هسبريس
HESPRESS

Exploratory Data Analysis Report for Hespress

Table of Contents

Report and Insights

The provided output contains information on the analysis of different datasets related to various topics. Each dataset represents a specific category or topic, and we have the following datasets:

1. `stories_politique.csv`: Contains stories related to politics.
2. `stories_medias.csv`: Contains stories related to the media.
3. `stories_regions.csv`: Contains stories related to regions.
4. `stories_sport.csv`: Contains stories related to sports.
5. `stories_marocains-du-monde.csv`: Contains stories related to Moroccans living abroad.
6. `stories_economie.csv`: Contains stories related to the economy.
7. `stories_orbites.csv`: Contains stories related to space.
8. `stories_faits-divers.csv`: Contains miscellaneous stories or incidents.
9. `stories_art-et-culture.csv`: Contains stories related to art and culture.
10. `stories_tamazight.csv`: Contains stories related to the Amazigh (Berber) language and culture.
11. `stories_societe.csv`: Contains stories related to society.

Let's analyze each dataset individually:

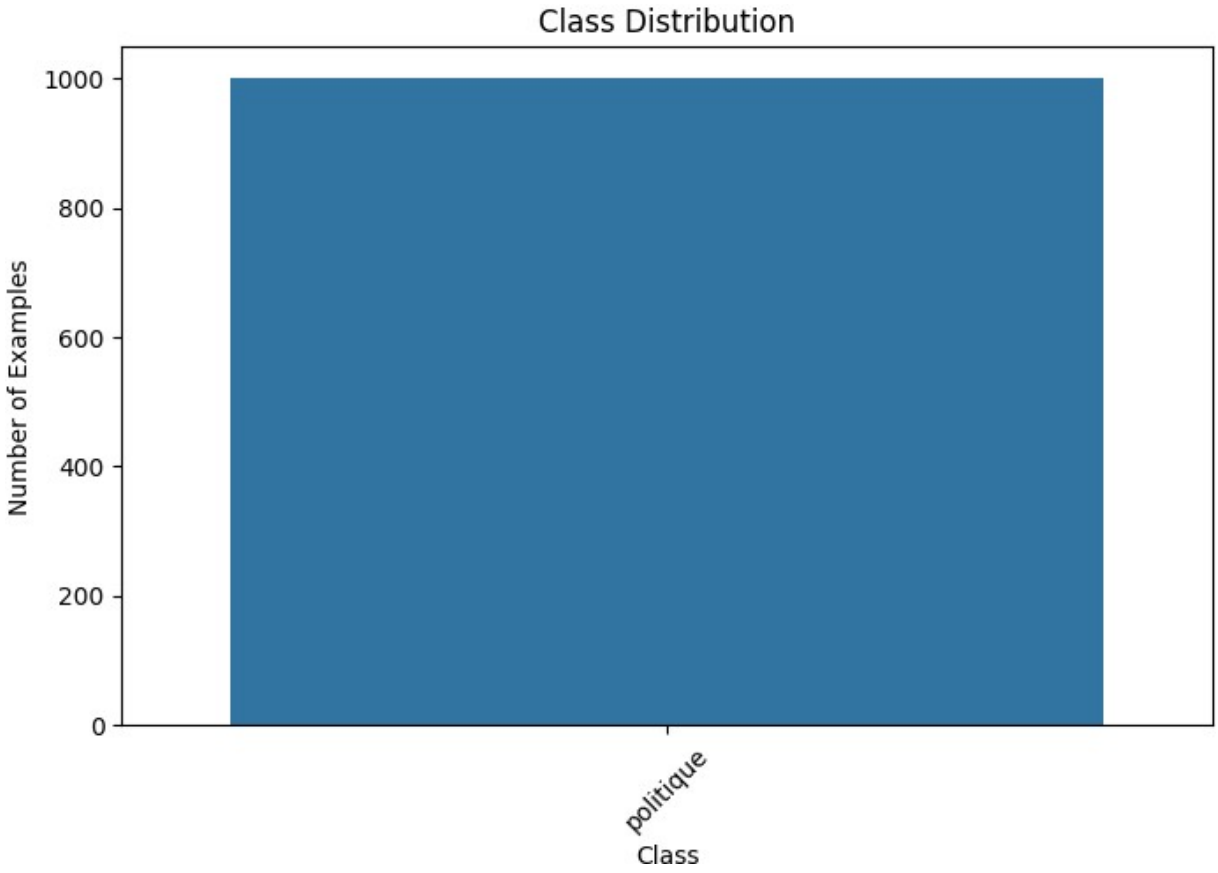
Dataset: stories_politique.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "politique" category.
- The top 10 frequent 2-grams (word pairs) in the stories are related to political terms, like "رئيس الحكومة" (Prime Minister), "مجلس النواب" (House of Representatives), and "الأمين العام" (General Secretary), indicating that the stories indeed cover political topics.

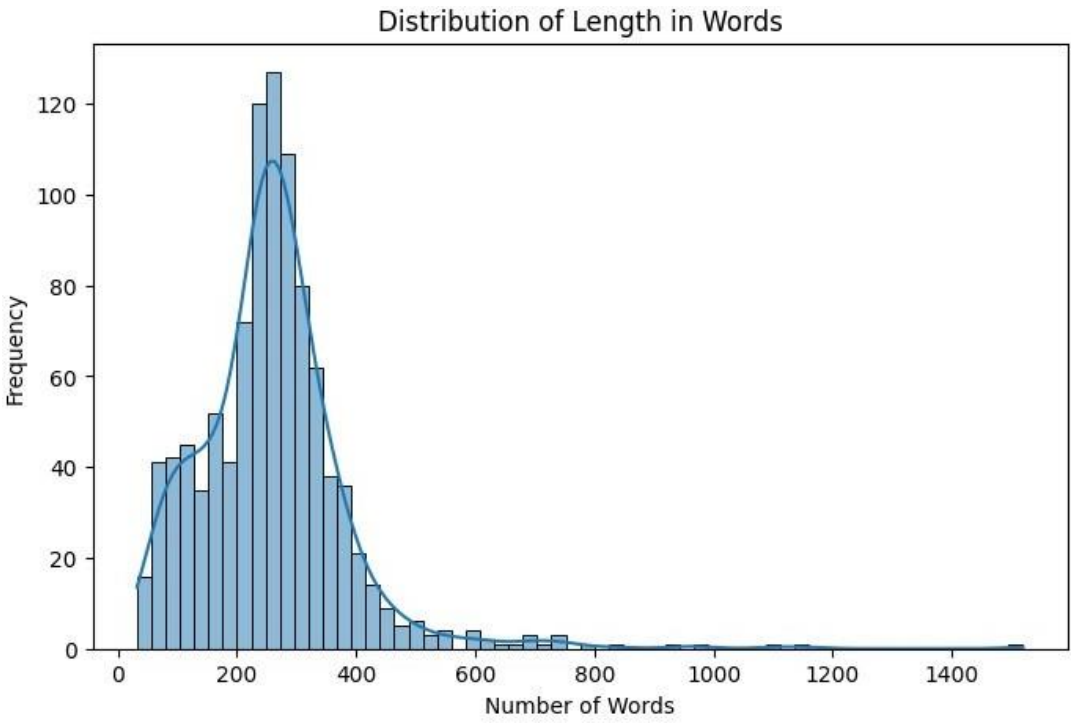
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "politique" category.
- Top 2-grams: Frequent word pairs include "العدالة والتنمية" "مجلس النواب," "رئيس الحكومة," and others.
- Length of Examples in Words: The average story contains around 258 words, with a minimum of 32 words and a maximum of 1520 words.
- Length of Examples in Letters: The average story contains around 1856 letters, with a minimum of 196 letters and a maximum of 10661 letters.

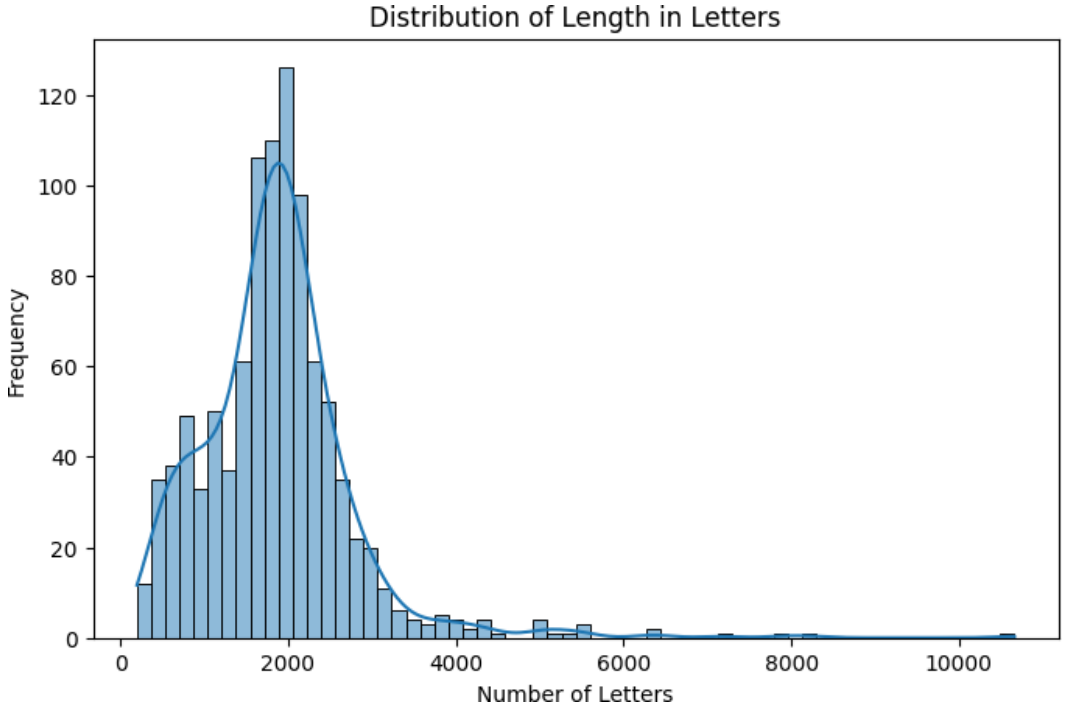
```
--- Individual File Analysis for stories_politique.csv ---

<class 'pandas.core.frame.DataFrame'>RangeIndex:
1000 entries, 0 to 999 Data columns (total 7
columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   author      1000 non-null   object
 4   text        1000 non-null   object
 5   publication 1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)memory
usage:  54.8+  KB None

Class Distribution:politique
1000

Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
441 :رئيس الحكومة
410 :مجلس النواب
402 :العدالة والتنمية
353 :الأمين العام
321 :محمد السادس
291 :الأصالة والمعاصرة
290 :الملك محمد
252 :مشروع القانون
217 :الدين العثماني
215 :سعد الدين

Length of Examples in Words:
count    1000.000000
mean      257.986000
std       128.889465
min        32.000000
25%       187.000000
50%       253.000000
75%       308.000000
max      1520.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      1855.676000
std        927.632054
min        196.000000
25%       1353.000000
50%       1842.500000
75%       2209.750000
max      10661.000000
Name: num_letters, dtype: float64
```



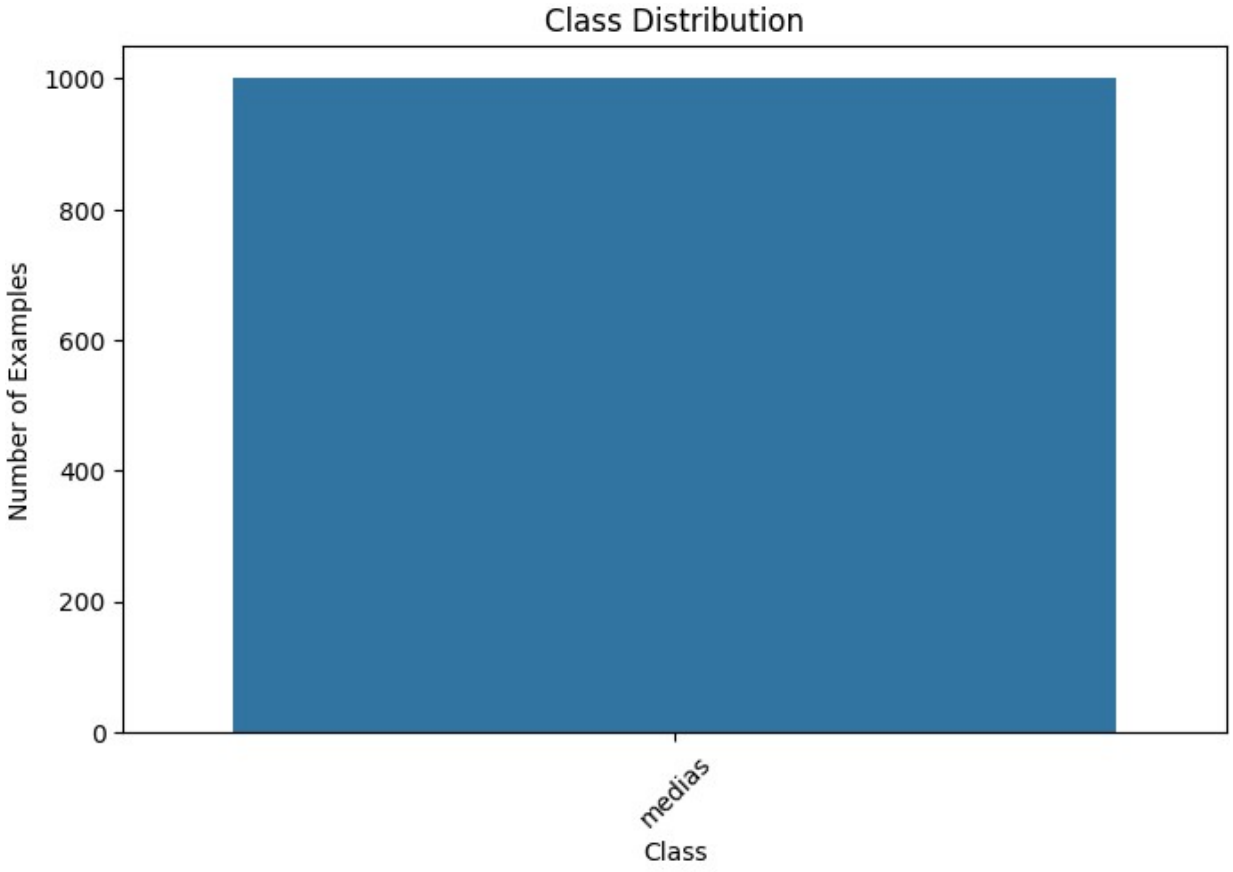
Dataset: stories_medias.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "medias" category.
- The top 10 frequent 2-grams (word pairs) in the stories are related to media topics, like "أخبار اليوم" (News of the day) and "الأحداث المغربية" (Moroccan events).

