems tailored to the intricacies of Android and iOS platforms.

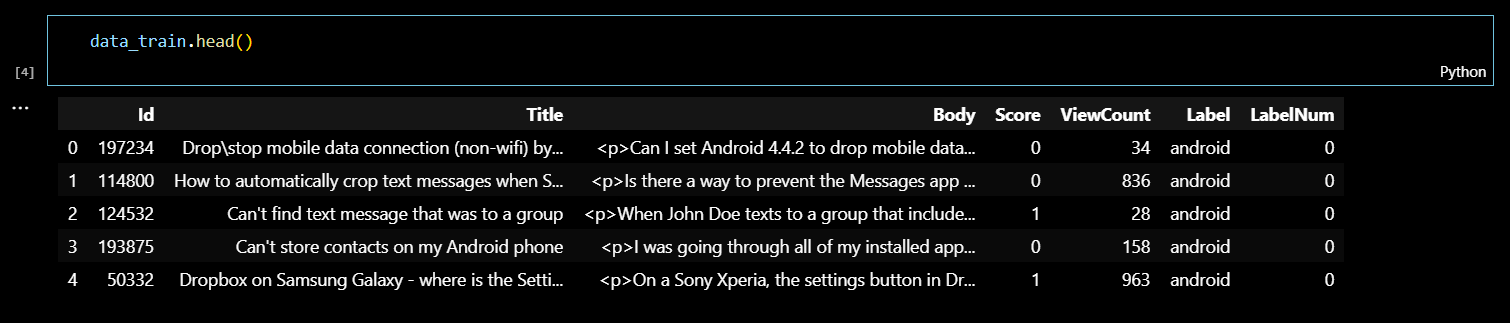
Conclusion

In conclusion, recent advancements in QA systems, exemplified by the works of Das and Saha (2022) and Tzu-Hsuan Lin et al. (2022), demonstrate the transformative potential of deep learning, NLP, and supervised learning techniques in automating information retrieval and knowledge dissemination across diverse domains. By leveraging these methodologies and insights, researchers can tackle the challenges of question classification in specialized domains such as Android and iOS platforms, paving the way for more accurate and efficient question answering systems tailored to the needs of modern mobile technology users.

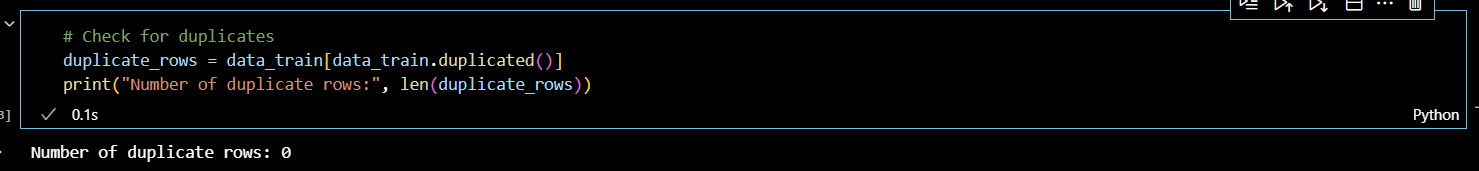
**Data analysis and insights**

For the following part, we will be discussing and analyzing all aspects of the dataset, while also extracting as many insights as possible.

* **Value analysis**  
  Starting with the basics, we used simple pandas methods to view the dataset dimensions, the first few rows, the data types used and null occurrences.As shown, the dataset has 51370 entries and 7 columns. Luckily, this dataset is clean from null values.

