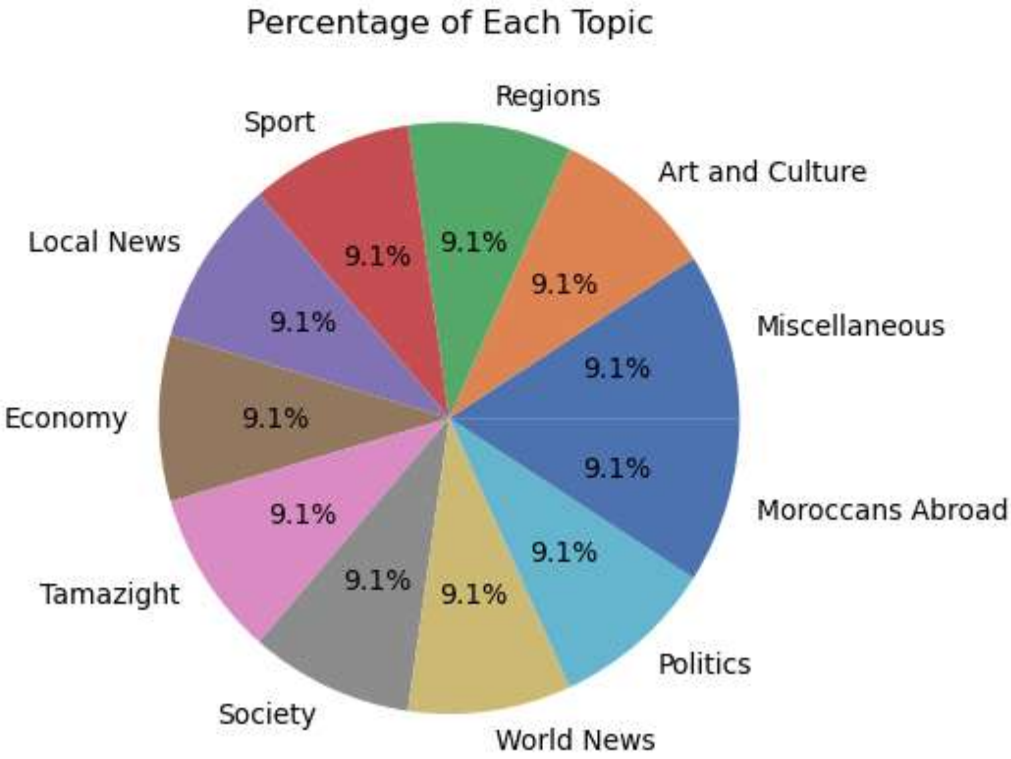
Hespress EDA

Insight #1

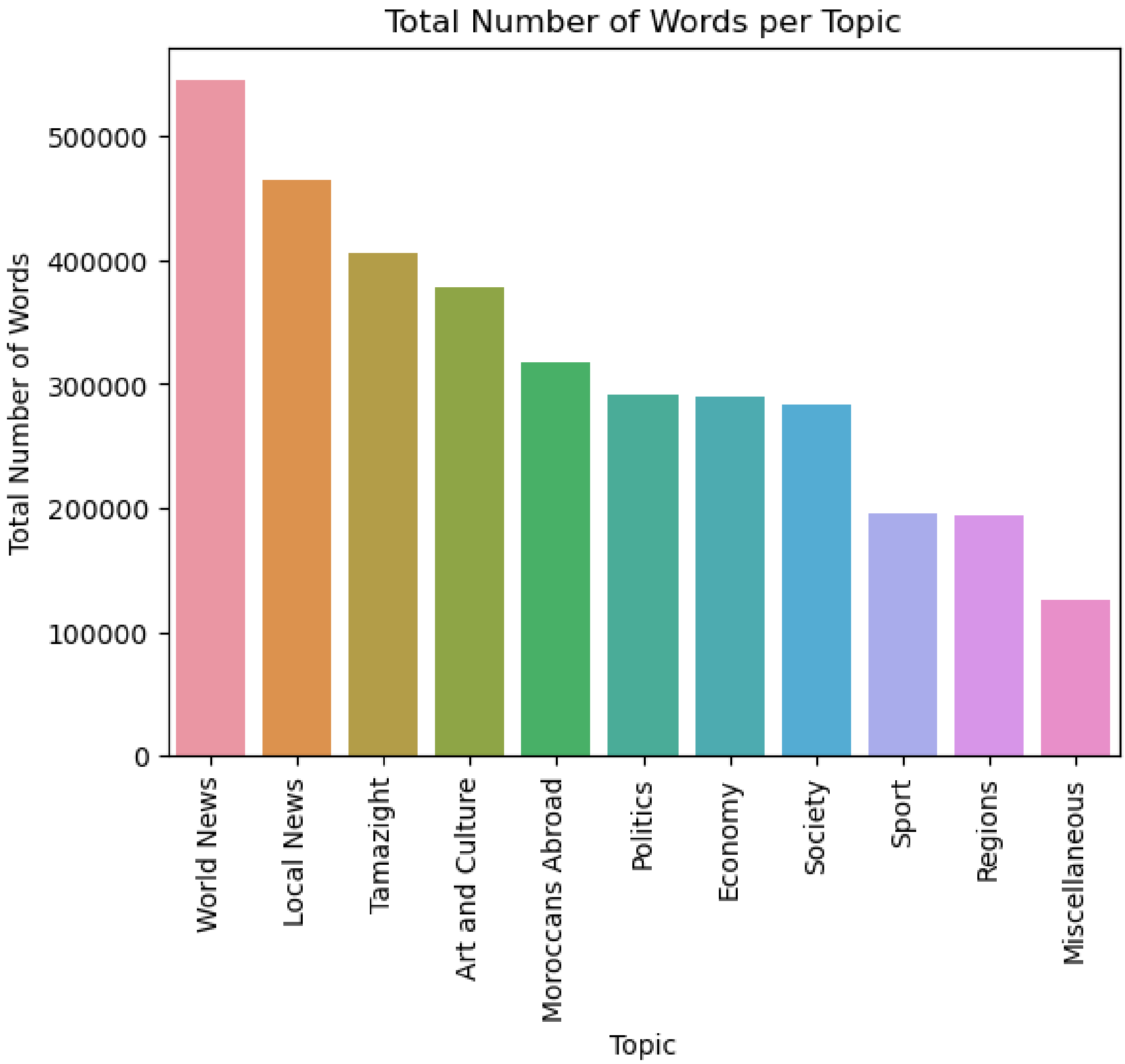


Using pie chart to plot the percentage of data for each topic/class and we the find that the data is balanced and all classes have equal number of records.

1000 records per topic for a total of 11,000 records.

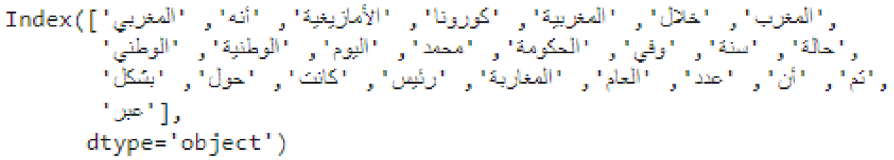
The names of each topic were mapped with English translation for better understanding.