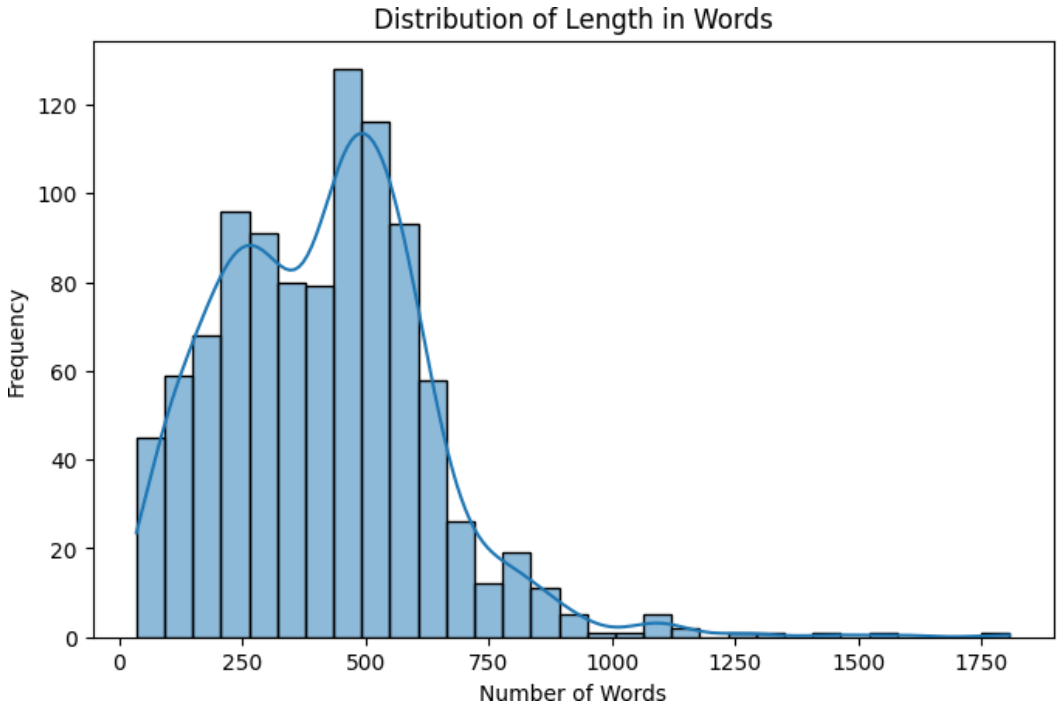
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "medias" category.
- Top 2-grams: Frequent word pairs include "فيروس كورونا" "الأحداث المغربية," "أخبار اليوم," and others.
- Length of Examples in Words: The average story contains around 415 words, with a minimum of 35 words and a maximum of 1807 words.
- Length of Examples in Letters: The average story contains around 2920 letters, with a minimum of 232 letters and a maximum of 12193 letters.

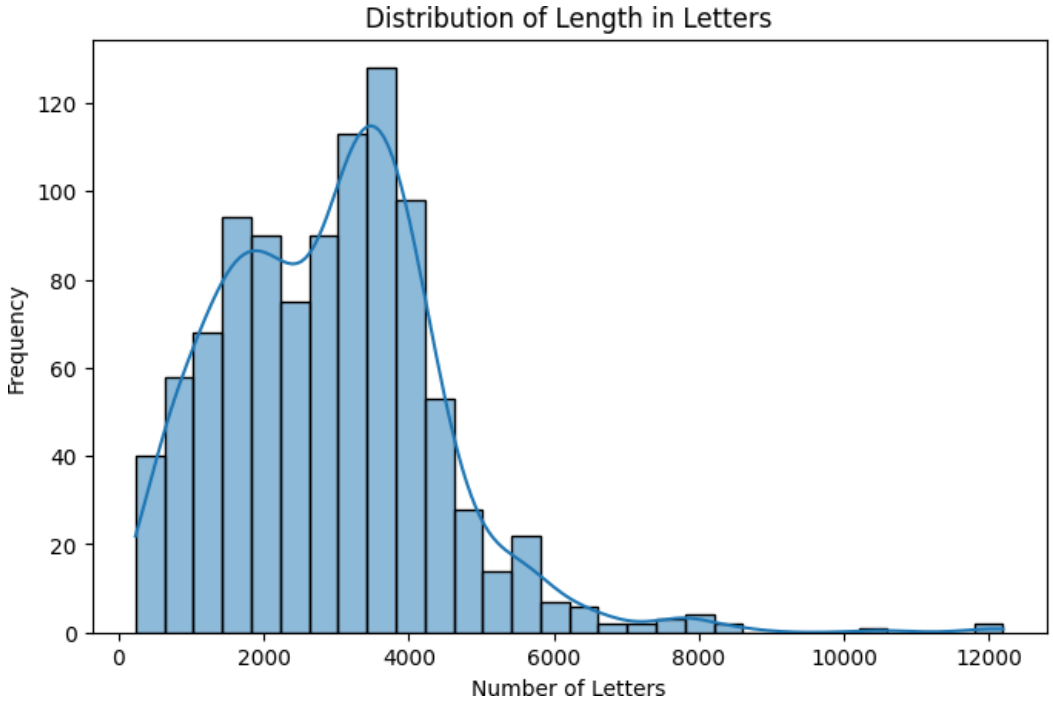
```
--- Individual File Analysis for stories_medias.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
medias    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
841 :أخبار اليوم
702 :الأحداث المغربية
330 :فيروس كورونا
328 :النيابة العامة
275 :محمد السادس
252 :وأضافت الجريدة
248 :التواصل الاجتماعي
243 :خبر آخر
226 :ووفق المنبر
225 :رصيف صحافة

Length of Examples in Words:
count    1000.000000
mean      414.966000
std       213.819339
min        35.000000
25%       250.750000
50%       426.500000
75%       545.000000
max      1807.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      2920.459000
std       1497.502209
min        232.000000
25%       1771.000000
50%       2982.500000
75%       3805.750000
max      12193.000000
Name: num_letters, dtype: float64
```



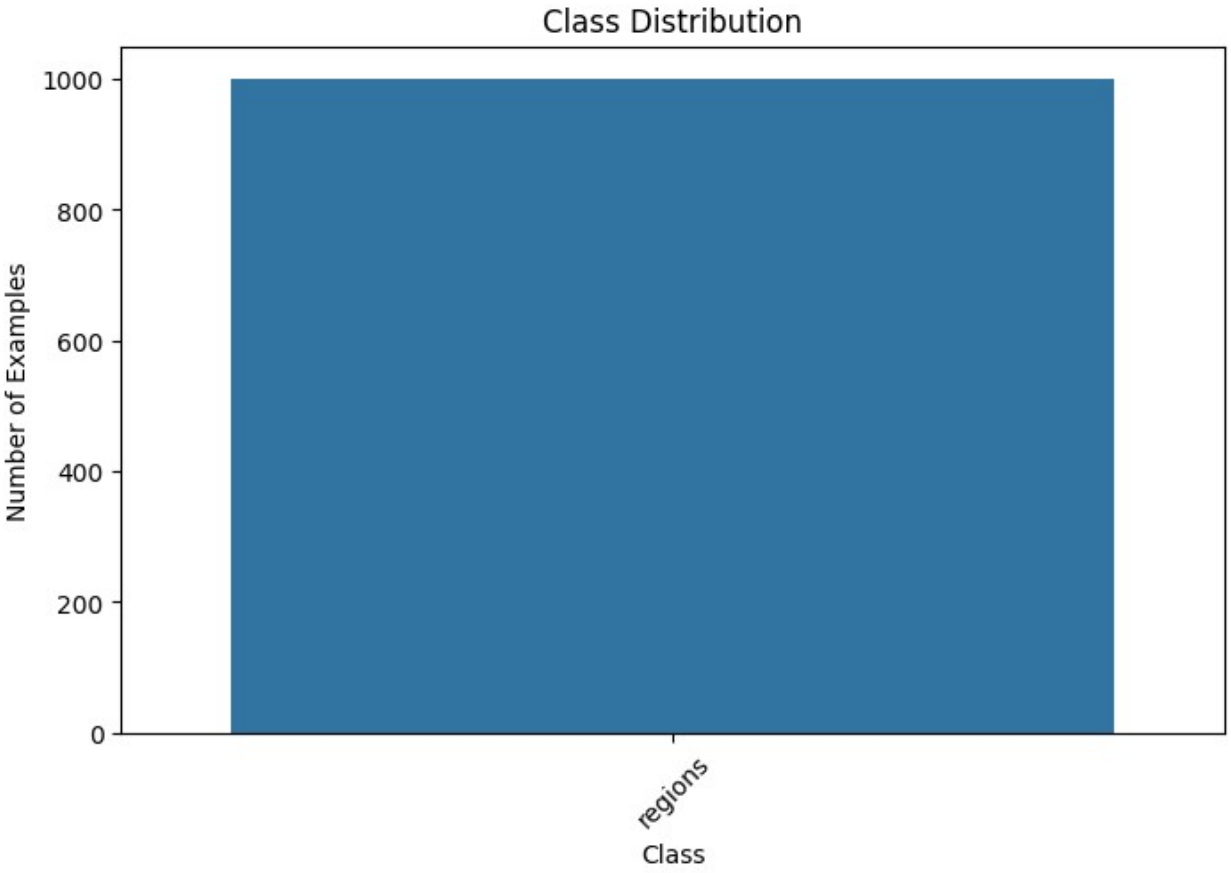
Dataset: stories_regions.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "regions" category.
- The top 10 frequent 2-grams (word pairs) in the stories are related to regional topics, like "كورونا المستجد" (New coronavirus) and "حالة بإقليم" (Case in the region).

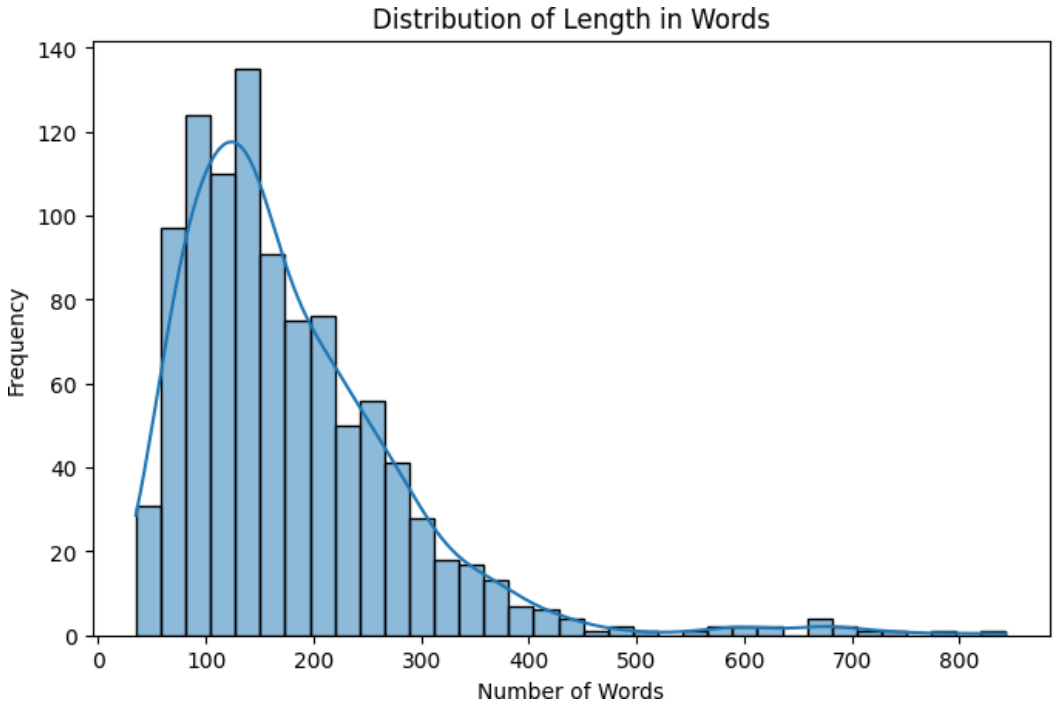
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "regions" category.
- Top 2-grams: Frequent word pairs include "19 كوفيد" "بفيروس كورونا," "كورونا المستجد," and others.
- Length of Examples in Words: The average story contains around 178 words, with a minimum of 35 words and a maximum of 843 words.
- Length of Examples in Letters: The average story contains around 1248 letters, with a minimum of 252 letters and a maximum of 5841 letters.