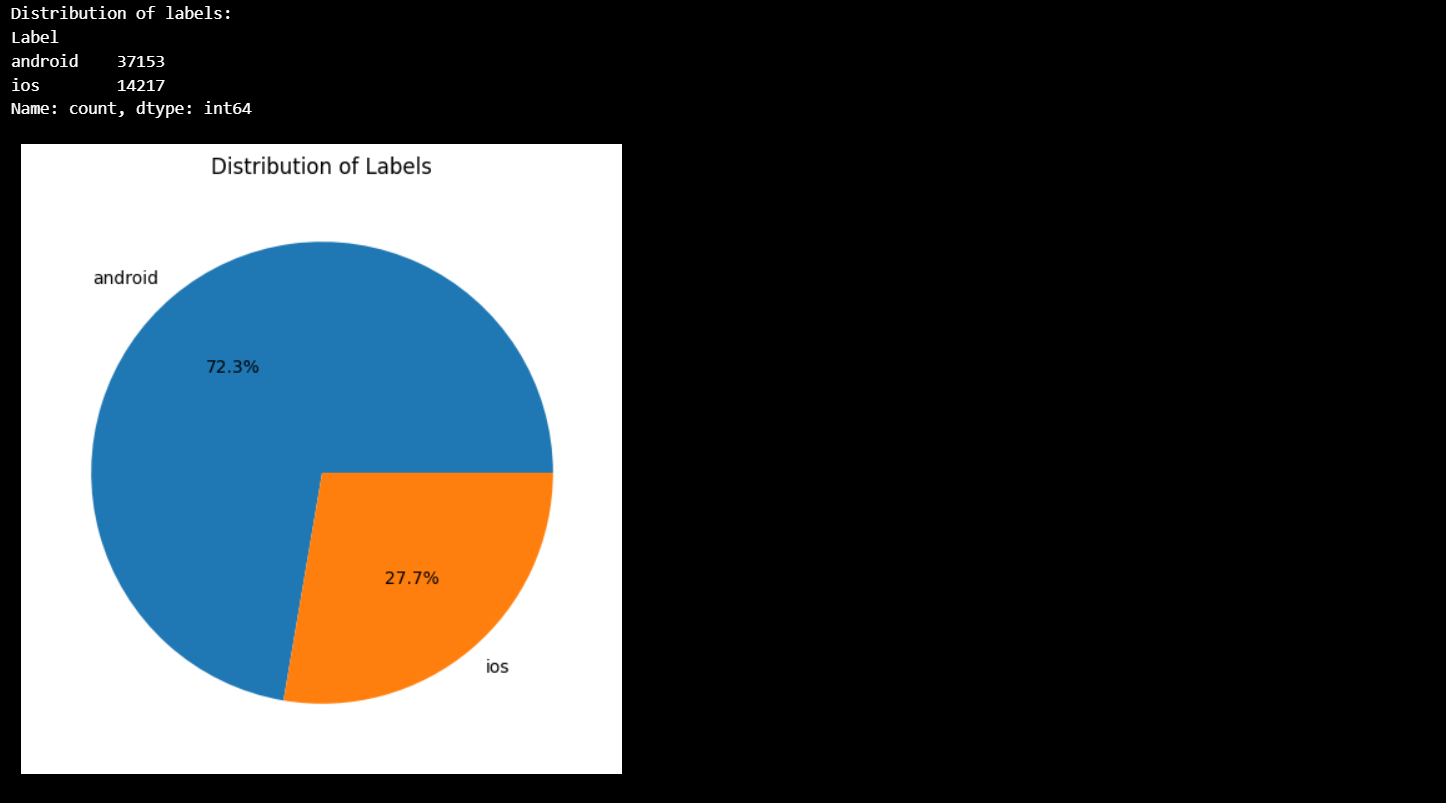
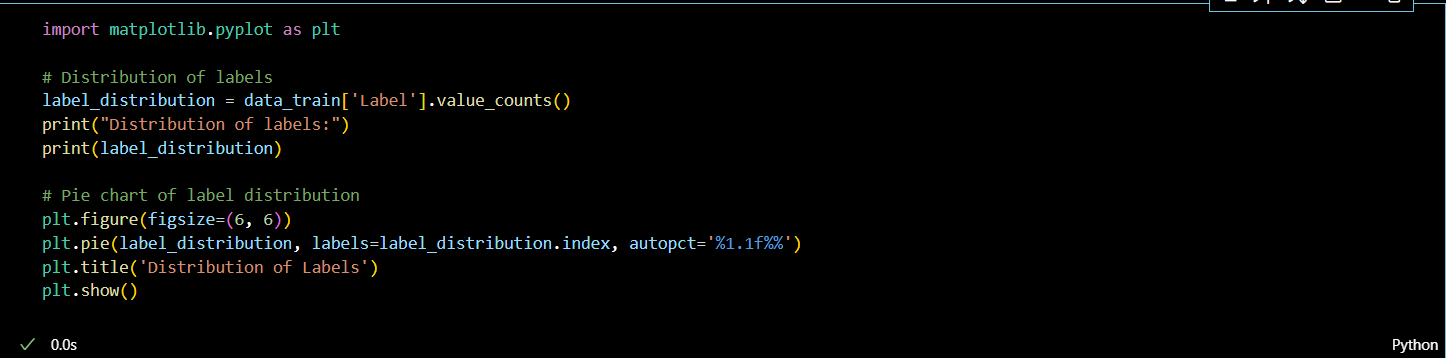
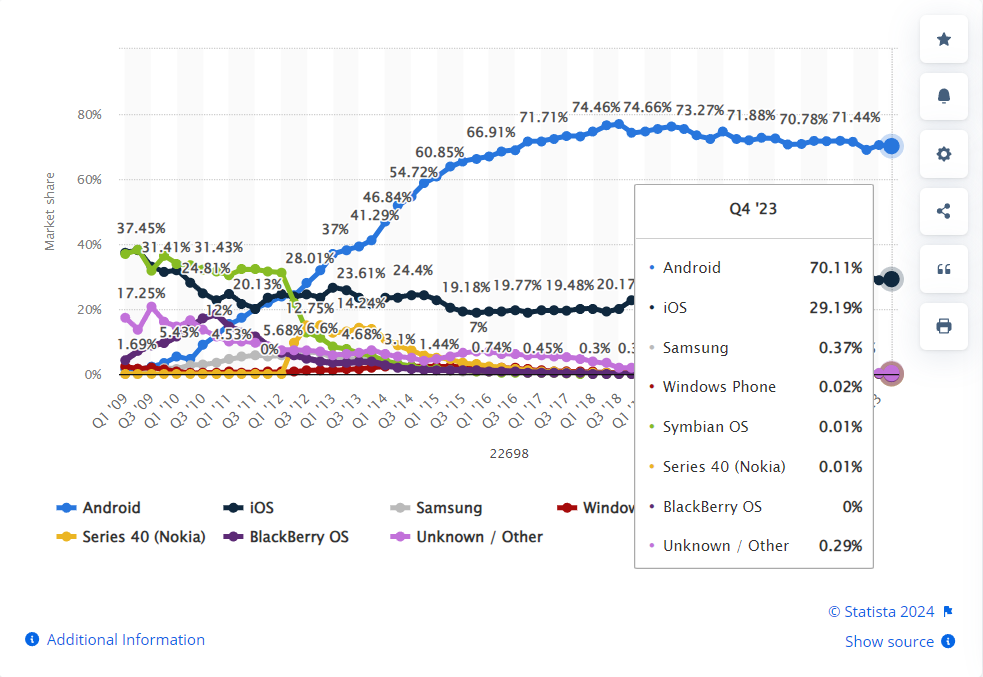
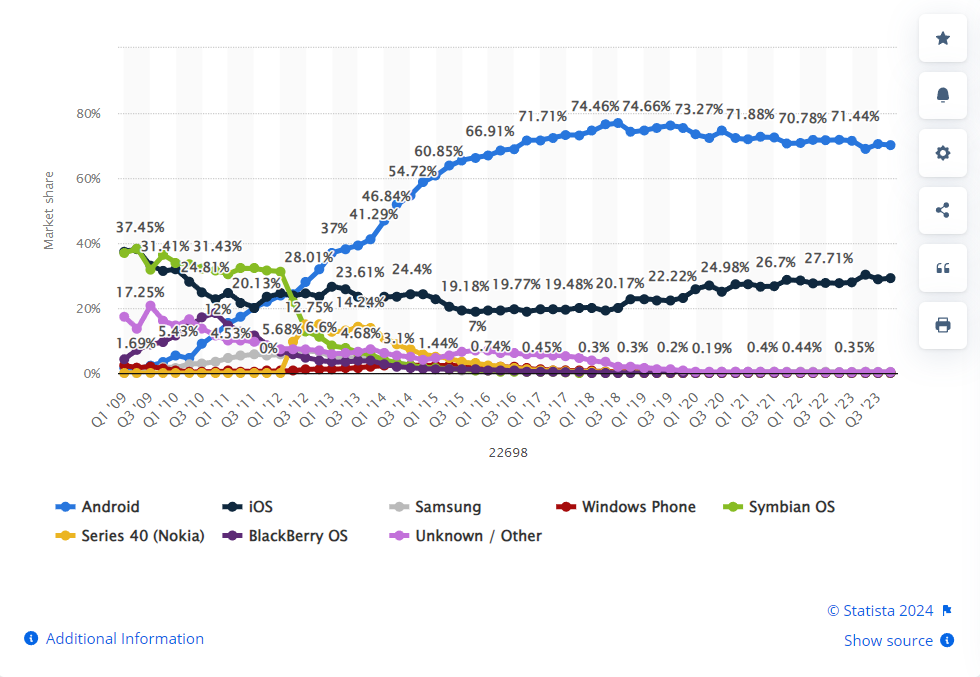
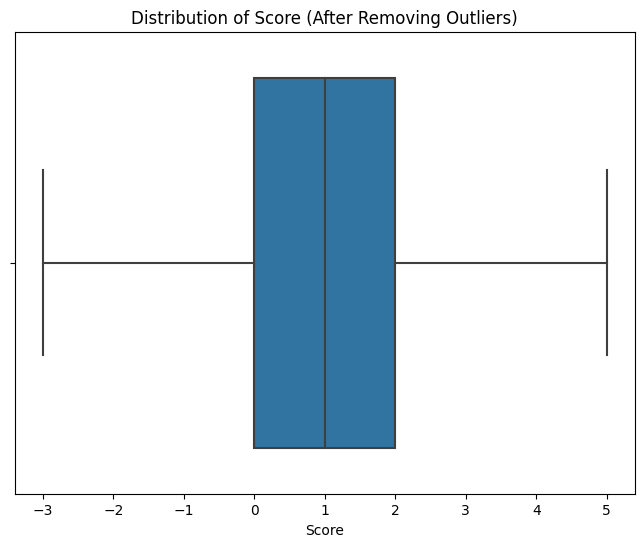
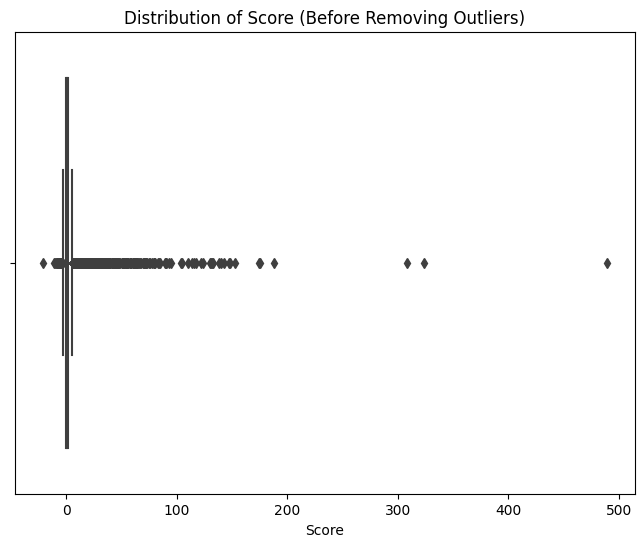
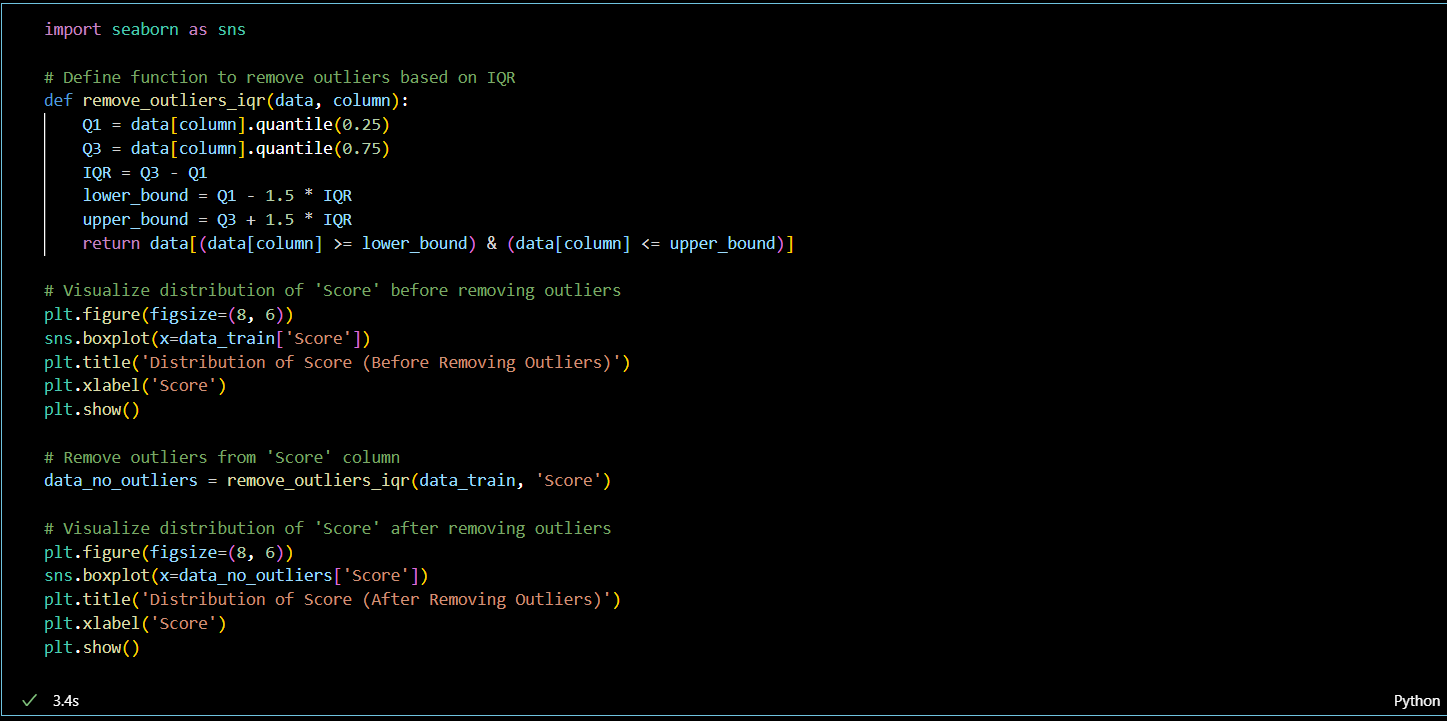
The following is just a clearer sample of the dataset.As seen, the seven columns are:

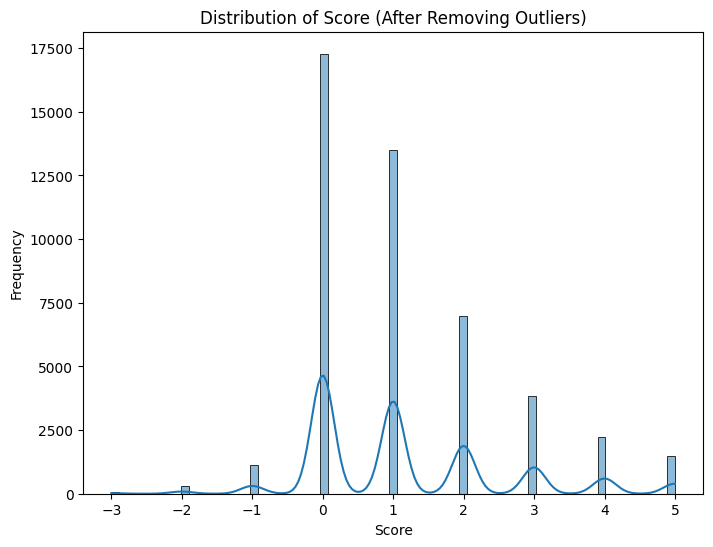
* Id: question id, each question has a unique value.
* Title: title of the question asked.
* Body: questions body related to the title.
* Score: score assigned to the the questions.
* ViewCount: number of views the question got.
* Label: classification label, whether the question is related to android or ios
* LabelNum: a one-hot encoding of the “Label” column, basically a binary classification indicator.

