

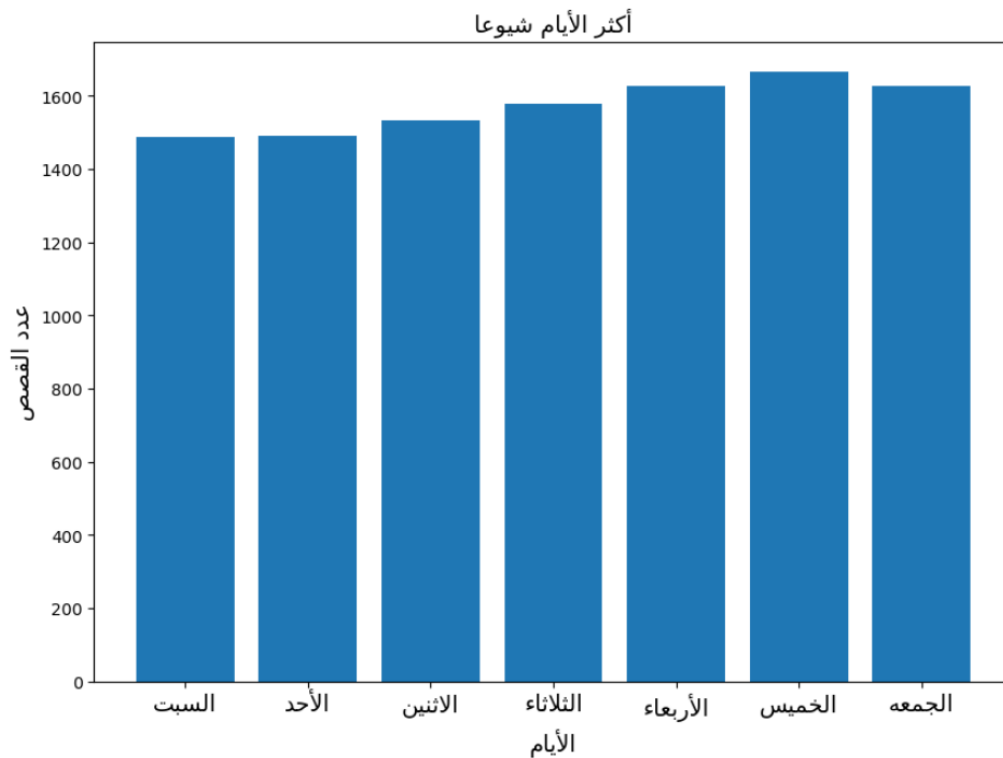
# Task 2 EDA

## 1. Number of stories per topic:

|                    |      |
|--------------------|------|
| art-et-culture     | 1000 |
| economie           | 1000 |
| faits-divers       | 1000 |
| marocains-du-monde | 1000 |
| medias             | 1000 |
| orbites            | 1000 |
| politique          | 1000 |
| regions            | 1000 |
| societe            | 1000 |
| sport              | 1000 |
| tamazight          | 1000 |

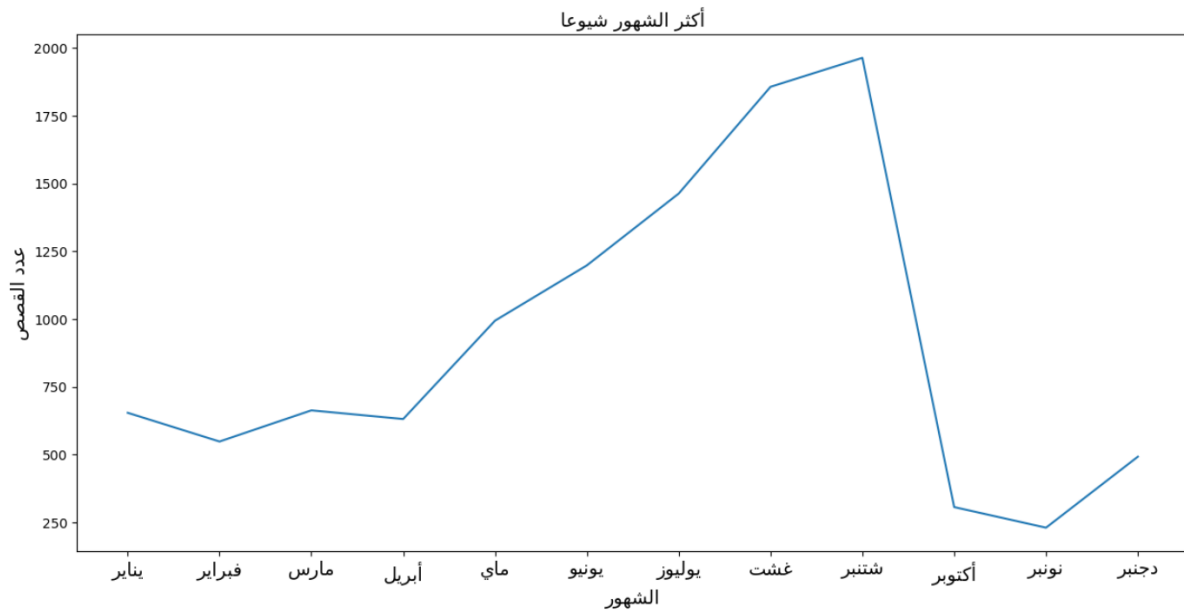
- All topics have the same number of stories, so no meaningful insights can be extracted from this question.