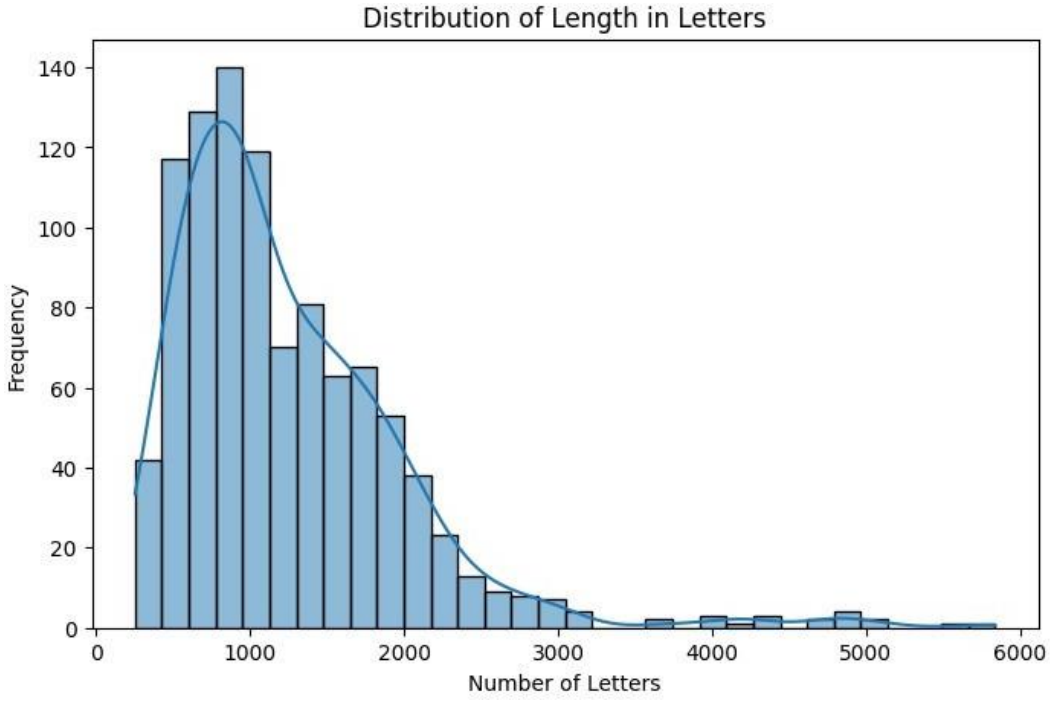
```
--- Individual File Analysis for stories_regions.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  1000 non-null   int64 
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
regions    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
515 كورونا المستجد
500 بفيروس كورونا
432 كوفيد 19
354 حالة بإقليم
351 الدار البيضاء
274 فيروس كورونا
263 بني ملال
227 المديرية الجهوية
220 الجهوية للصحة
207 إصابة جديدة

Length of Examples in Words:
count    1000.000000
mean      177.859000
std       106.605947
min        35.000000
25%       104.000000
50%       151.000000
75%       222.500000
max       843.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count    1000.00000
mean     1247.73300
std       752.44526
min       252.00000
25%       732.75000
50%      1054.00000
75%      1610.25000
max      5841.00000
Name: num_letters, dtype: float64
```



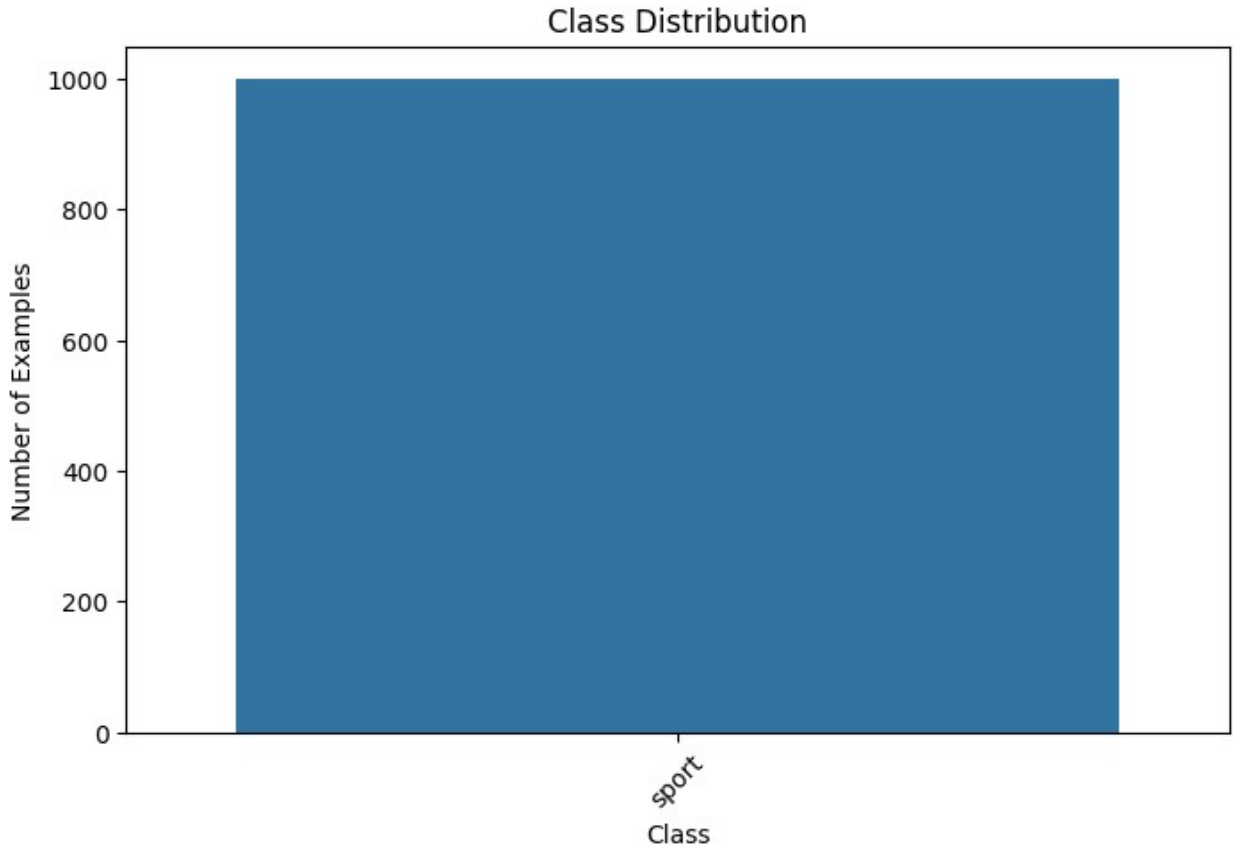
Dataset: stories_sport.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "sport" category.
- The top 10 frequent 2-grams (word pairs) in the stories are related to sports, such as "لكرة القدم" (Football), "الملكية المغربية" (Moroccan Royal Family), and "كورونا المستجد" (New coronavirus).

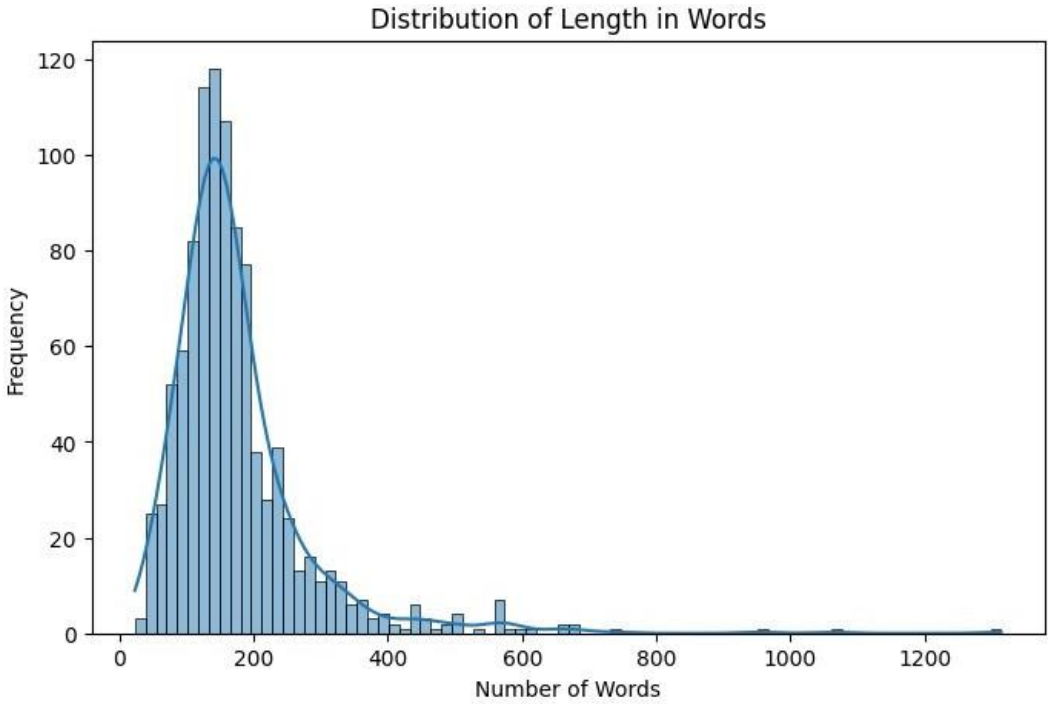
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "sport" category.
- Top 2-grams: Frequent word pairs include "الدولي المغربي" "فيروس كورونا," "لكرة القدم," and others.
- Length of Examples in Words: The average story contains around 176 words, with a minimum of 23 words and a maximum of 1314 words.
- Length of Examples in Letters: The average story contains around 1200 letters, with a minimum of 142 letters and a maximum of 8681 letters.

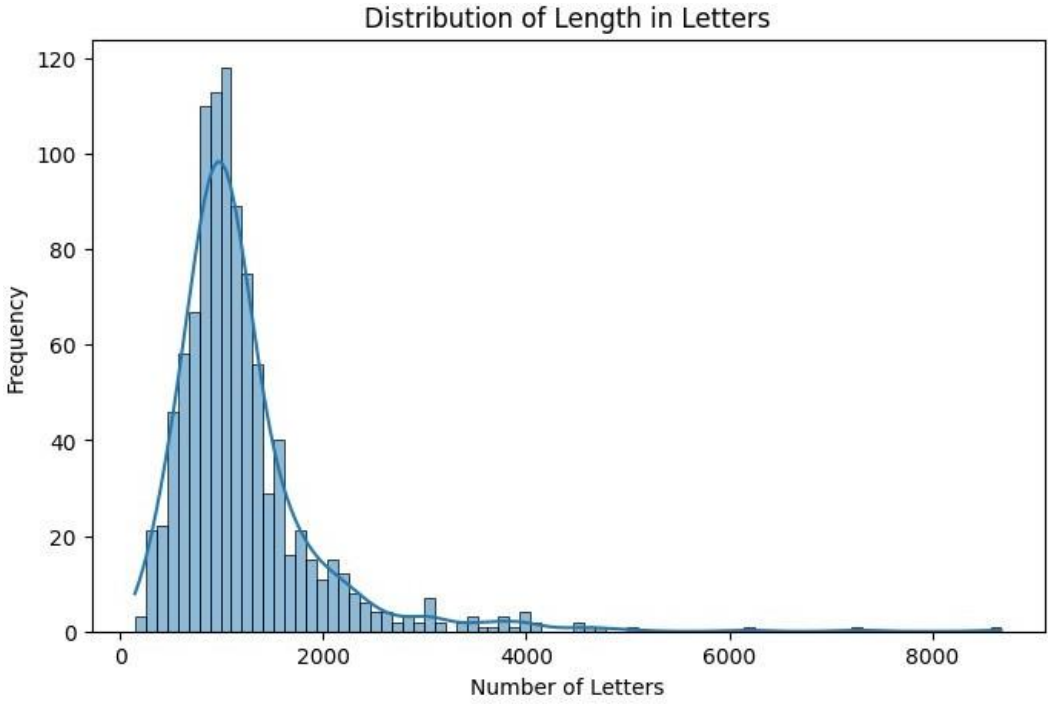
```
--- Individual File Analysis for stories_sport.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  1000 non-null   int64 
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
sport    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
1038 :لكرة القدم
372 :فيروس كورونا
317 :الدولي المغربي
284 :كرة القدم
250 :البالغ العمر
239 :الموسم الحالي
229 :كورونا المستجد
193 :الملكية المغربية
174 :المغربية لكرة
148 :الطاقم التقني

Length of Examples in Words:
count    1000.000000
mean      175.528000
std       108.658547
min        23.000000
25%       118.000000
50%       150.500000
75%       197.000000
max      1314.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count    1000.000000
mean     1199.855000
std       741.178671
min       142.000000
25%       809.000000
50%      1027.000000
75%      1342.500000
max      8681.000000
Name: num_letters, dtype: float64
```



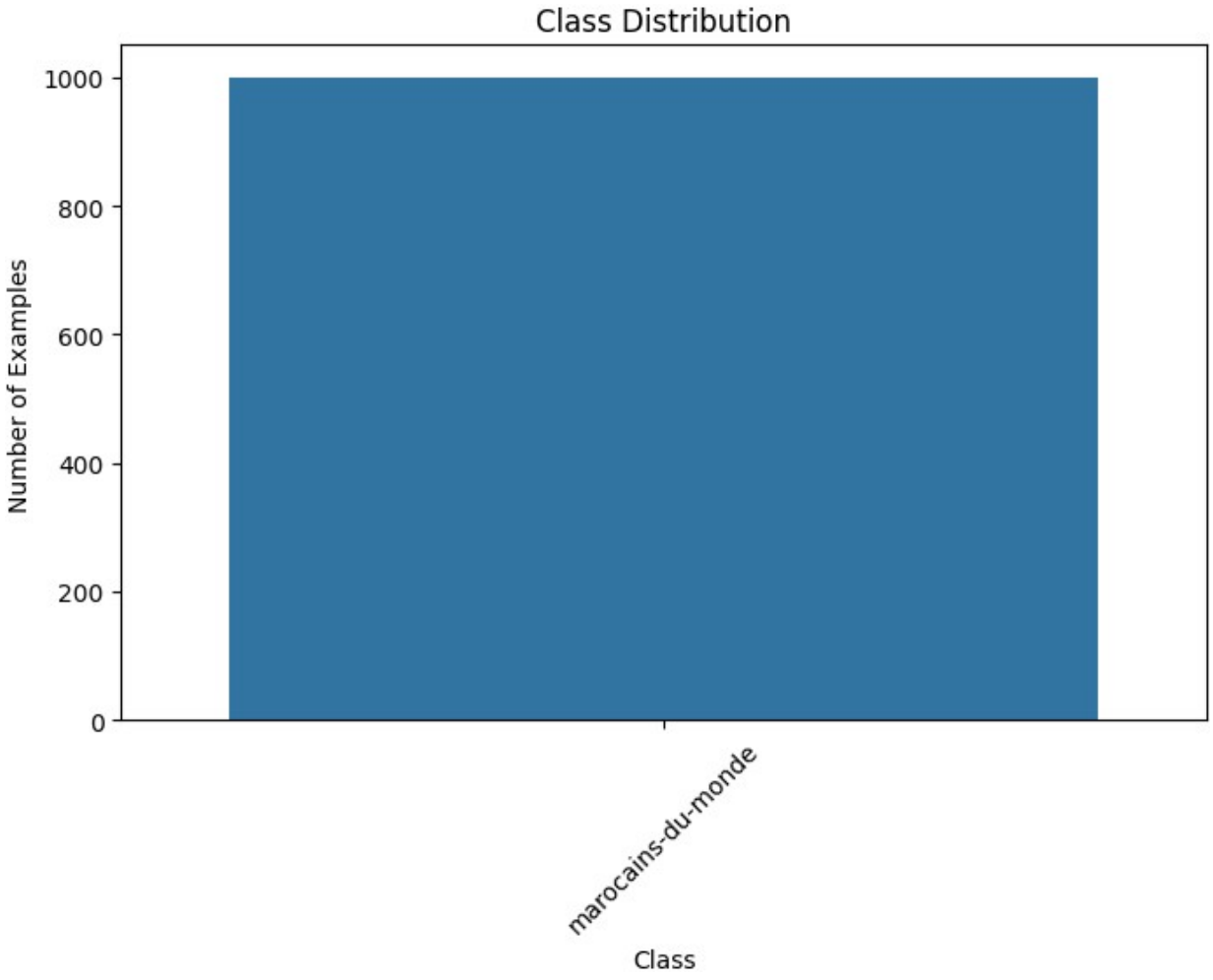
Dataset: stories_marocains-du-monde.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "marocains-du-monde" category (Moroccans living abroad).
- The top 10 frequent 2-grams (word pairs) in the stories are related to Moroccans living abroad, such as "الجالية المغربية" (Moroccan community) and "مغاربة العالم" (Moroccans of the world).