Insight #2

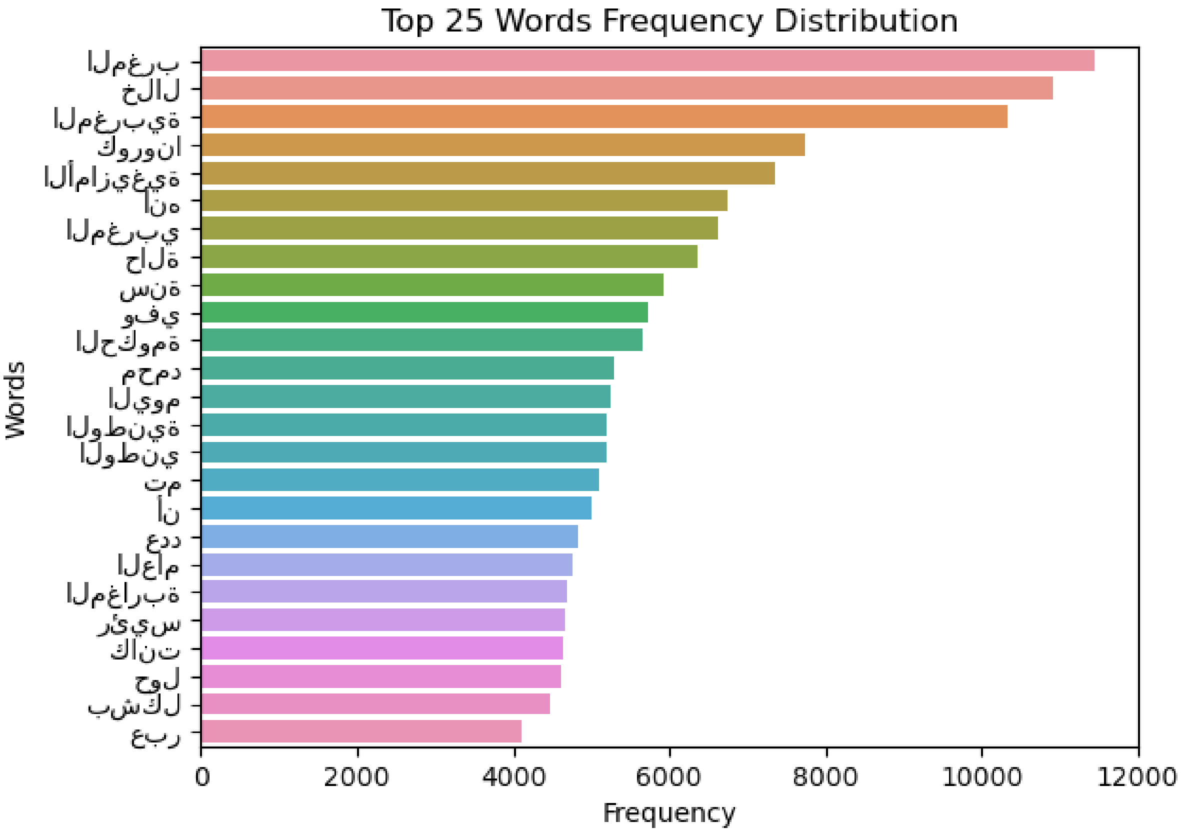


Using bar chart to plot the total number of words per topic we find that the topic “World News” (“orbites”) has the greatest total number of words (~550k+) and “Local News” (“medias”) is the second indicating that on average news stories have the greatest number of words per story, on the other side the topic “Miscellaneous” (“faits-divers”) has the least total number of words (~150k+).

Insight #3



```
Index(['المغرب', 'خلال', 'المغربية', 'كورونا', 'الأمازيغية', 'أنه', 'المغربي',
       'حالة', 'سنة', 'وفي', 'الحكومة', 'محمد', 'اليوم', 'الوطنية', 'الوطني',
       'تم', 'أن', 'عدد', 'العام', 'المغاربة', 'رئيس', 'كانت', 'حول', 'بشكل',
       'عبر'],
      dtype='object')
```



After removing stop words from stories and calculating the frequencies of each word, we use horizontal bar chart to plot the top 25 most frequent words and we find the top word is “المغرب” which isn’t surprising because it is a Moroccan dataset.

The word “كورونا” is also up there ranked the 4th on the list due to the recent pandemic.

We also find the name “محمد” the 13th word and the 1st name on the list, the name is one of the most common names in the world.