## 2. Most common days the stories have been published:



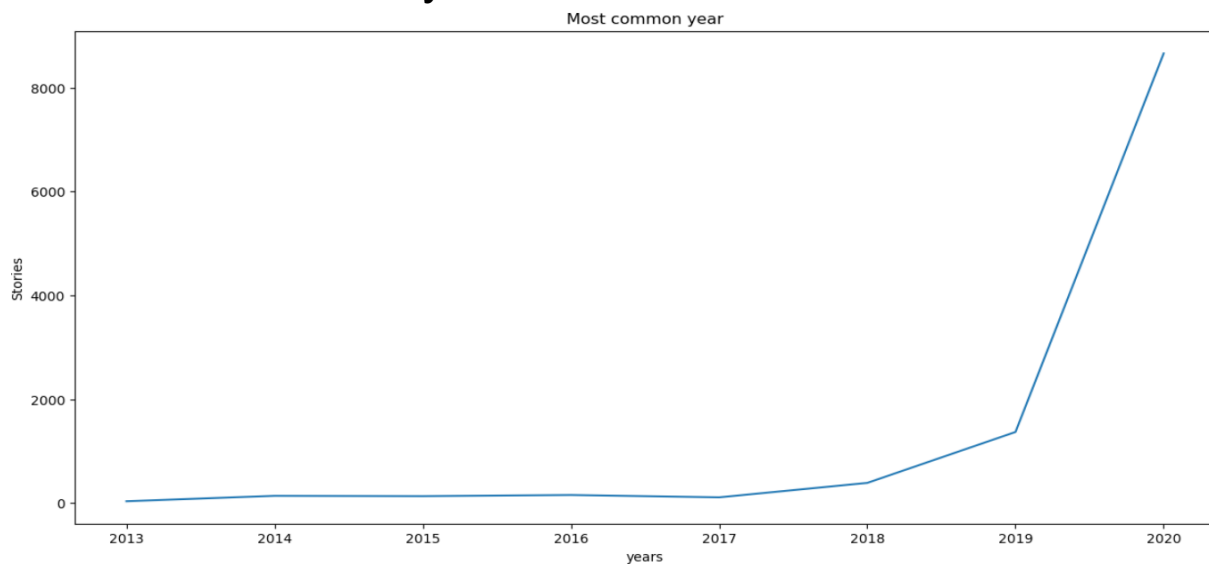
- No insights can be extracted from days since approximately they have the same number of stories.

### 3. Most common months



- We can find that at the start and the end of the year, there are few numbers of stories compared to the number in the middle of the year, it starts to increase from April to September

### 4. Most common year



- Over the years we've found that the number of stories started to increase in 2013 and it stayed the same from 2014 to the end of 2017 and it started to increase terribly from 2018.

## 5. Average number of words and letters per topic

|                    | N_Words | N_letters |
|--------------------|---------|-----------|
| topic              |         |           |
| art-et-culture     | 426.650 | 2675.401  |
| economie           | 345.076 | 2177.350  |
| faits-divers       | 149.244 | 928.050   |
| marocains-du-monde | 374.791 | 2373.735  |
| medias             | 545.222 | 3413.936  |
| orbites            | 643.044 | 4026.461  |
| politique          | 336.223 | 2150.037  |
| regions            | 223.271 | 1422.024  |
| societe            | 332.185 | 2100.177  |
| sport              | 227.919 | 1398.471  |
| tamazight          | 471.819 | 3000.346  |