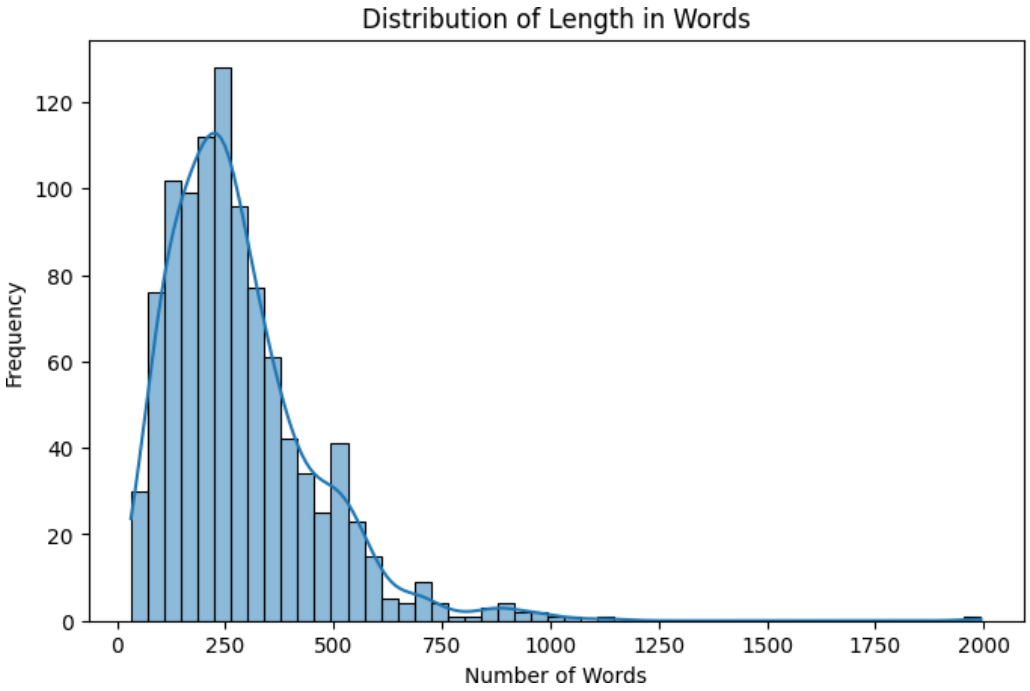
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "marocains-du-monde" category.
- Top 2-grams: Frequent word pairs include "المقيمين بالخارج" "مغاربة العالم," "الجالية المغربية," and others.
- Length of Examples in Words: The average story contains around 283 words, with a minimum of 32 words and a maximum of 1994 words.
- Length of Examples in Letters: The average story contains around 2028 letters, with a minimum of 225 letters and a maximum of 13864 letters.

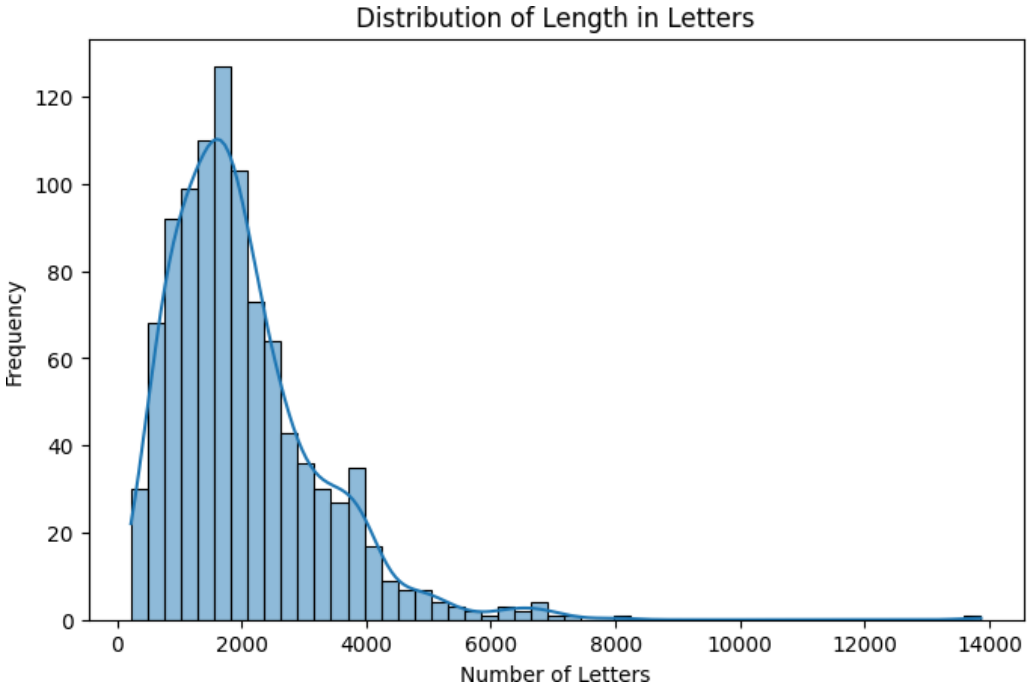
```
--- Individual File Analysis for stories_marocains-du-monde.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
marocains-du-monde    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
610 :الجالية المغربية
476 :مغاربة العالم
372 :المقيمين بالخارج
354 :المغاربة العالقين
273 :أفراد الجالية
222 :المغربية المقيمة
221 :فيروس كورونا
218 :المغاربة المقيمين
176 :محمد السادس
163 :بالمغاربة المقيمين

Length of Examples in Words:
count    1000.000000
mean      283.390000
std       173.448443
min        32.000000
25%       164.000000
50%       247.000000
75%       357.250000
max      1994.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      2027.658000
std       1231.865623
min        225.000000
25%       1186.000000
50%       1785.500000
75%       2549.500000
max      13864.000000
Name: num_letters, dtype: float64
```



Dataset: stories_economie.csv

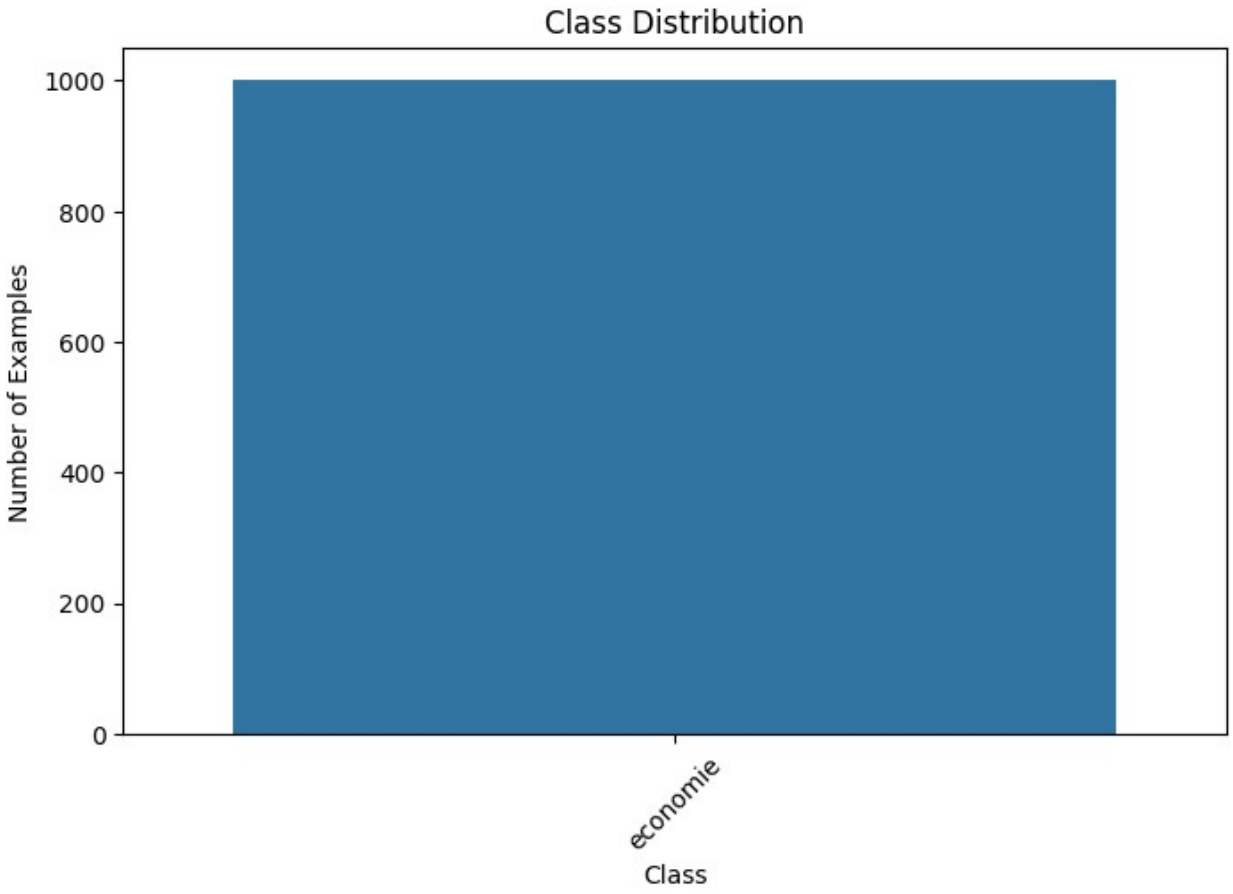
- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "economie" category (economy).
- The top 10 frequent 2-grams (word pairs) in the stories are related to economic terms, such as "مليار درهم" (Billion dirhams) and "جائحة كورونا" (Coronavirus pandemic).

Here are some insights and summary statistics:

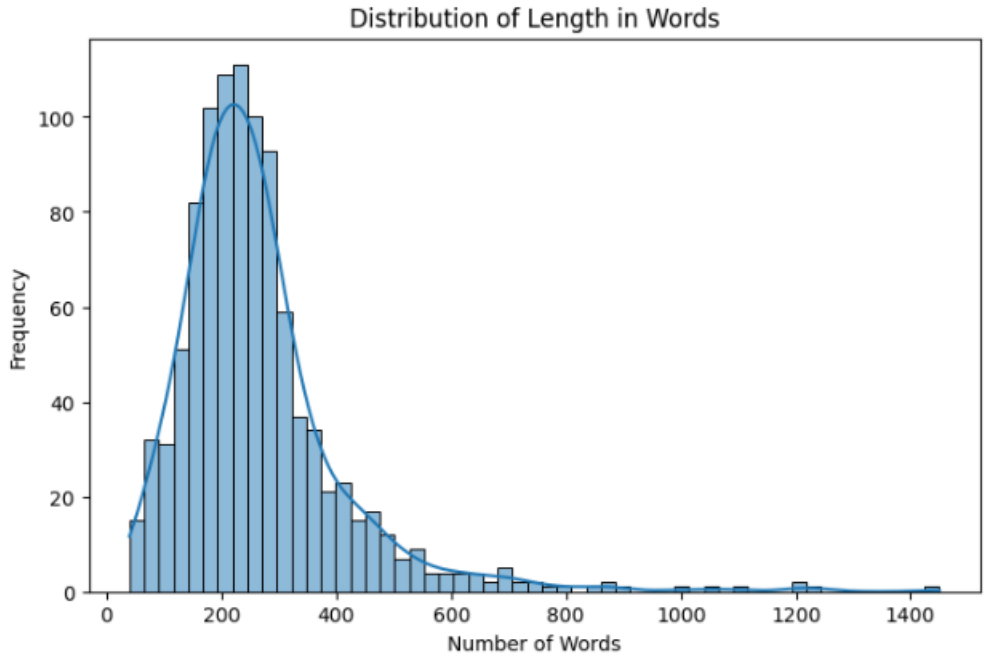
Class Distribution: All 1000 entries belong to the "economie" category.

- Top 2-grams: Frequent word pairs include "كورونا المستجد" "فيروس كورونا," "مليار درهم," and others.
- Length of Examples in Words: The average story contains around 265 words, with a minimum of 40 words and a maximum of 1450 words.
- Length of Examples in Letters: The average story contains around 1876 letters, with a minimum of 264 letters and a maximum of 10768 letters.

