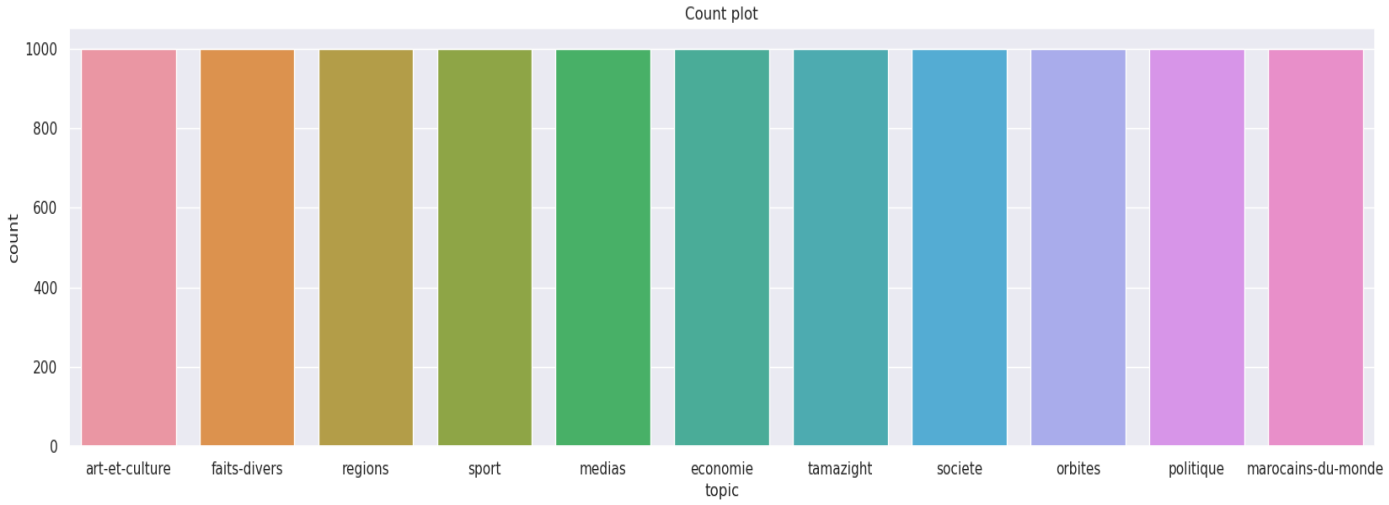
Task 2 – EDA
Hespress dataset

About Dataset

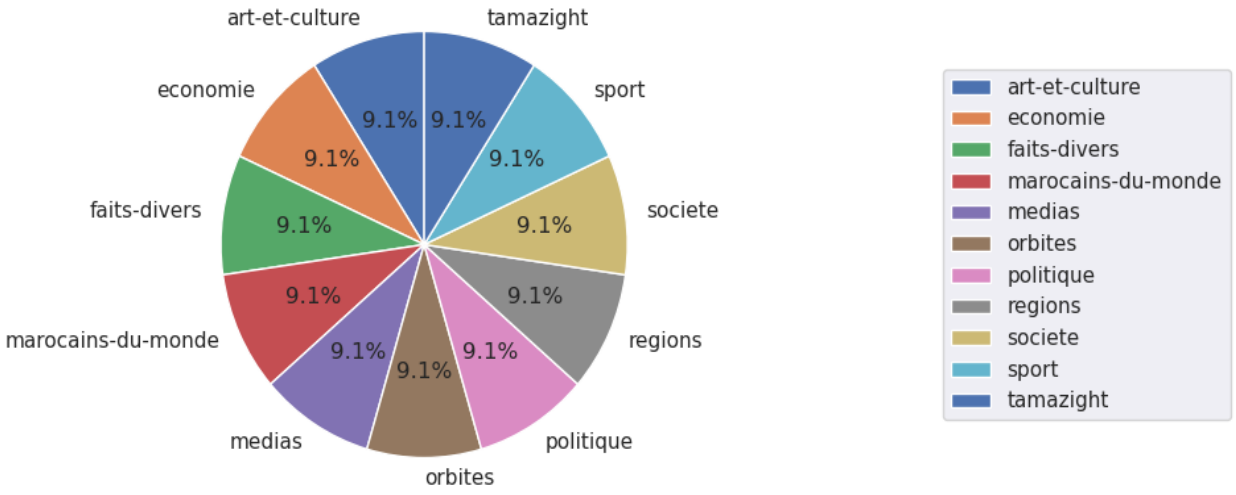
Scraped over 10 000 stories from Hespress with details such as the author, the publishing date, the topic. Also scraped all the comments associated with them which resulted in over 300K comments, this is very interesting for sentimental analysis, understanding the general opinion, and maybe even predicting election results, since with each comment a score by the readers is associated.