Also, no duplicates are present in the dataset.So, for the primary analysis, it is clear that the dataset doesn’t contain any troubling entries.

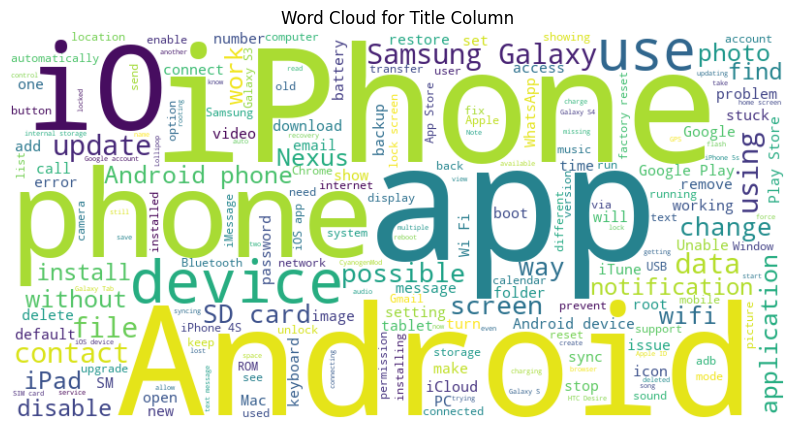
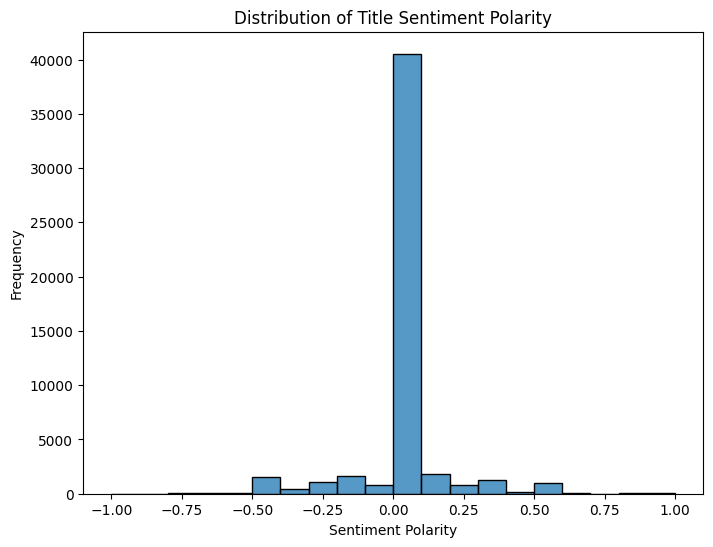
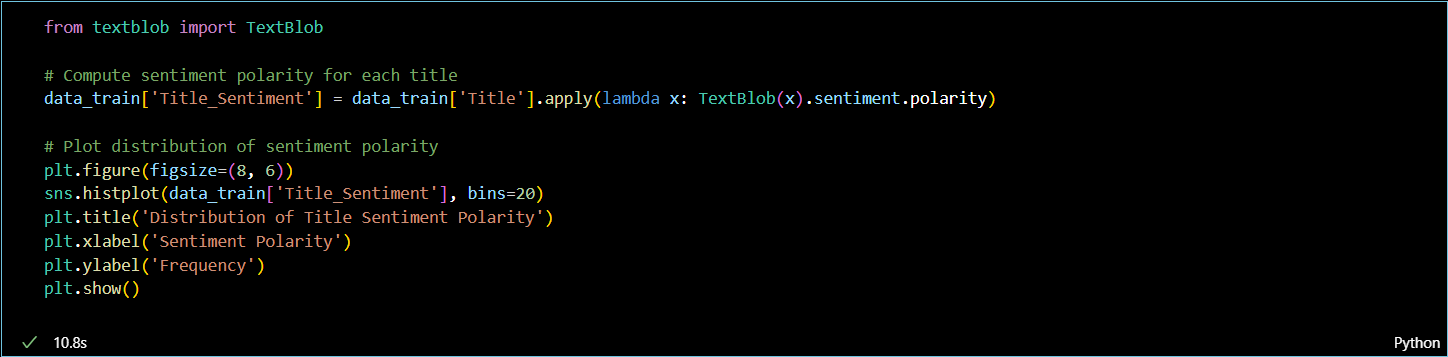
* **Exploratory Data Analysis (EDA)**

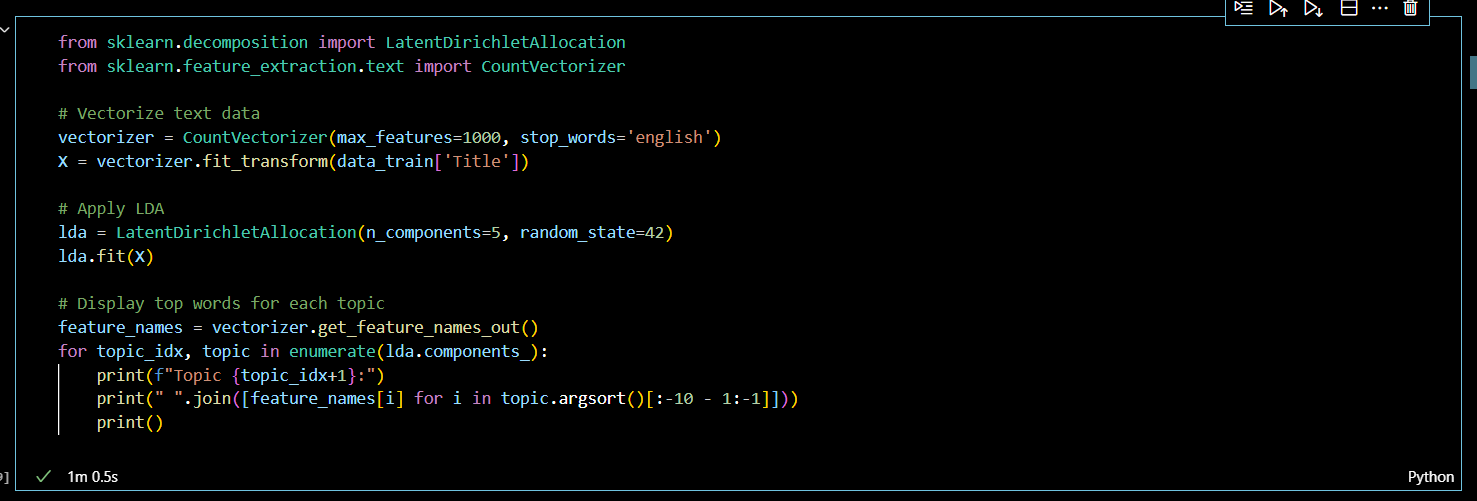
The exploratory data analysis will include analysis of different columns and visualization to help drive all useful insights that can be captured from the dataset.