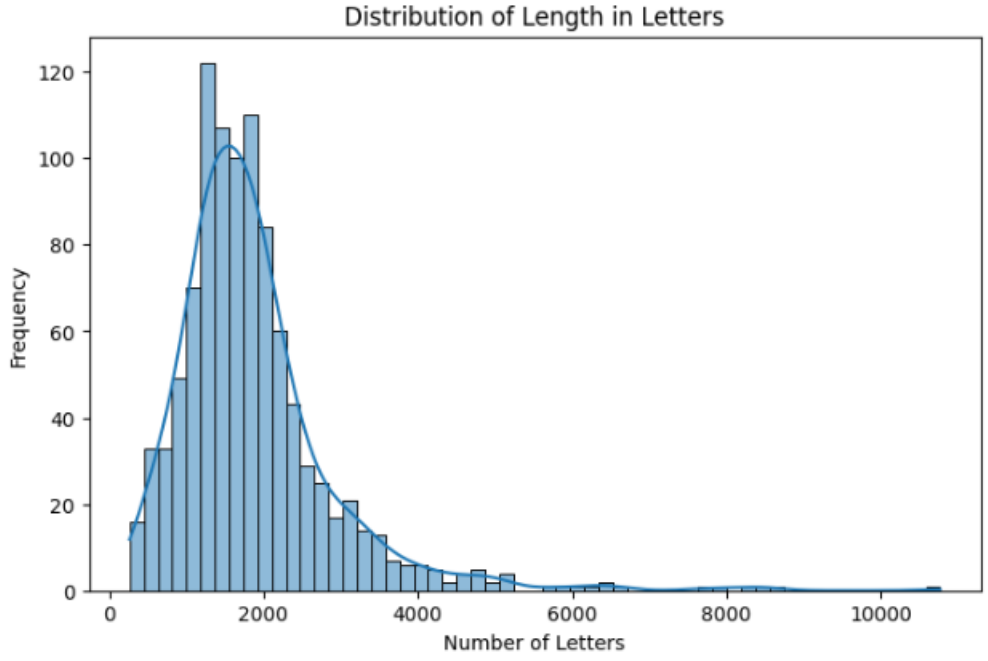
```
--- Individual File Analysis for stories_economie.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
economie    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
586 :مليار درهم
508 :فيروس كورونا
312 :كورونا المستجد
304 :كوفيد 19
279 :الحجر الصحي
256 :جائحة كورونا
251 :مليارات درهم
246 :مليون درهم
238 :السنة الجارية
210 :الاقتصاد والمالية

Length of Examples in Words:
count    1000.000000
mean      264.703000
std       147.975178
min        40.000000
25%       179.000000
50%       238.500000
75%       309.000000
max      1450.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      1876.200000
std       1056.735247
min        264.000000
25%       1267.000000
50%       1685.000000
75%       2194.500000
max      10768.000000
Name: num_letters, dtype: float64
```



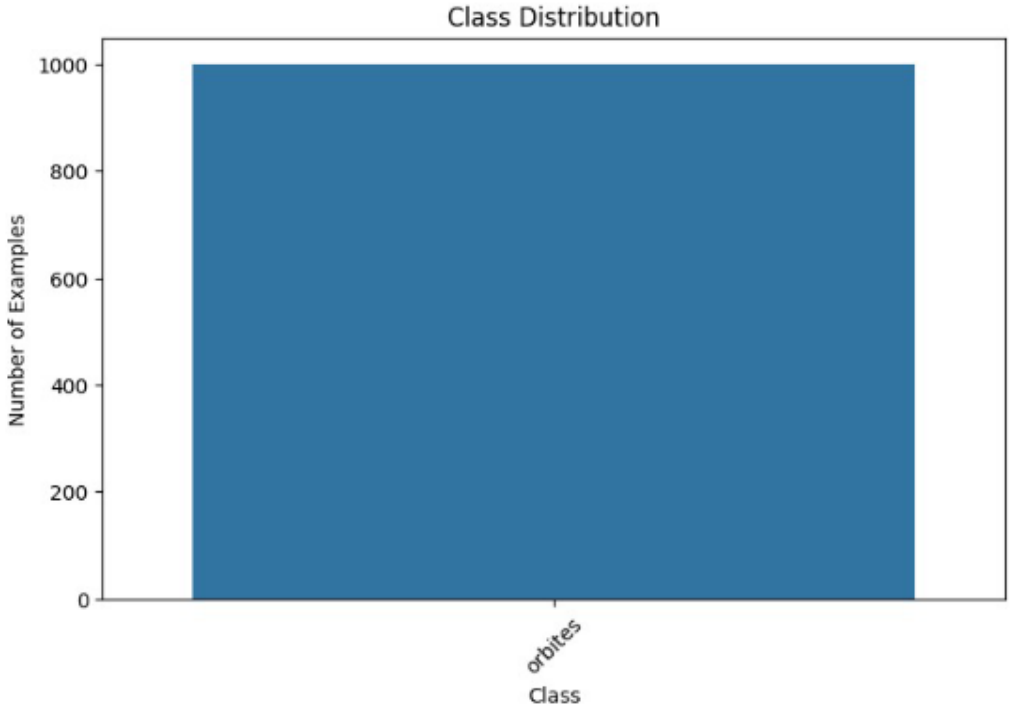
Dataset: stories_orbites.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "orbites" category (space-related).
- The top 10 frequent 2-grams (word pairs) in the stories include terms related to space and events like "فيروس كورونا" (Coronavirus) and "الأمم المتحدة" (United Nations).

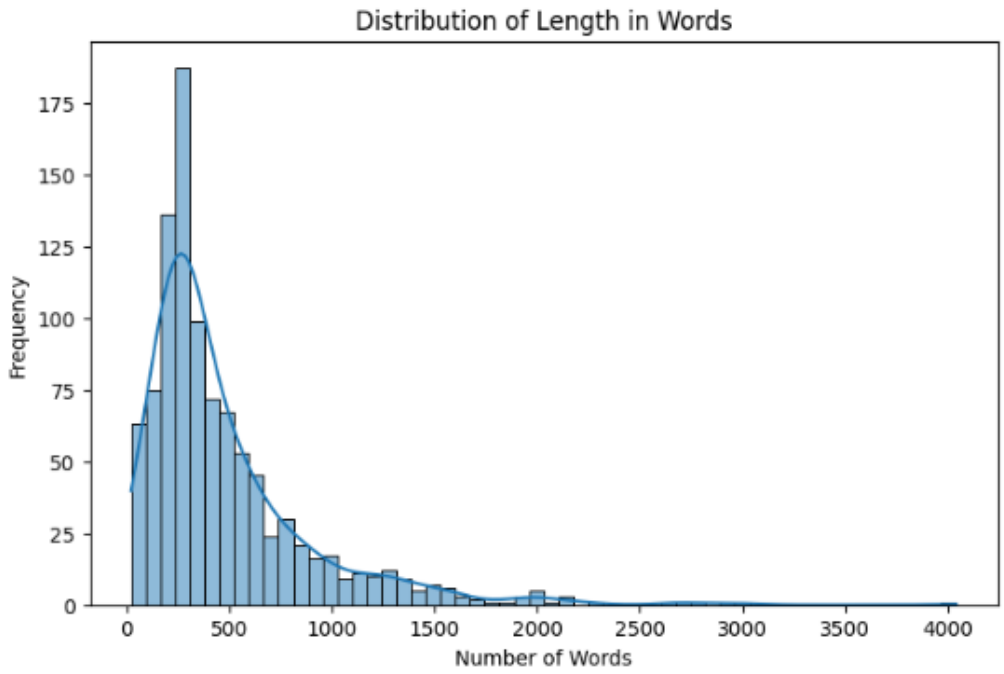
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "orbites" category.
- Top 2-grams: Frequent word pairs include "محمد السادس" ",19 "كوفيد" ",فيروس كورونا," and others.
- Length of Examples in Words: The average story contains around 481 words, with a minimum of 21 words and a maximum of 4040 words.
- Length of Examples in Letters: The average story contains around 3413 letters, with a minimum of 123 letters and a maximum of 27217 letters.

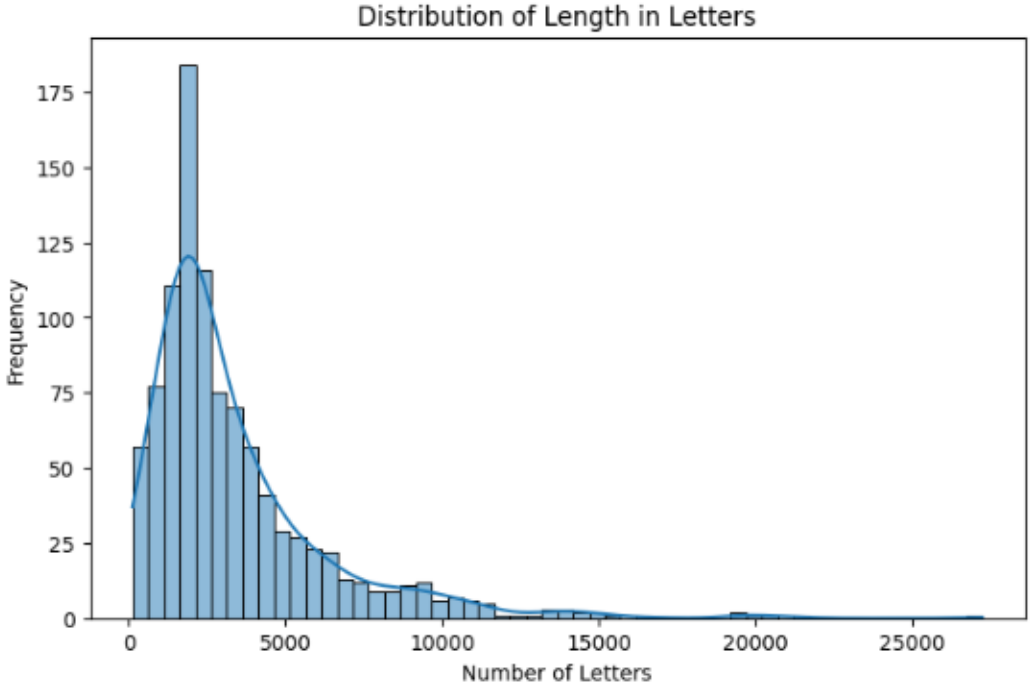
```
--- Individual File Analysis for stories_orbites.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  1000 non-null   int64 
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
orbites    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
414 :فيروس كورونا
402 :19 كوفيد
330 :محمد السادس
289 :الولايات المتحدة
277 :الملك محمد
246 :جائحة كورونا
234 :الحجر الصحي
212 :الأمم المتحدة
195 :كورونا المستجد
192 :صلى الله

Length of Examples in Words:
count    1000.000000
mean      480.672000
std       429.339582
min        21.000000
25%       229.750000
50%       333.500000
75%       594.250000
max      4040.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      3412.555000
std       3006.370769
min        123.000000
25%       1640.000000
50%       2395.500000
75%       4170.000000
max      27217.000000
Name: num_letters, dtype: float64
```



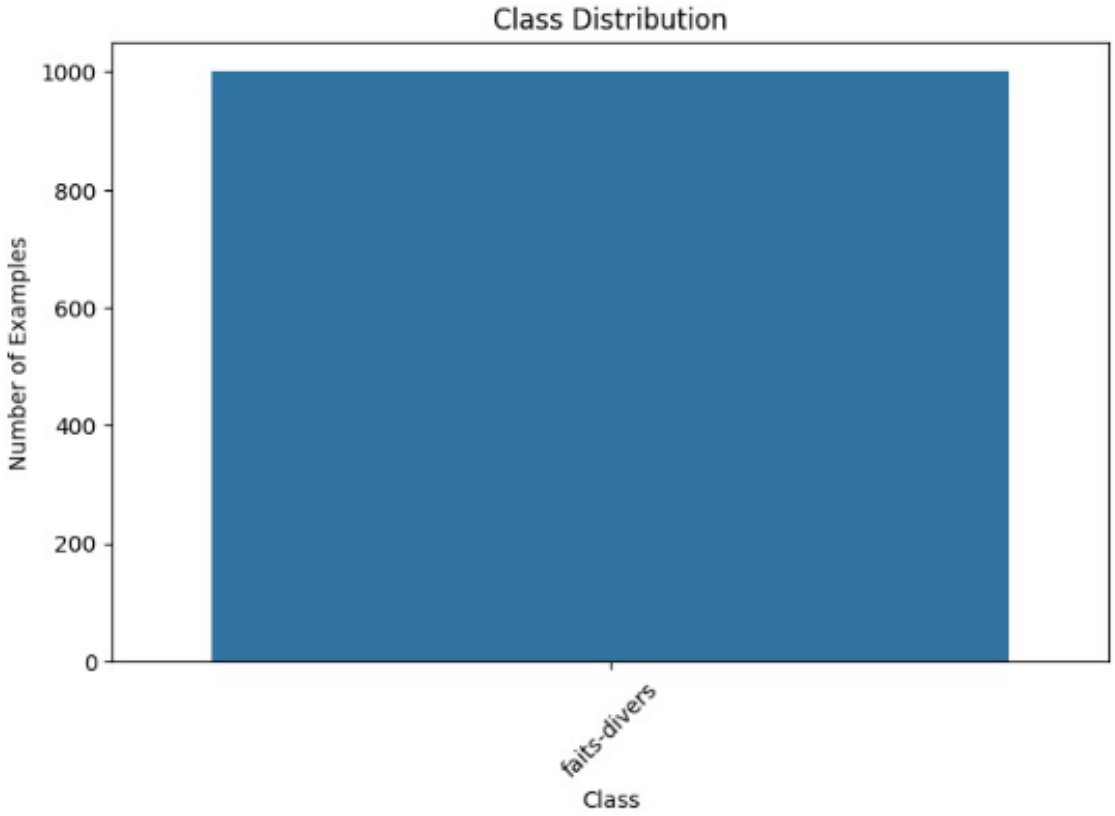
Dataset: stories_faits-divers.csv

- The dataset contains 1000 entries and 7 columns.
- The class distribution shows that all examples belong to the "faits-divers" category (miscellaneous incidents).
- The top 10 frequent 2-grams (word pairs) in the stories include terms related to incidents and authorities like "النيابة العامة" (Public Prosecution) and "الشرطة القضائية" (Judicial Police).

