# Task2 EDA on **Hespress Data set**

The data describe the content of a Moroccan newspaper called **Hespress**. The Data consists of the title and the content for what is the site called story which is a kind of text speaking under that title and classify it according to its context to one of the categories either art, economics, news, Moroccans of the world, etc. also it mentions the one who wrote the story and when he wrote it.

At first sight, the Data got no missing values

A number of numbers and symbols

Description automatically generated

# The Number of examples of each class

A pie chart with different colored circles

Description automatically generatedA graph with blue and white stripes

Description automatically generatedFrom these charts, it’s obvious that there are class balances among different types of classes.

# The Word count and character length for story content in each class

## 

Here I will add chars that offer the distribution of the number of words for each story and the number of characters for each class using Box Plot and Kde Plot. As we got 11 class to classify among them so I will try to give some insights about some that look odd in some way.

1)Art and Culture:

A diagram of a graph and a diagram of a graph

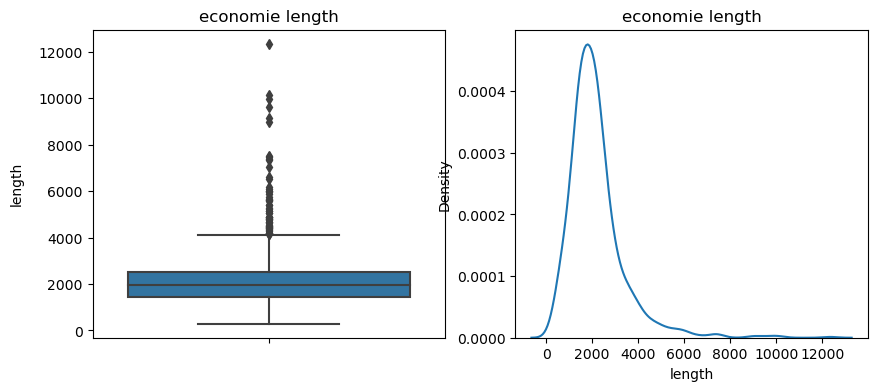
Description automatically generated

A diagram of a graph and a diagram of a graph

Description automatically generated

This shows that the average words per story in Art and Culture class are about 500 words with a number of char about 2500 characters.

2) Economic:

A diagram of a graph

Description automatically generated

Here we see that the average number of words in the Economic Class is about 300 words which is less than the first class with 2000 characters this might give that the vocabulary in this class might be long according that the number of characters is near to great difference in the number of words.3)

3) Miscellaneous facts:

\* Here I’d like to say that the language used might be French in the data frame except the data itself is in Arabic.

A diagram of a graph and a diagram of a graph

Description automatically generatedA diagram of a graph and a diagram of a graph

Description automatically generated

Here the same gives intuition about the number of words and their length but it also gives the probability for each number of the word in the story shown in the KDE plot.

4) Moroccans of the world:

A diagram of a graph and a diagram of a graph

Description automatically generatedA diagram and graph of a graph

Description automatically generated

Here it might be in the range of economic but also the number of characters is large according the number of characters in the first class.

5) Media:

A diagram and graph of data

Description automatically generatedA diagram of a graph and a diagram of a graph

Description automatically generated

The range of words is getting higher in this class it reaches 600 words for each story on average, unlike other classes.

6) Orbits

A diagram of a graph and a diagram of a graph

Description automatically generatedA diagram of a graph and a diagram of a graph

Description automatically generated

7)Politics:

A diagram of a graph and a diagram of a graph

Description automatically generated

A diagram of a graph and a diagram of a graph

Description automatically generated

8) Regions:

A diagram of a graph and a diagram of a graph

Description automatically generated

A diagram of a graph and a diagram of a graph

Description automatically generated

9)Societe :

A diagram of a graph

Description automatically generated

A graph and diagram of a word

Description automatically generated

10) Sport:

A graph and diagram of a word

Description automatically generated:

A diagram of a graph

Description automatically generated

A diagram of a graph and a diagram of a graph

Description automatically generated11) Tamazight:

A graph and a diagram

Description automatically generated

# Most Frequent N-grams

First, let’s define Ngram: the continuous sequence of words that have.

## Bigram

A group of rows of colorful bars

Description automatically generated

## Trigram:

A group of colorful bars

Description automatically generated

Final note the data has so many values that are wrong in the Date columns there are a lot of different representations of the month name for example he wrote “غشت” which really refers to "اغسطس" there may be more errors but according to the shortage of time I will stop here if it is required to do more it is very welcomed to mail with extended time to do more.