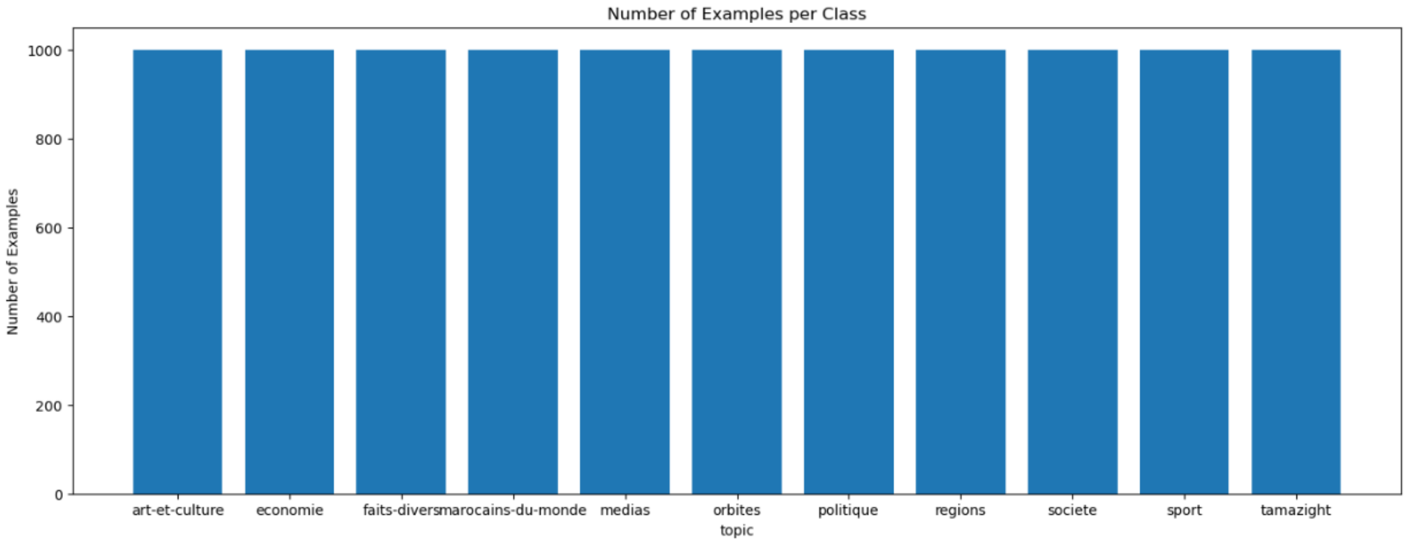
Task 2: Exploratory Data Analysis:

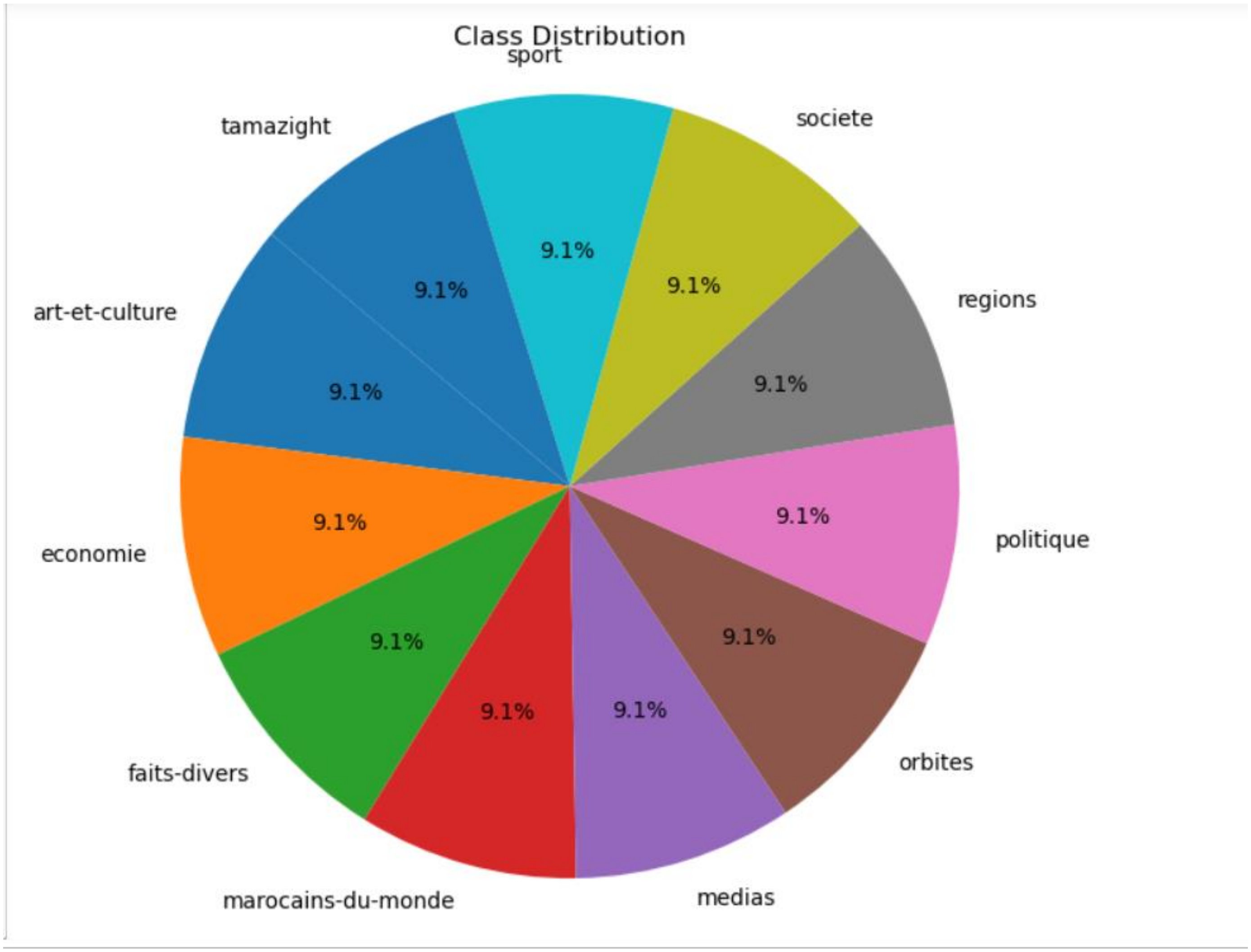
Number of Examples per class:

The code generates a bar plot that shows the number of examples (data points) belonging to each class or category in the 'topic' column of the DataFrame combined_df. The x-axis represents the class labels, the y-axis represents the number of examples, and each bar in the plot corresponds to a specific class with its respective count. This type of visualization helps to understand the class distribution and can be useful for tasks like exploring the class imbalance in a dataset.



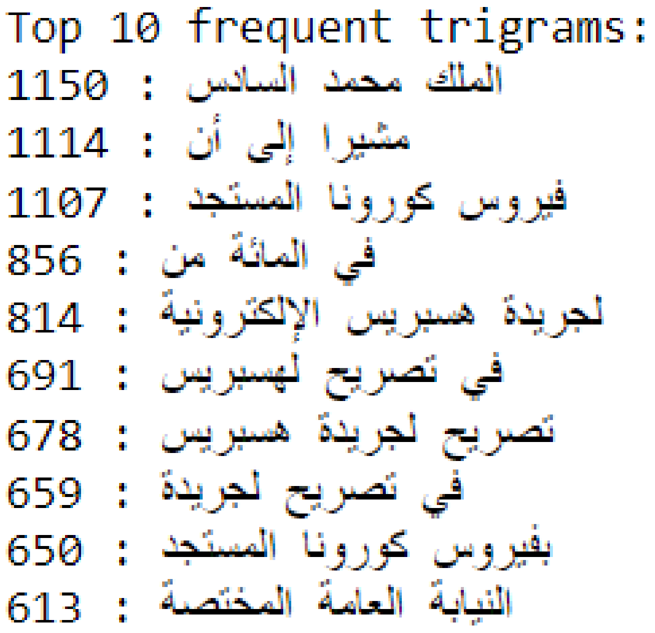
Class Distribution:

This part of the code creates a simple pie chart to visualize the distribution of classes in a dataset. This code generates a pie chart that represents the class distribution of the 'topic' column in the DataFrame combined_df. Each slice of the pie represents a unique class, and the size of each slice is proportional to the number of occurrences of that class in the dataset. The percentage labels on each slice show the proportion of each class relative to the whole dataset. Pie charts are useful for visualizing the relative frequencies or proportions of different categories in a dataset.