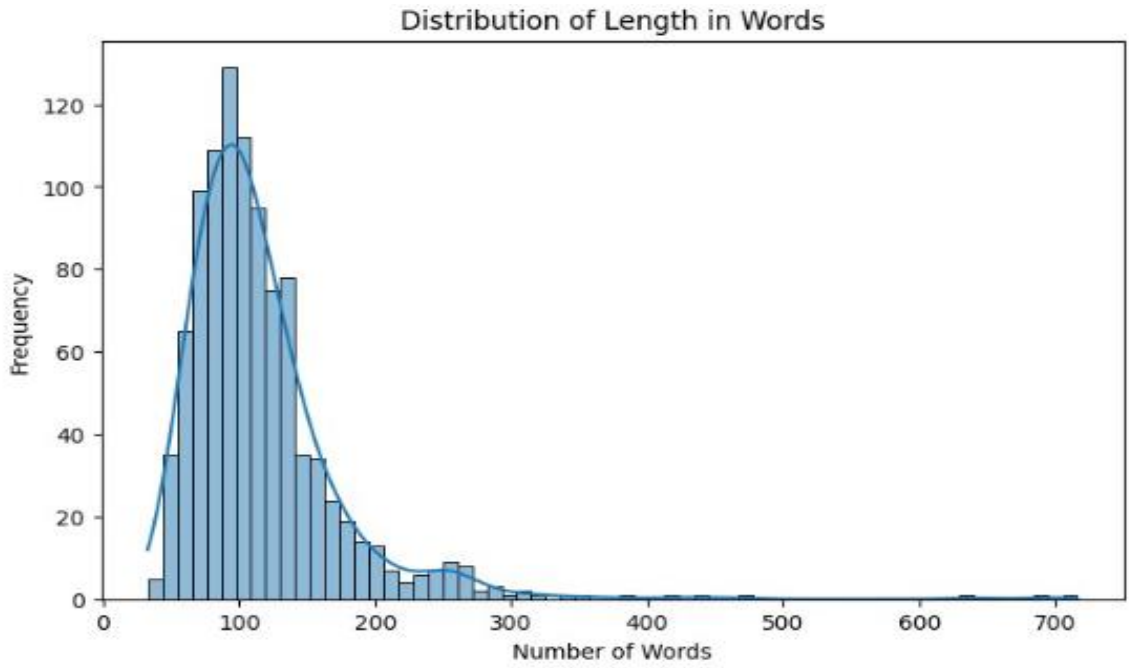
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "faits-divers" category.
- Top 2-grams: Frequent word pairs include "الحراسة النظرية" "العامة المختصة," "النيابة العامة," and others.
- Length of Examples in Words: The average story contains around 116 words, with a minimum of 33 words and a maximum of 716 words.
- Length of Examples in Letters: The average story contains around 803 letters, with a minimum of 264 letters and a maximum of 4862 letters.

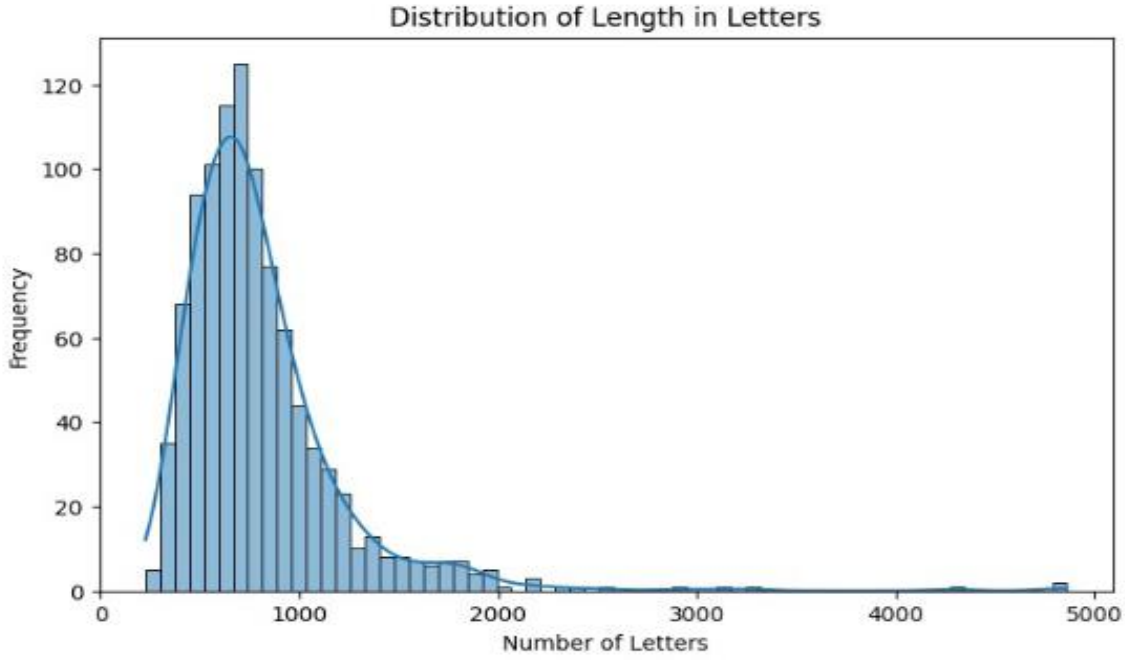
```
--- Individual File Analysis for stories_faits-divers.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
faits-divers    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
730 :النيابة العامة
554 :العامة المختصة
408 :الحراسة النظرية
343 :الدرك الملكي
239 :تدبير الحراسة
239 :مصادر هسبريس
228 :إشراف النيابة
225 :الشرطة القضائية
221 :للأمن الوطني
215 :رهن إشارة

Length of Examples in Words:
count    1000.0000
mean      116.4600
std        60.3919
min        33.0000
25%        80.0000
50%       104.0000
75%       134.2500
max       716.0000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count    1000.000000
mean      802.796000
std       418.832727
min       226.000000
25%       553.000000
50%       712.500000
75%       922.250000
max      4862.000000
Name: num_letters, dtype: float64
```

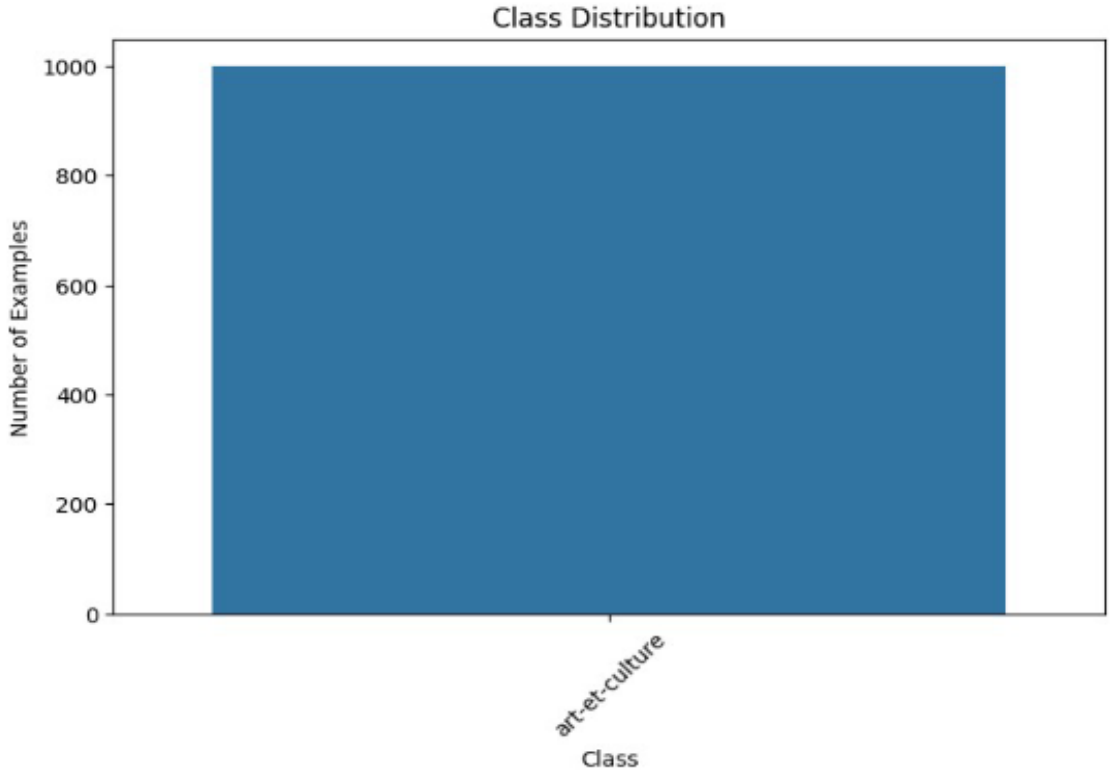


- The class distribution shows that all examples belong to the "art-et-culture" category (art and culture).
- The top 10 frequent 2-grams (word pairs) in the stories include terms related to culture, art, and social media like "وزارة الثقافة" (Ministry of Culture) and "التواصل الاجتماعي" (Social media).

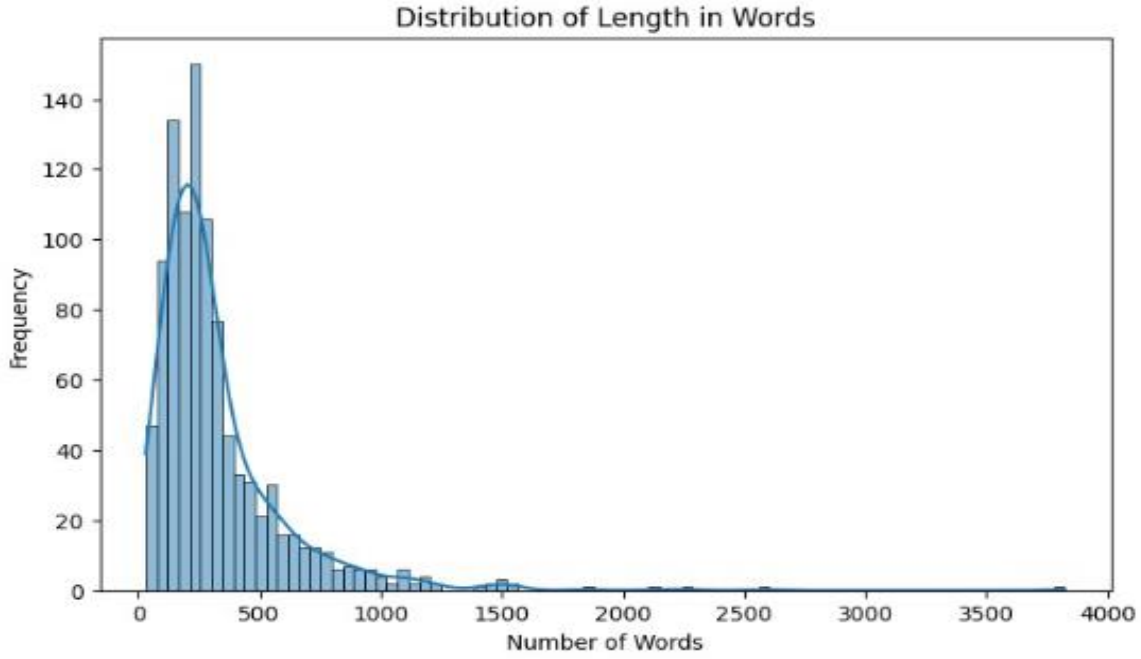
Here are some insights and summary statistics:

- Class Distribution: All 1000 entries belong to the "art-et-culture" category.
- Top 2-grams: Frequent word pairs include "وزارة الثقافة," "فيروس كورونا," "التواصل الاجتماعي" and others.
- Length of Examples in Words: The average story contains around 327 words, with a minimum of 31 words and a maximum of 3823 words.
- Length of Examples in Letters: The average story contains around 2299 letters, with a minimum of 217 letters and a maximum of 25279 letters.