* **Label distribution**The dataset holds 37153 questions related to android (72.3%) and 14217 questions related to ios (27.7%) , which clearly shows that the dataset is biased towards android questions. This bias is logical because per the latest statistics provided by Statista , android market share is 70.11% while IOS is 29.19%.  
  But, this bias is to be considered during the training of the model.
* **Score distribution**The left box-plot shows the distribution of score before removing outliers, and the right box-plot shows the distribution of values after removing outliers. Outliers were removed using the Interquartile Range (IQR) method. It is a measure of statistical dispersion, or spread, which is used to identify the extent of spread in the middle 50% of a dataset. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the data.

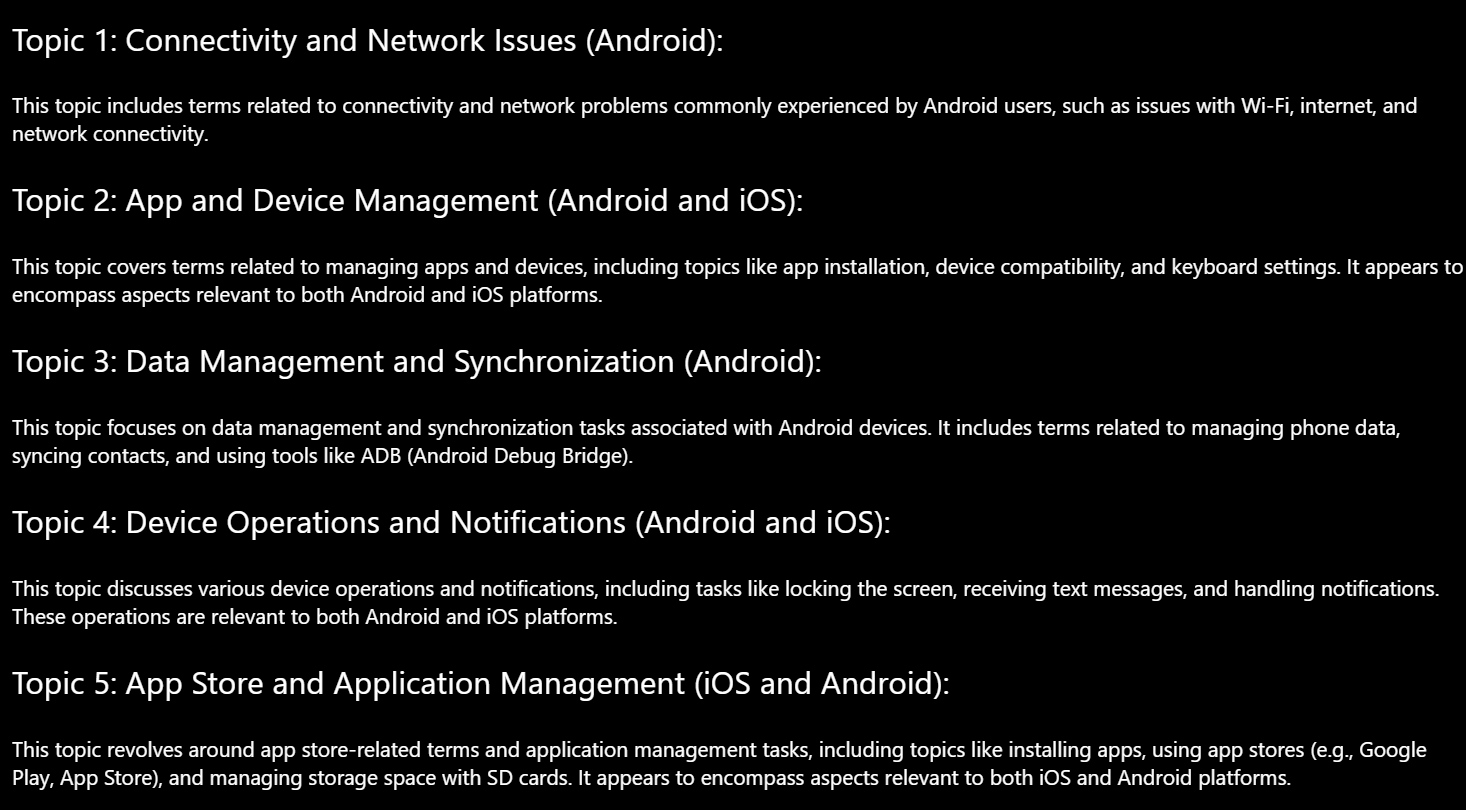
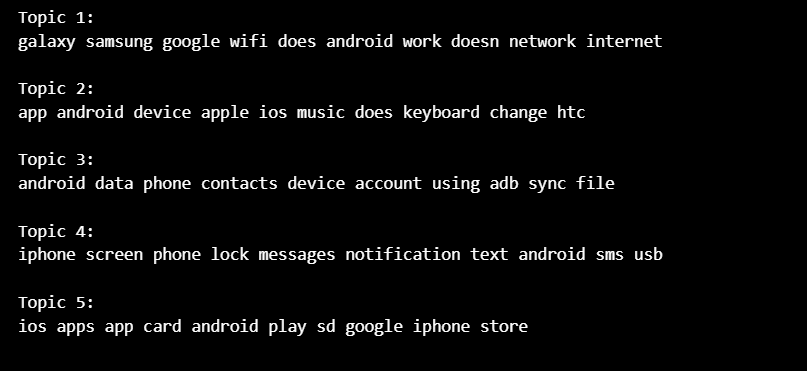


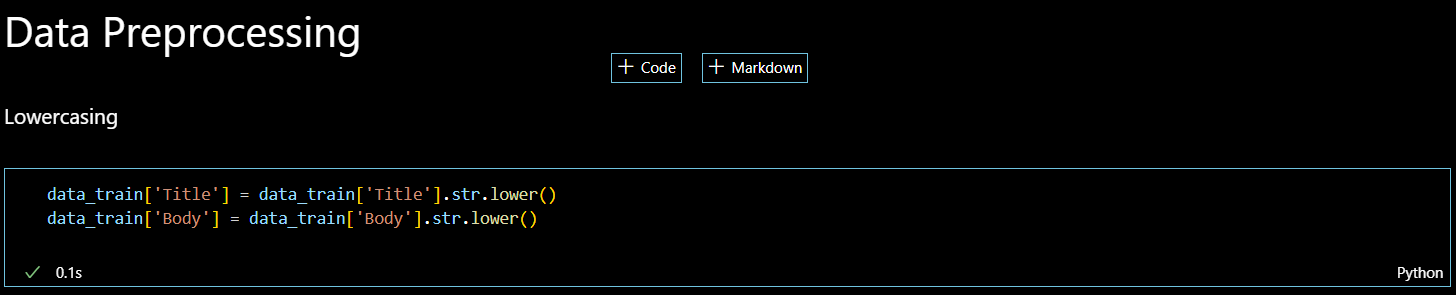
The above histogram visualizes the distribution of scores after removing outliers.