Number of examples per class

```
sns.set(rc={'figure.figsize': [20, 5]}, font_scale=1)
sns.countplot(x=df["topic"])
plt.title("Count plot")
plt.show()
```




```
plt.pie(df.topic.value_counts(),autopct='%1.1f%%',startangle=90, labels = np.unique(df.topic))
plt.legend(title="",loc="center left", bbox_to_anchor =(1.5, 0, 1, 1));
```



Top frequent n-grams generally

```
trigrams = ngrams(filtered_words, 3)
trigram_freq = Counter(trigrams)
top_trigrams = trigram_freq.most_common(10)
print("Top frequent trigrams:", top_trigrams)
```

Top frequent trigrams: [(('مؤكدة', 'جديدة', 'إصابة'), 71), (('ساعة', '24', 'خلال'), 67), (('جديدة', 'مؤكدة', 'بكورونا'), 62), (('مؤكدة', 'بكورونا', 'خلال'), 61), (('بكورونا', 'خلال', '24'), 61), (('محمد', 'الملك', 'السادس'), 53), (('التفاصيل', 'الحالة', 'الوبائية'), 46), (('الحالة', 'الوبائية', 'جهة'), 31), (('جهة', 'درعة', 'كانفيلاك'), 30), (('الوبائية', 'جهة', 'درعة'), 29)]

Top frequent n-grams per class

```
class_ngrams = {}
n = 3
for label in set([label for label in data.topic]):
    class_sentences = " ".join([sent.split('/')[0] for sent in list(data.title + '/' + data.topic) if sent.split('/')[1] == label])
    class_words = nltk.word_tokenize(class_sentences)
    class_trigrams = ngrams(class_words, n)
    class_trigram_freq = Counter(class_trigrams)
    class_top_trigrams = class_trigram_freq.most_common(3)
    class_ngrams[label] = class_top_trigrams

print("Top frequent n-grams per class:", class_ngrams)
```

Top frequent n-grams per class: {'economie': [(('البيضاء', 'الدار', 'البورصة'), 7), (('زمن', 'في', 'كورونا'), 6), (('كورونا', 'أزمة', 'تداعيات'), 5)], 'marocains-du-monde': [(('في', 'عالقون', 'مغاربة'), 8), (('قافلة', 'التبناك', 'الوحيد'), 8), (('في', 'العالقين', 'المغاربة'), 7)], 'faits-divers': [(('الحُور', 'على', 'جثة'), 9), (('سير', 'حادثة', 'في'), 9), (('هنك', 'عرض', 'قاصر'), 6)], 'tamazight': [(('وا', 'الأمازيغية'), 29), (('29 ,(', ')', 'الفرنسي'), 29), (('والاستمرار', 'الأمازيغية', 'السنة'), 23)], 'societe': [(('مؤكدة', 'جديدة', 'إصابة'), 71), (('ساعة', '24', 'خلال'), 64), (('جديدة', 'مؤكدة', 'بكورونا'), 62)], 'medias': [(('المغرب', 'الصحافة', 'رصيف'), 28), (('الصحافة', 'رصيف', 'كورونا'), 21), (('الصحافة', 'رصيف', 'الريسوني'), 12)], 'regions': [(('الوبائية', 'الحالة', 'التفاصيل'), 46), (('الحالة', 'الوبائية', 'في'), 31), (('الوبائية', 'في', 'جهة'), 31)], 'sport': [(('تجربة', 'جديدة', 'في'), 10), (('جديدة', 'في', 'يقترب'), 7)], 'art-et-culture': [(('كورونا', 'زمن', 'في'), 8), (('نم', 'في', 'ة'), 6), (('الشعر', 'في', 'بيت'), 4)], 'politique': [(('البحرية', 'الحدود', 'ترسيم'), 14), (('السادس', 'محمد', 'الملك'), 13), (('المشروع', 'غير', 'الإثراء'), 7)], 'orbites': [(('السادس', 'محمد', 'الملك'), 26), (('كورونا', 'بعد', 'ما'), 8), (('السادس', 'محمد', 'يشيد'), 6)]}

Lengths of examples in words and letters

```
word_lengths = [len(example.split()) for example in data.title]
letter_lengths = [len(example) for example in data.title]

# Print the results
for i, example in enumerate(data.title):
    print(f"Example {i+1}: '{example}'")
    print(f"Number of words: {word_lengths[i]}")
    print(f"Number of letters: {letter_lengths[i]}")
```

```
Example 1: 'بيت الشعر يسائل وزير الثقافة عن كواليس سوداء'
Number of words: 8
Number of letters: 44
Example 2: 'مهرجان سينما المؤلف يستحضر روح تريا جبران'
Number of words: 7
Number of letters: 42
Example 3: 'فيلم بدون عنف لهشام العسري كعب الحذاء وواقع مؤلم للنساء'
Number of words: 10
Number of letters: 55
Example 4: 'تنين ووهان  مريم أيت أحمد توقّع أولى روايات الجائحة بالمغرب'
Number of words: 10
Number of letters: 60
```

Show the most common words

```
top = Counter([word for sent in filtered_words for word in sent.split()])
temp = pd.DataFrame(top.most_common(20))
temp = temp.iloc[1:,:]
temp.columns = ['Common_words','count']
temp.style.background_gradient(cmap='Purples')
```

	Common_words	count
1	المغرب	904
2	الصحافة	580
3	رصيف	520
4	الأمازيغية	419
5	مغاربة	371
6	جديدة	293
7	الحكومة	271
8	المغربي	241
9	المغاربة	239
10	المغربية	225
11	بالمغرب	221
12	الملك	221
13	إصابة	199
14	مغربي	185
15	مغربية	177
16	وزارة	148
17	العثماني	146
18	البيضاء	141
19	المملكة	140

```
fig = px.bar(temp, x="count", y="Common_words", title='Most Commmon Words', orientation='h',
             width=700, height=700,color='Common_words')
fig.show()
```