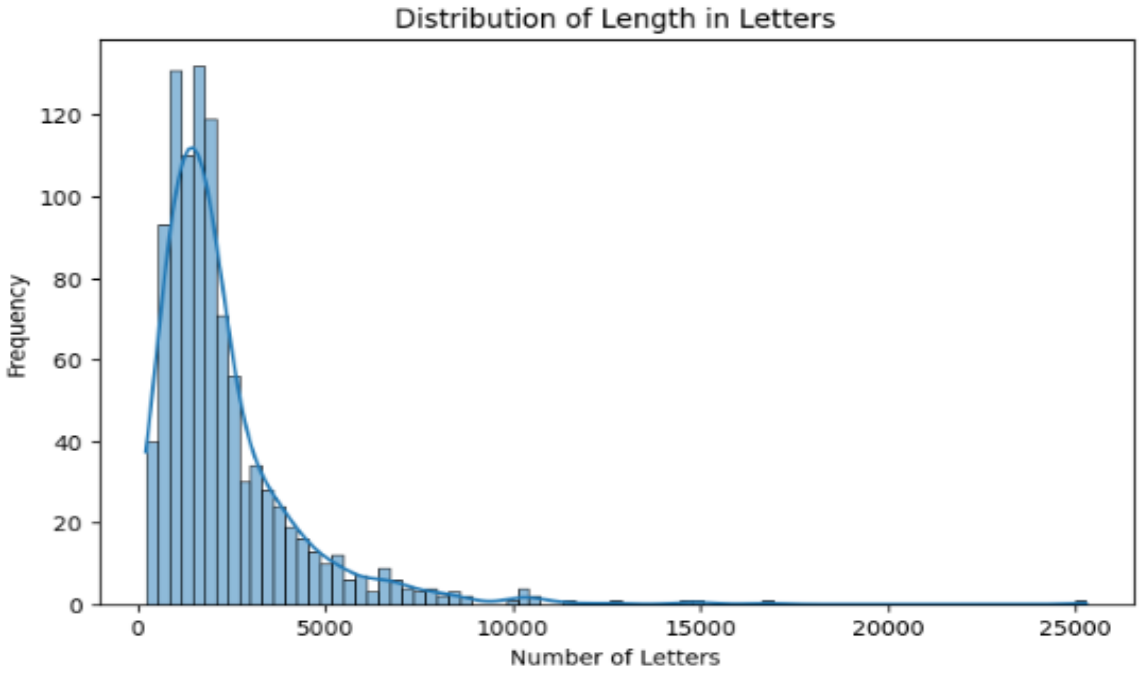
```
--- Individual File Analysis for stories_art-et-culture.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
art-et-culture    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
151 :وزارة الثقافة
138 :فيروس كورونا
135 :والشباب والرياضة
132 :الثقافة والشباب
127 :التواصل الاجتماعي
116 :اللغة العربية
100 :محمد السادس
95 :جائحة كورونا
95 :الدار البيضاء
87 :وزير الثقافة

Length of Examples in Words:
count    1000.000000
mean      327.024000
std       292.535564
min        31.000000
25%       158.000000
50%       245.500000
75%       384.250000
max      3823.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      2298.729000
std       2014.367807
min        217.000000
25%       1103.500000
50%       1742.000000
75%       2668.250000
max      25279.000000
Name: num_letters, dtype: float64
```



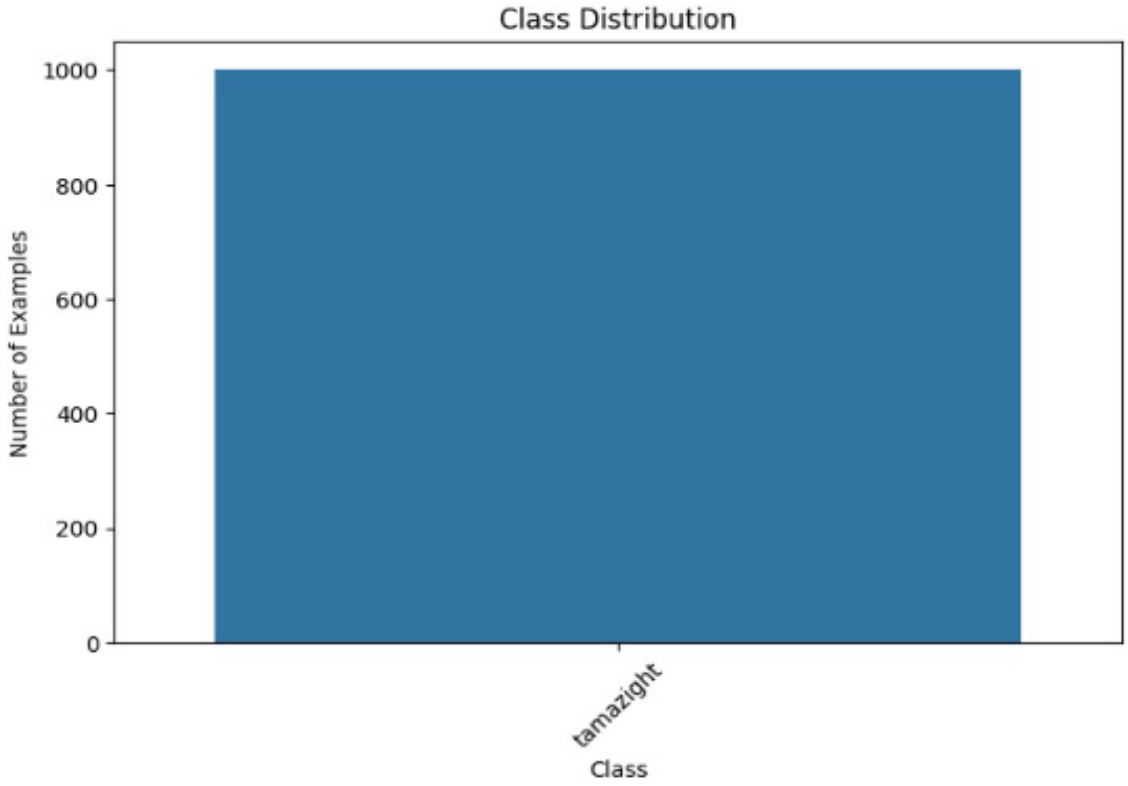
Dataset: stories_tamazight.csv

- The 'stories_tamazight.csv' dataset comprises 1,000 entries and 7 columns. A
- ll examples in this dataset are categorized under the "tamazight" topic, demonstrating a balanced class distribution.
- Top 10 frequent 2-grams (word pairs) in the stories primarily revolve around Amazigh (Berber) language and cultural themes. Noteworthy 2-grams include "اللغة الأمازيغية" (The Amazigh language), "الثقافة الأمازيغية" (Amazigh culture), and other phrases that reflect the rich Amazigh heritage.

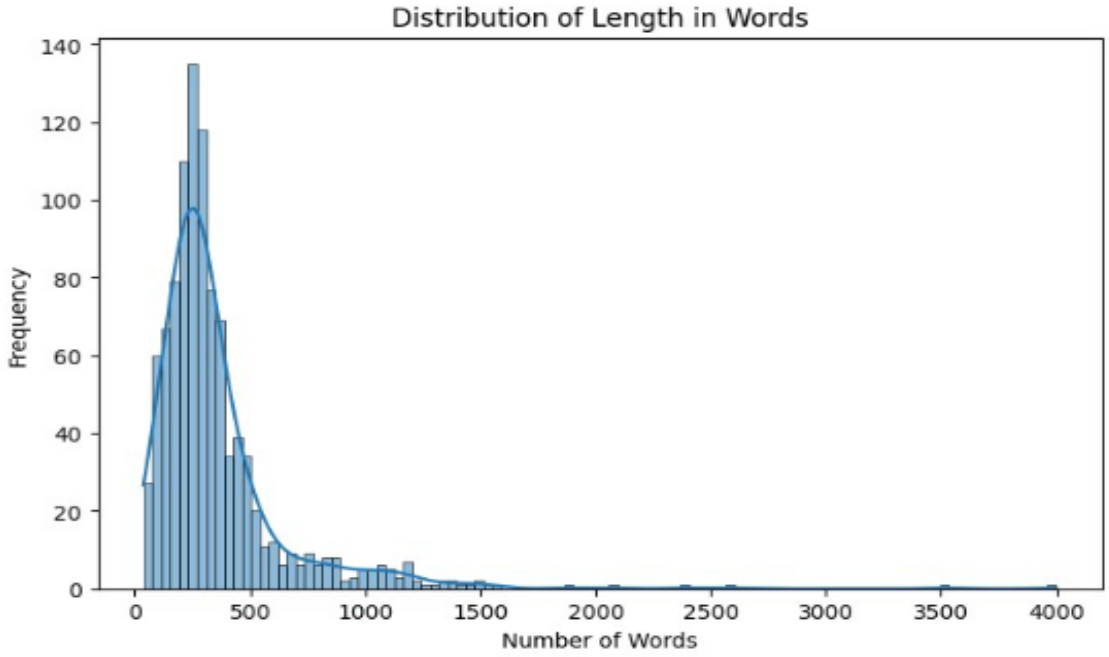
Here are some insights and summary statistics:

- Class Distribution: All 1,000 entries belong to the "tamazight" category.
- Top 2-grams: Frequent word pairs highlight Amazigh language and cultural themes.
- Length of Examples in Words: Stories range from short to long narratives, with an average of approximately 357 words.
- Length of Examples in Letters: The average story consists of around 2333 letters, revealing the diversity in story lengths.

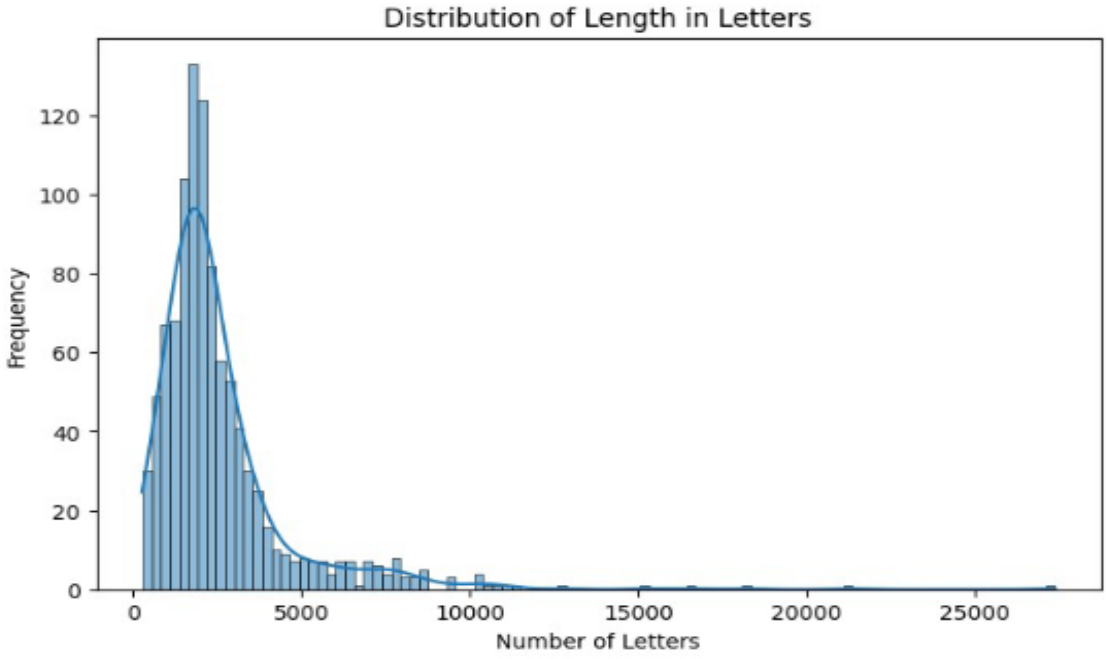
```
--- Individual File Analysis for stories_tamazight.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
tamazight    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
900 :اللغة الأمازيغية
622 :القانون التنظيمي
548 :للثقافة الأمازيغية
521 :الطابع الرسمي
501 :الملكي للثقافة
449 :الحركة الأمازيغية
446 :الرسمي للأمازيغية
417 :المعهد الملكي
381 :السنة الأمازيغية
322 :مشروع القانون

Length of Examples in Words:
count    1000.000000
mean      357.165000
std       315.515775
min        38.000000
25%       200.750000
50%       275.000000
75%       395.000000
max      3999.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      2563.571000
std       2181.851074
min        264.000000
25%       1467.250000
50%       2004.000000
75%       2846.500000
max      27364.000000
Name: num_letters, dtype: float64
```

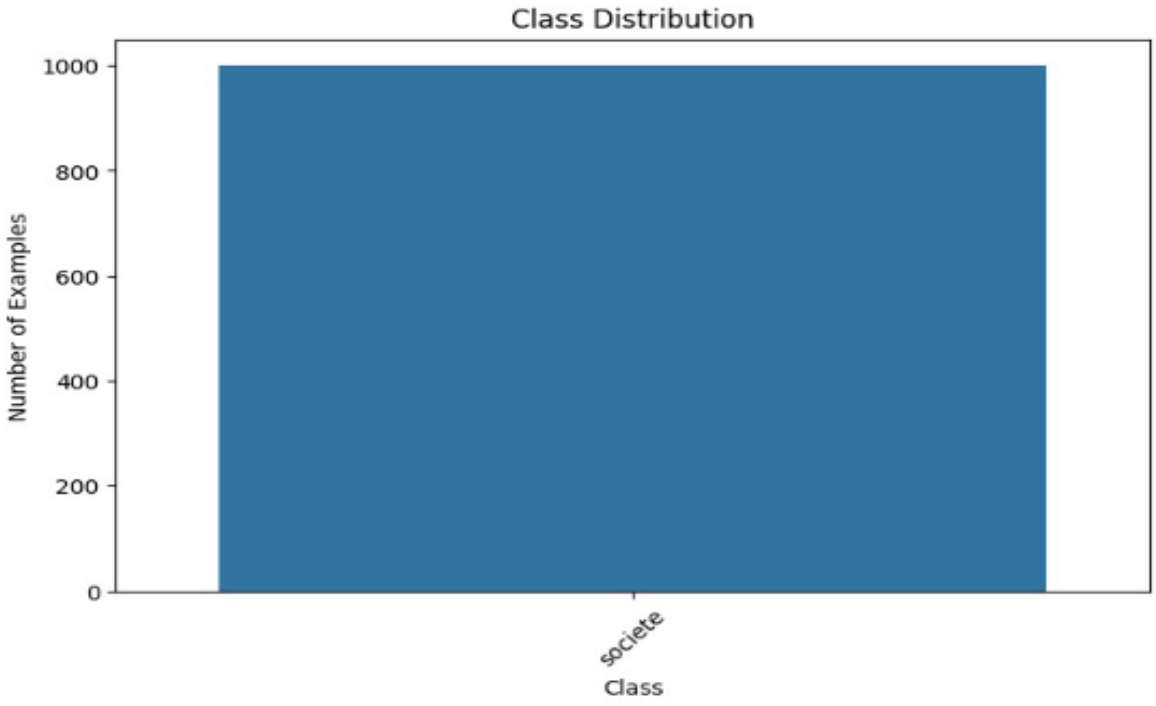


- The 'stories_societe.csv' dataset includes 1,000 entries and 7 columns.
- Similar to the previous dataset, all examples fall under the "societe" topic, demonstrating a uniform class distribution.
- The top 10 frequent 2-grams (word pairs) in the stories predominantly pertain to societal topics and issues. Notable 2-grams comprise "وزارة الصحة" (Ministry of Health), "كوفيد 19" (COVID-19), and other phrases related to social and health matters, indicating that the stories encompass a wide range of societal themes.