* **Word-cloud**Word-cloud was used in the first image to visualize the most used words in the “Title” column, and then re-used to visualize the words related to each OS class separately.
* **Sentiment polarity analysis**The above plot shows that most questions have a sentiment polarity of 0, which is neutral. That is logical as these questions are mostly technical.
* **Topic classification**

****

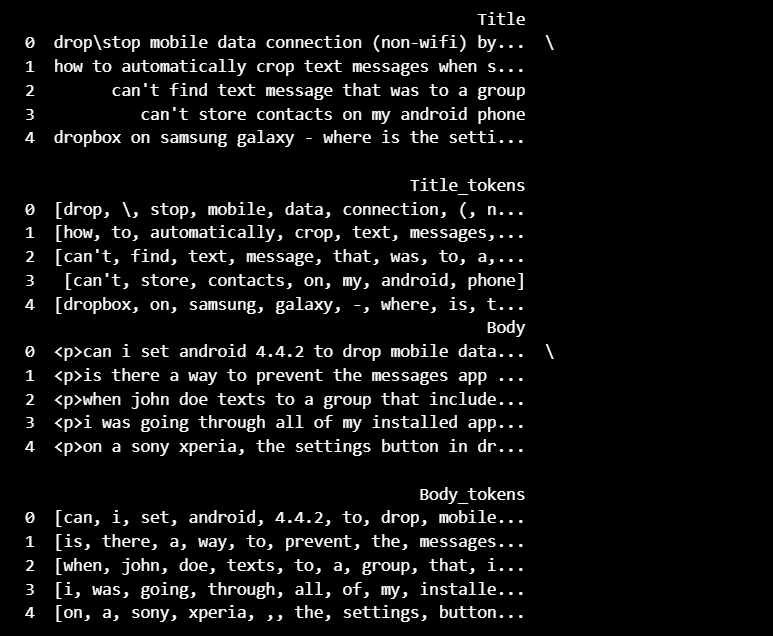
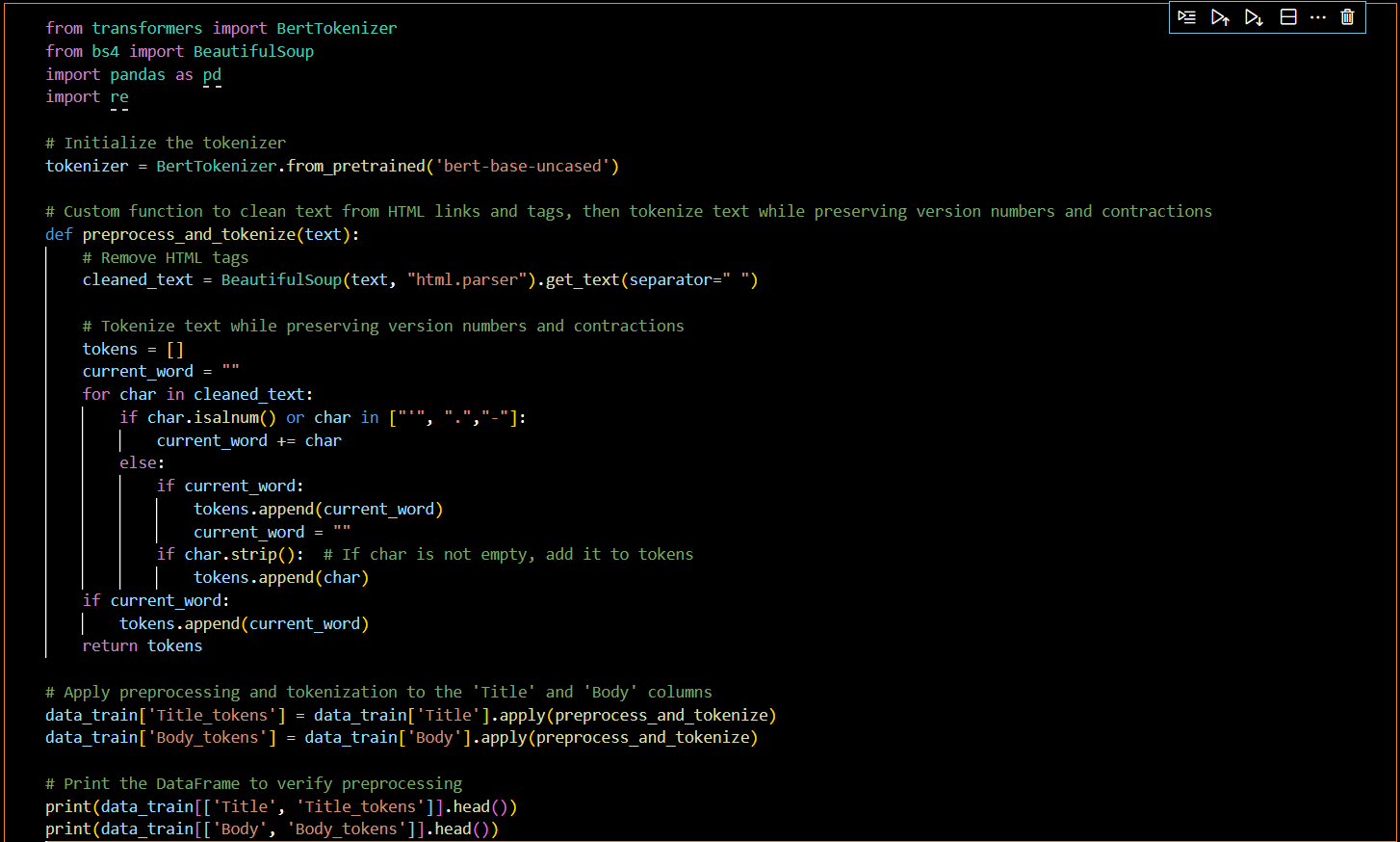
We utilized Latent Dirichlet Allocation (LDA) to uncover latent topics within text data. We first vectorized the text using CountVectorizer, then trained the LDA model to identify topics based on word distributions. Finally, we displayed the top words associated with each topic, providing insight into the main themes captured by the model.