Top frequent n-grams generally and per class:

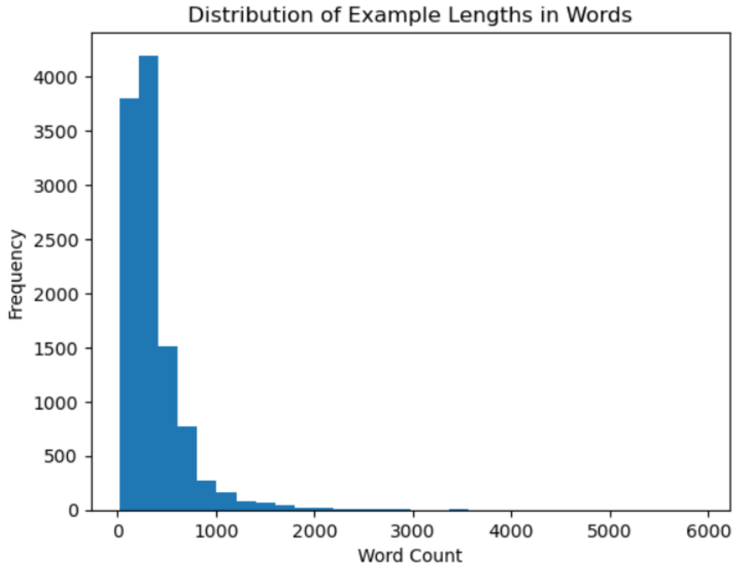
This figure performs text analysis to extract and find the most frequent trigrams (three consecutive words) in a given dataset. This code uses scikit-learn's CountVectorizer to analyze a text dataset and extract the most frequent trigrams. It then prints the top 10 most frequent trigrams along with their respective frequencies. The code is useful for identifying common sequences of three consecutive words that occur frequently in the provided text data.

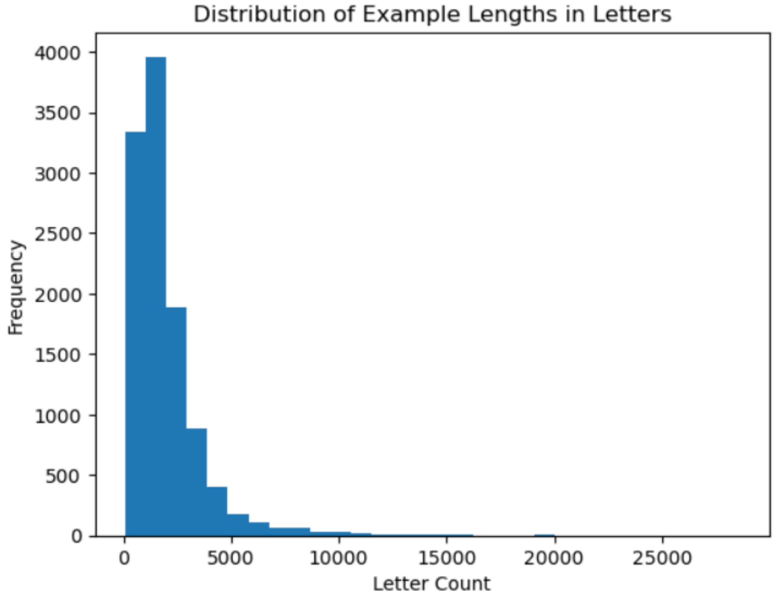


```
Top 10 frequent trigrams:
1150 : الملك محمد السادس
1114 : مشيرا إلى أن
1107 : فيروس كورونا المستجد
856 : في المائة من
814 : لجريدة هسبريس الإلكترونية
691 : في تصريح لهسبريس
678 : تصريح لجريدة هسبريس
659 : في تصريح لجريدة
650 : بفيروس كورونا المستجد
613 : النيابة العامة المختصة
```

lengths of examples in words and letters:

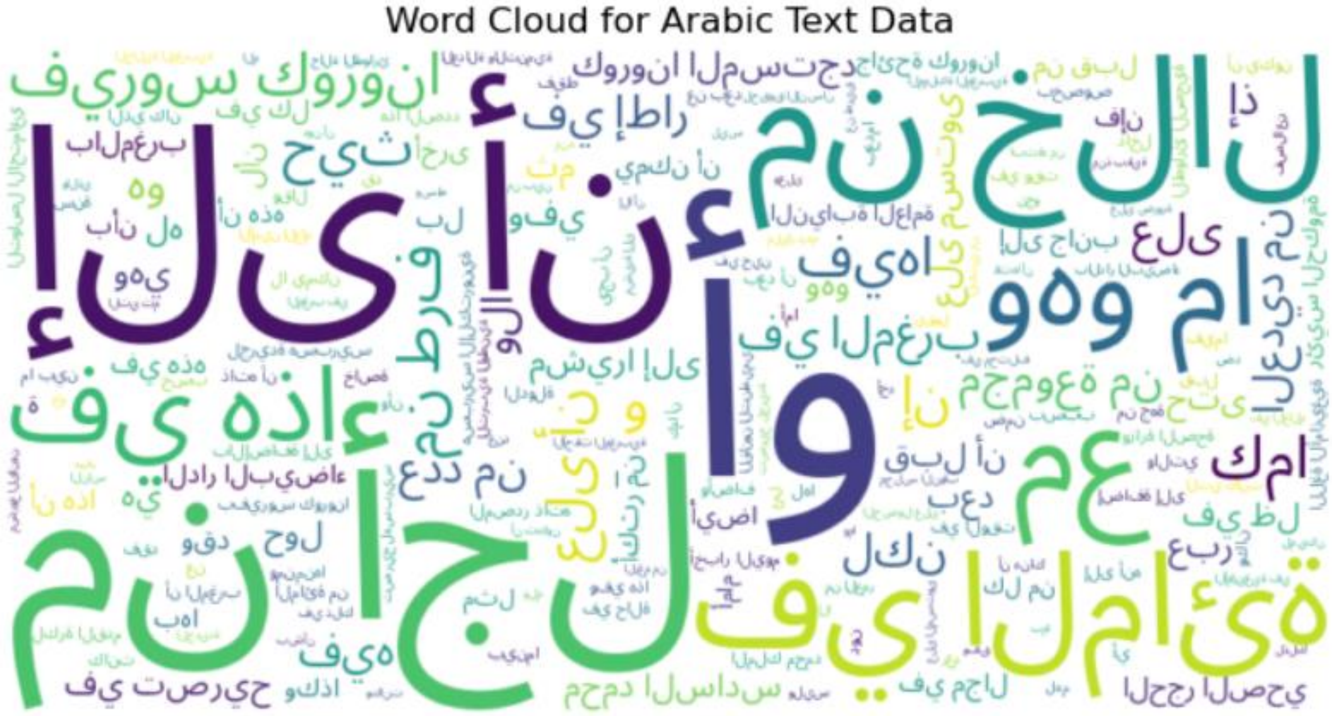
This part snippet performs analysis on the text data in the combined_df DataFrame. It calculates the word count and letter count for each text example and then creates two histograms to visualize the distribution of example lengths (in words and letters). This code calculates the word count and letter count for each text example in the DataFrame and visualizes the distribution of example lengths using histograms. The first histogram shows the frequency of example lengths in terms of word count, and the second histogram shows the frequency of example lengths in terms of letter count. These visualizations are useful for understanding the distribution of text lengths in the dataset.





Word cloud for Arabic text data

This code generates and displays a word cloud for Arabic text data. A word cloud is a visualization that shows the most frequently occurring words in a text by representing them as larger words.
The word cloud visually represents the most frequently occurring words in the text data, with larger words indicating higher frequency. The word cloud helps to get a quick overview of the important words and themes present in the Arabic text data.


Word Cloud for Arabic Text Data