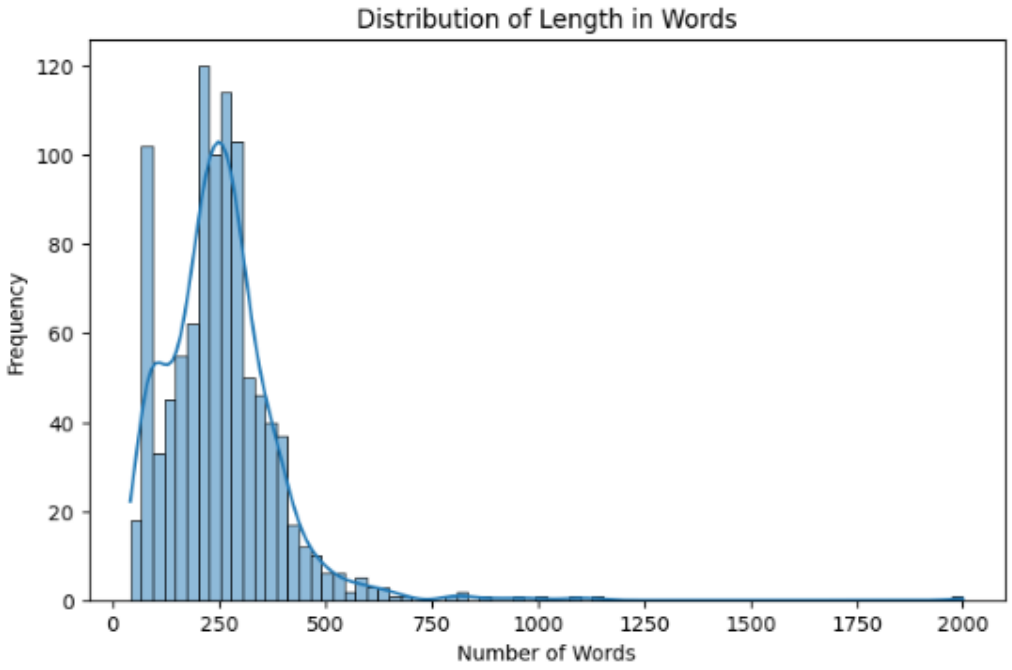
Here are some insights and summary statistics:

- Class Distribution: All 1,000 entries belong to the "societe" category.
- Top 2-grams: Frequent word pairs highlight various societal topics and health-related matters.
- Length of Examples in Words: Stories vary in length, with an average of approximately 253 words.
- Length of Examples in Letters: The average story consists of around 1803 letters, indicating diversity in the complexity and depth of societal narratives.

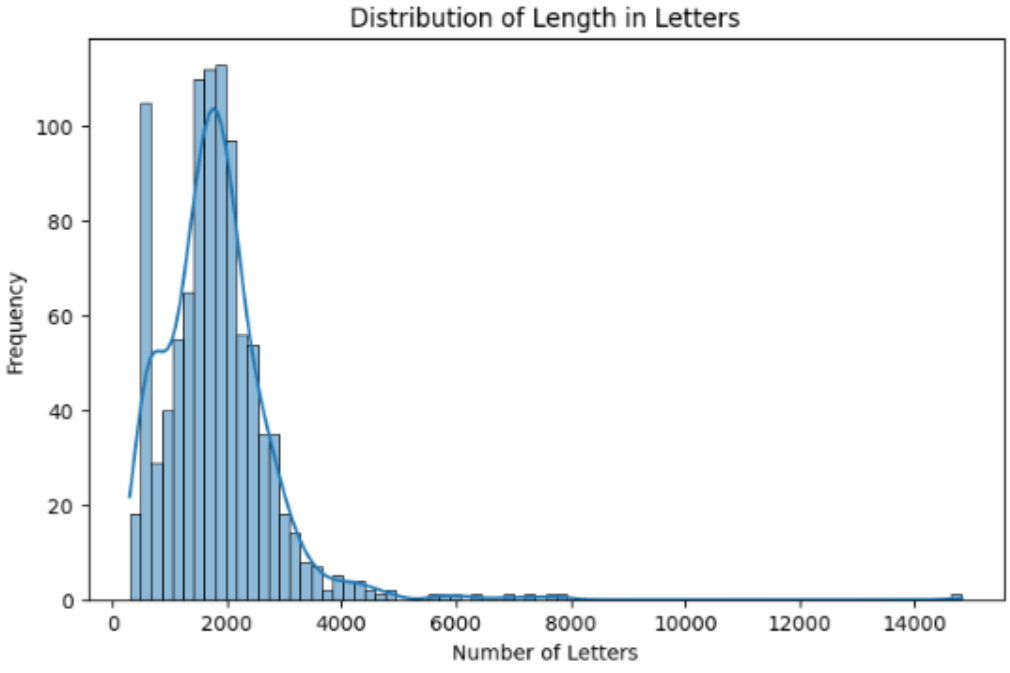
```
--- Individual File Analysis for stories_societe.csv ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  1000 non-null   int64
 1   id          1000 non-null   object
 2   title       1000 non-null   object
 3   date        1000 non-null   object
 4   author      1000 non-null   object
 5   story       1000 non-null   object
 6   topic       1000 non-null   object
dtypes: int64(1), object(6)
memory usage: 54.8+ KB
None

Class Distribution:
societe    1000
Name: topic, dtype: int64
```




```
Top 10 Frequent 2 -grams:
471 :وزارة الصحة
428 :19 كوفيد
347 :التربية الوطنية
333 :فيروس كورونا
289 :بفيروس كورونا
271 :الحجر الصحي
256 :كورونا المستجد
222 :وزارة التربية
212 :هسبريس الإلكترونية
191 :الدخول المدرسي

Length of Examples in Words:
count    1000.000000
mean      253.744000
std       138.229359
min        42.000000
25%       173.000000
50%       244.000000
75%       306.000000
max      2000.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count     1000.000000
mean      1803.464000
std        989.322978
min        303.000000
25%       1244.500000
50%       1739.500000
75%       2186.250000
max      14827.000000
Name: num_letters, dtype: float64
```



Summary of each dataset analysis:

I have performed analysis on various CSV files, each representing different topics or classes (e.g., politique, medias, regions, sport, etc.). Here are the insights from the analysis:

1. Dataset Information:

- The dataset contains a total of 1000 entries in each CSV file.
- The data has 7 columns: 'Unnamed: 0', 'id', 'title', 'date', 'author', 'story', and 'topic'.
- The 'topic' column represents the class distribution for each file.

2. Class Distribution:

- Each CSV file represents a specific topic, and each topic has 1000 examples.
- The class distribution shows the number of examples for each topic.

3. Top 10 Frequent 2-grams:

- For each CSV file, the top 10 most frequent 2-grams have been identified based on the 'story' content.
- The 2-grams represent pairs of adjacent words that appear most frequently in the text.

4. Length of Examples:

- The length of examples is analyzed in terms of word count and letter count.
- The word count statistics (mean, standard deviation, min, max) for each file indicate the average and variation in the number of words per example.
- The letter count statistics (mean, standard deviation, min, max) for each file indicate the average and variation in the number of letters per example.

Overall, the analysis gives an overview of the content and distribution of each CSV file, providing valuable insights into the data and its characteristics for each topic.

Overview Analysis of All 'stories' Files

Here are the insights from the analysis:

1. Dataset Information:

- The combined dataset contains a total of 11,000 entries.
- Each entry has 7 columns: 'Unnamed: 0', 'id', 'title', 'date', 'author', 'story', and 'topic'.
- The 'topic' column represents the class distribution for each entry.

2. Class Distribution:

- Each 'stories' file corresponds to a specific topic, and there are 1,000 examples for each topic.
- The class distribution shows the number of examples for each topic, and all topics have an equal number of examples.

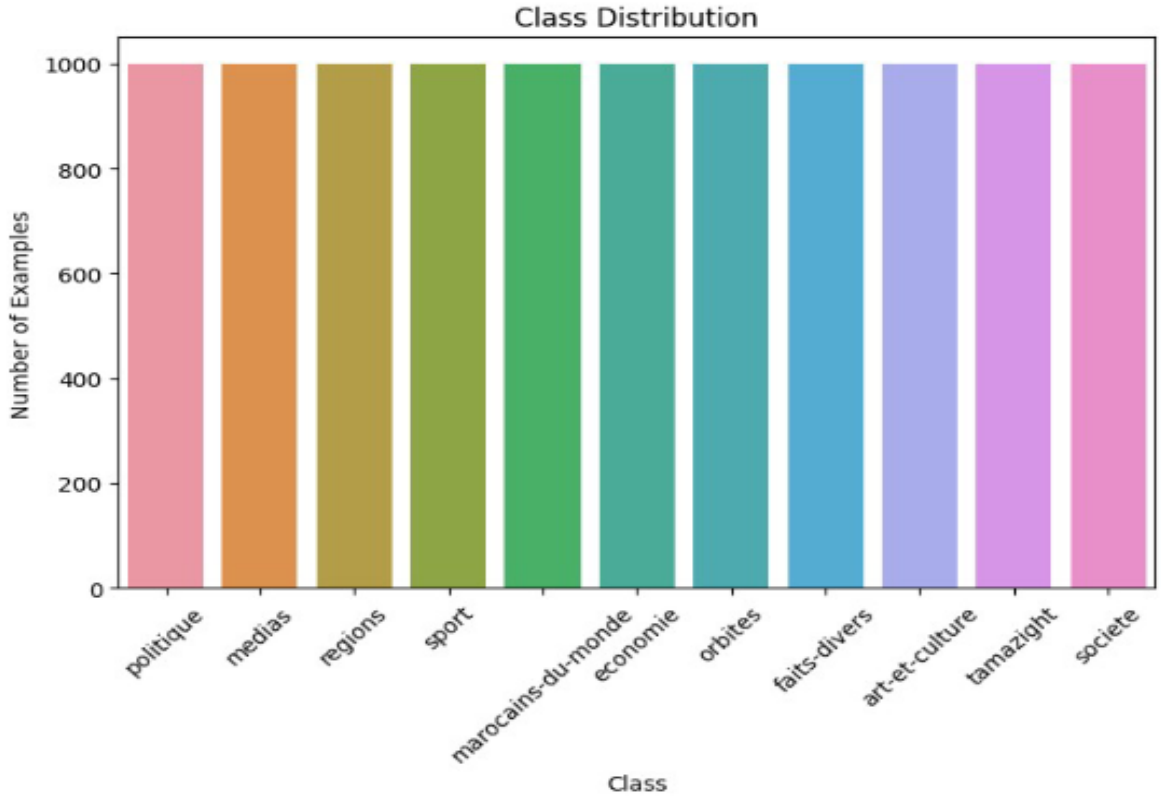
3. Top 10 Frequent 2-grams:

- The top 10 most frequent 2-grams have been identified based on the 'story' content across all 'stories' files.
- The most common 2-grams include phrases related to "فيروس كورونا" (Coronavirus), " كوفيد 19" (COVID-19), and other related terms.

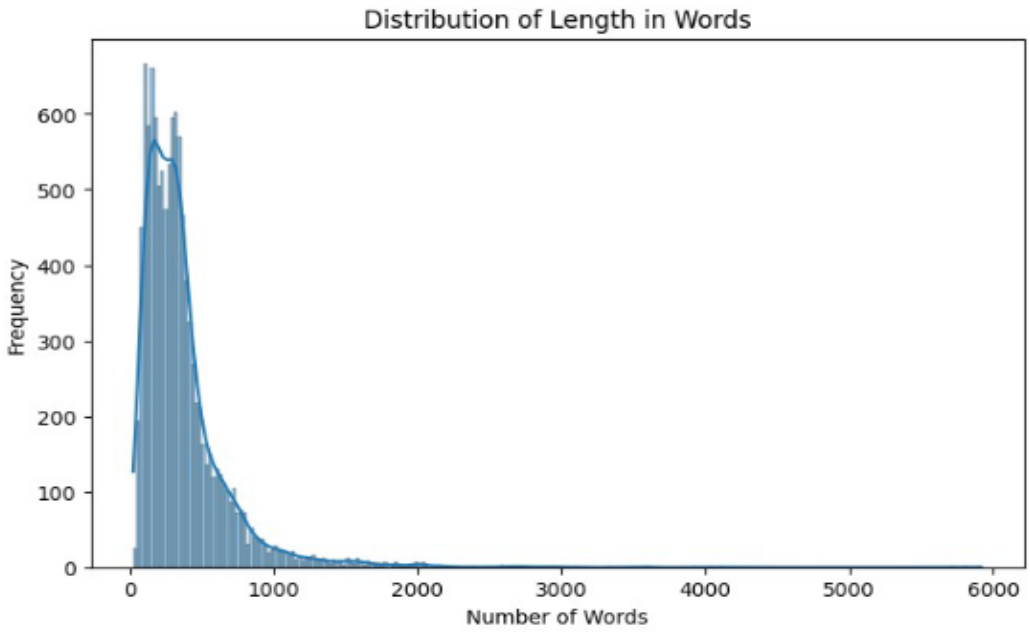
4. Length of Examples:

- The length of examples is analyzed in terms of word count and letter count.
- The 'num_words' column provides statistics (count, mean, standard deviation, min, 25%, 50%, 75%, max) for the number of words in each example.
- The 'num_letters' column provides statistics (count, mean, standard deviation, min, 25%, 50%, 75%, max) for the number of letters in each example.

Overall, the analysis gives an overview of the content and distribution of the 'stories' dataset, providing valuable insights into the data and its characteristics. The dataset contains various topics, and each example has a different length in terms of words and letters. The top frequent 2-grams suggest the prevalence of certain phrases and topics in the stories.




```
Top 10 Frequent 2 -grams:
2839 :فيروس كورونا
2247 :كوفيد 19
1943 :كورونا المستجد
1561 :محمد السادس
1509 :الحجر الصحي
1443 :الدار البيضاء
1262 :النيابة العامة
1256 :جائحة كورونا
1232 :رئيس الحكومة
1230 :بفيروس كورونا

Length of Examples in Words:
count    11000.000000
mean       370.494909
std        330.396957
min         24.000000
25%        173.000000
50%        298.000000
75%        439.000000
max       5920.000000
Name: num_words, dtype: float64
```




```
Length of Examples in Letters:
count    11000.000000
mean      2333.271636
std       2032.348676
min        134.000000
25%       1101.000000
50%       1889.000000
75%       2780.250000
max      34371.000000
Name: num_letters, dtype: float64
```