****

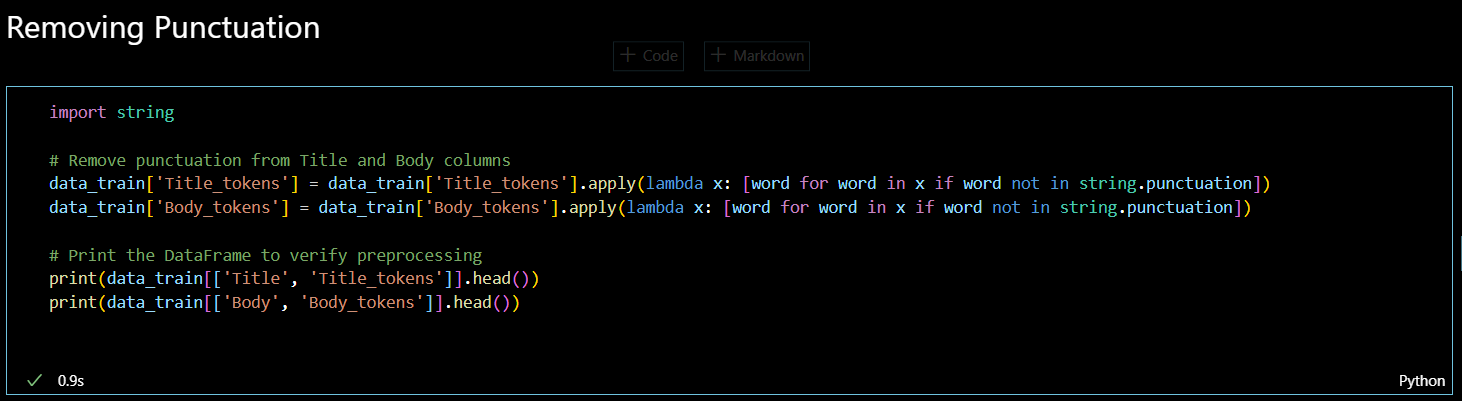
* **Data pre-processing**
* **Lowercasing**

The lowercasing is done to normalize the words for future tokenization and model training.

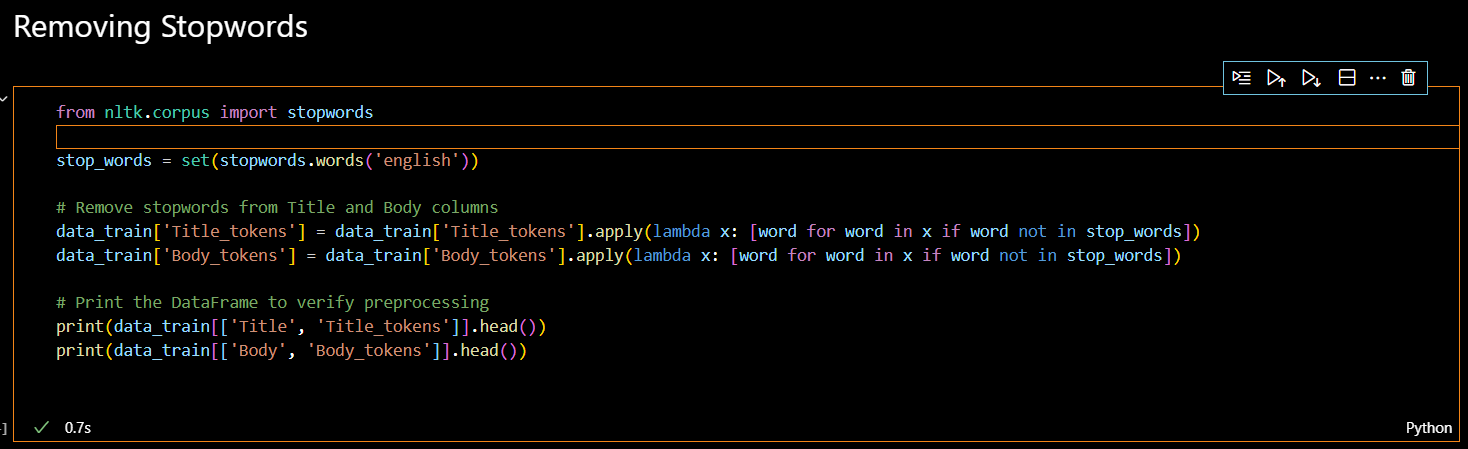
* **Tokenization**



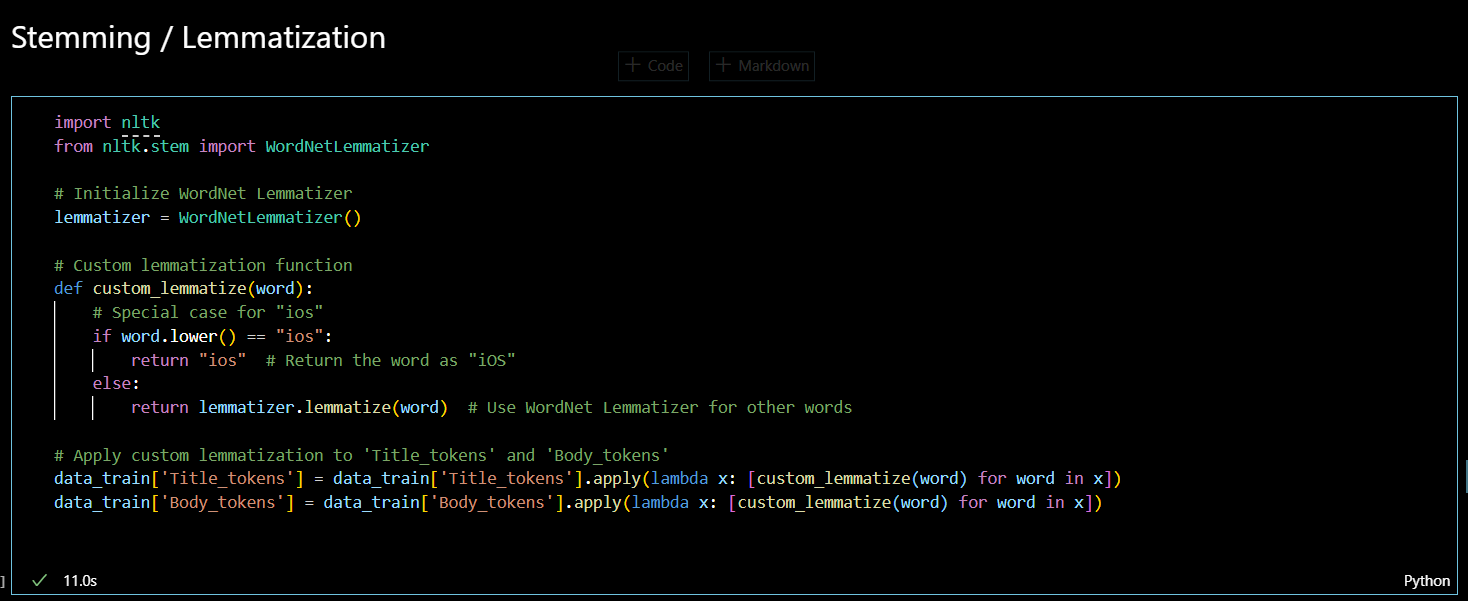
Tokenization is done to the dataset after lowercasing, it produces a list of all the words used in the dataset.