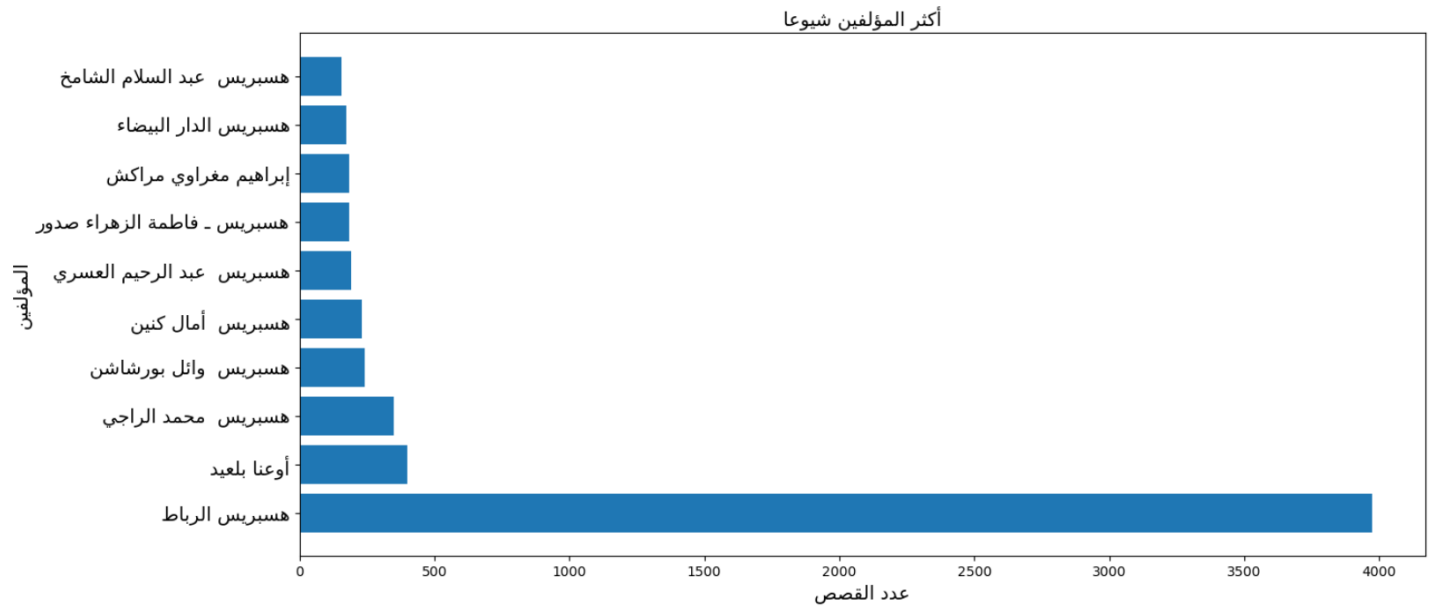
- Orbites topic has the most average number of words (643) and letters (4026)

## 6. Top frequent 2-grams generally and per class

- Generally:
  - The most repeated title 2-gram is: ('رصف', 'صحف') (count: 520)
  - The most repeated story 2-gram is: ('يرس', 'كورو') (count: 3873)
- Per regions:
  - The most repeated title 2-gram is: ('لالي', 'لاني') (count: 61)
  - The most repeated story 2-gram is: ('يرس', 'كورو') (count: 767)
- Per medias:
  - The most repeated title 2-gram is: ('رصف', 'صحف') (count: 520)
  - The most repeated story 2-gram is: ('الخير', 'اليوم') (count: 757)

- Per art-et-culture:
  - The most repeated title 2-gram is: ('وزر', 'ثقّف') (count: 20)
  - The most repeated story 2-gram is: ('وزر', 'ثقّف') (count: 322)
- Per orbites:
  - The most repeated title 2-gram is: ('جئح', 'كورو') (count: 36)
  - The most repeated story 2-gram is: ('يرس', 'كورو') (count: 456)
- Per politique:
  - The most repeated title 2-gram is: ('جلس', 'نوب') (count: 29)
  - The most repeated story 2-gram is: ('جلس', 'نوب') (count: 646)
- Per economie:
  - The most repeated title 2-gram is: ('ازم', 'كورو') (count: 38)
  - The most repeated story 2-gram is: ('يرس', 'كورو') (count: 534)
- Per societe:
  - The most repeated title 2-gram is: ('غرب', 'سجل') (count: 83)
  - The most repeated story 2-gram is: ('وزر', 'صحة') (count: 663)
- Per tamazight:
  - The most repeated title 2-gram is: ('سنة', 'امازيغية') (count: 65)
  - The most repeated story 2-gram is: ('لغة', 'امازيغية') (count: 1085)
- Per marocains-du-monde:
  - The most repeated title 2-gram is: ('غرب', 'علق') (count: 87)
  - The most repeated story 2-gram is: ('غرب', 'قيم') (count: 862)
- Per faits-divers:
  - The most repeated title 2-gram is: ('روح', 'خدر') (count: 31)
  - The most repeated story 2-gram is: ('انيب', 'عمة') (count: 738)
- Per sport:
  - The most repeated title 2-gram is: ('كرة', 'قدم') (count: 29)
  - The most repeated story 2-gram is: ('الكر', 'قدم') (count: 1006)

## 7. Most common Authors



- As we can see (هسبريس الرباط) is the most common author for stories, I think because Rabat is Morocco's capital or maybe the headquarters of this company is at Rabat.