* **Removing punctuation**

Cleaning the tokens by removing punctuations.

* **Removing Stopwords**

Cleaning the tokens by removing stop words that will not add much to the context.

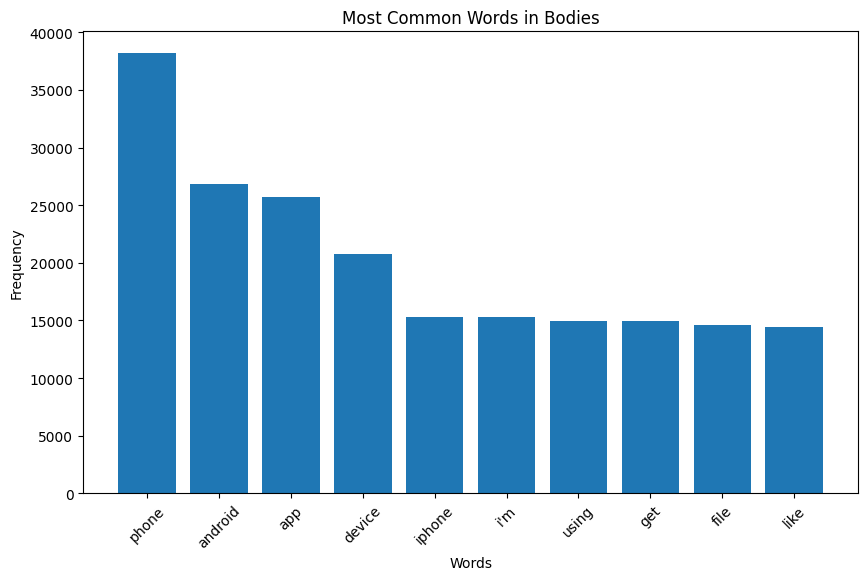
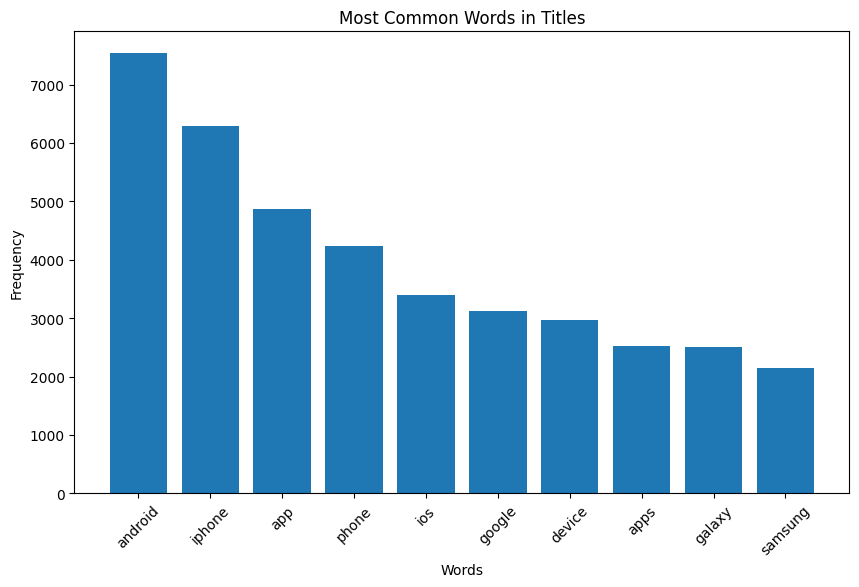
* **Stemming / Lemmatization**

Returning the tokens to their stem word.

* **Joining tokens** 

Reconstructing the sentences by joining the final tokens.

* **Token Analysis**



After applying all the cleaning methods for tokens, the above histogram visualizes the most frequent tokens in both the Body and Title columns. Which shows a successful tokenization as these words are mostly associated with technical questions related to android and ios.

**Critical Insights**

1. Imbalance in Android vs. iOS Content: The dataset heavily leans towards Android-related questions, reflecting the dominance of Android in the market compared to iOS. This skew raises concerns about potential differences in user interests or available information, highlighting the need for careful model training to maintain balance.

2. Effect of Extreme Values on Scores: Removing outliers significantly changes the distribution of scores, indicating that extreme values play a significant role in shaping overall trends. Understanding these outliers' nature and impact is crucial for accurate modeling and interpretation.

3. Digging Deeper into Word Frequency: While word clouds provide initial insights into common themes, a deeper analysis of word frequency could reveal underlying patterns and topics. Uncovering and understanding these subtleties can enhance analysis and model performance.

4. Exploring Neutral Sentiment: The predominance of neutral sentiment reflects the dataset's technical nature. However, examining sentiment fluctuations and their relationship with user engagement could uncover valuable insights for optimizing content and enhancing user satisfaction.

5. Ensuring Data Integrity through Pre-processing: Thorough pre-processing techniques, such as advanced tokenization to maintain sequence integrity and handling contractions, are essential for preserving data accuracy and minimizing information loss. Multiple problems were encountered as contraction tokenization and version tokenization (e.g. 4.4.2). Which needed special and advanced tokenization.

In summary, examining label biases, understanding outlier effects, conducting nuanced word frequency analysis, exploring sentiment trends, and employing meticulous pre-processing practices are crucial for extracting valuable insights and ensuring reliable model performance in real-world datasets.

**References:**

* Das, Arijit & Saha, Diganta. (2022). Question Answering System Using Deep Learning in the Low Resource Language Bengali. 10.1002/9781119857686.ch10.
* Tzu-Hsuan Lin, Yu-Hua Huang, Alan Putranto. Intelligent question and answer system for building information modeling and artificial intelligence of things based on the bidirectional encoder representations from transformers model, Automation in Construction, Volume 142, 2022, 104483, ISSN 0926-5805,<https://doi.org/10.1016/j.autcon.2022.104483>.
* Beta, Tatenda & Magadza, Tirivangani. (2023). Question-And-Answer System Using Natural Language Processing.
* LAVANYA, S. “End to End Question-Answering System Using NLP and SQuAD Dataset.” *Analytics Vidhya*, https://www.analyticsvidhya.com/blog/2021/11/end-to-end-question-answering-system-using-nlp-and-squad-dataset/. Accessed 9 March 2024.