import libraries

In [4]:
```
import pandas as pd
import matplotlib.pyplot as plt
import os
import nltk
from nltk.tokenize import word_tokenize
from nltk.util import ngrams
import matplotlib.pyplot as plt
from collections import Counter
import string
```

In [5]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\LENOVO\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[5]:
```
True
```

read data and Concatenate all csv files

In [6]:
```
# Assuming all CSV files are in the same directory
data_directory = "C:/Users/LENOVO/Downloads/archive/stories/"
csv_files = [file for file in os.listdir(data_directory) if file.endswith('.csv')]

# Initialize an empty list to store DataFrames from each CSV
dfs = []

for file in csv_files:
    file_path = os.path.join(data_directory, file)
    df = pd.read_csv(file_path)
    dfs.append(df)

# Concatenate all DataFrames into a single DataFrame
combined_data = pd.concat(dfs, ignore_index=True)
```

In [7]:
```
combined_data.head()
```

Out[7]:

	Unnamed: 0	id	title	date	author	story	topic
0	0	f06aa998054e11eba66e646e69d991ea	بيت الشعر" يسائل وزير الثقافة" عن كوابيس سوداء	الجمعة 02 أكتوبر 2020 - 23:19	هسبريس من الرباط	وجه "بيت الشعر في المغرب" إلى وزير الثقافة وال...	art-et-culture
1	1	f1cf1b9c054e11ebb718646e69d991ea	مهرجان "سينما المؤلّف" يستحضر روح ثريا جبران	الجمعة 02 أكتوبر 2020 - 07:26	هسبريس من الرباط	في ظلّ استمرار حالة الطوارئ الصحية المرتبطة بج...	art-et-culture
2	2	f2d282a4054e11eb800f646e69d991ea	فيلم "بدون عنف" لهشام العسري .."كعب الحذاء ووا...	الجمعة 02 أكتوبر 2020 - 04:00	عفيفة الحسينات*	تشير مشاهدة فيلم قصير ضمن الثلاثية الأخيرة للم...	art-et-culture
3	3	f3f46cac054e11eba403646e69d991ea	تنين ووهان" .. مريم أيت أحمد توقّع أولى "روا...	الجمعة 02 أكتوبر 2020 - 02:00	حاورَها: وائل بورشاشن	مِن قلب أيّام "الحَجْر"، رأتِ النّورَ الفصول ...	art-et-culture
4	4	f50f0476054e11eba31b646e69d991ea	مسكر يتخلّى عن دعم "الوزارة" بسبب "الجمهور	الخميس 01 أكتوبر 2020 - 19:40	هسبريس من الرباط	أعلن الفنان المغربيّ سعيد مسكر تخليه عن مبلغ ا...	art-et-culture

In [8]:
```
combined_data.columns
```

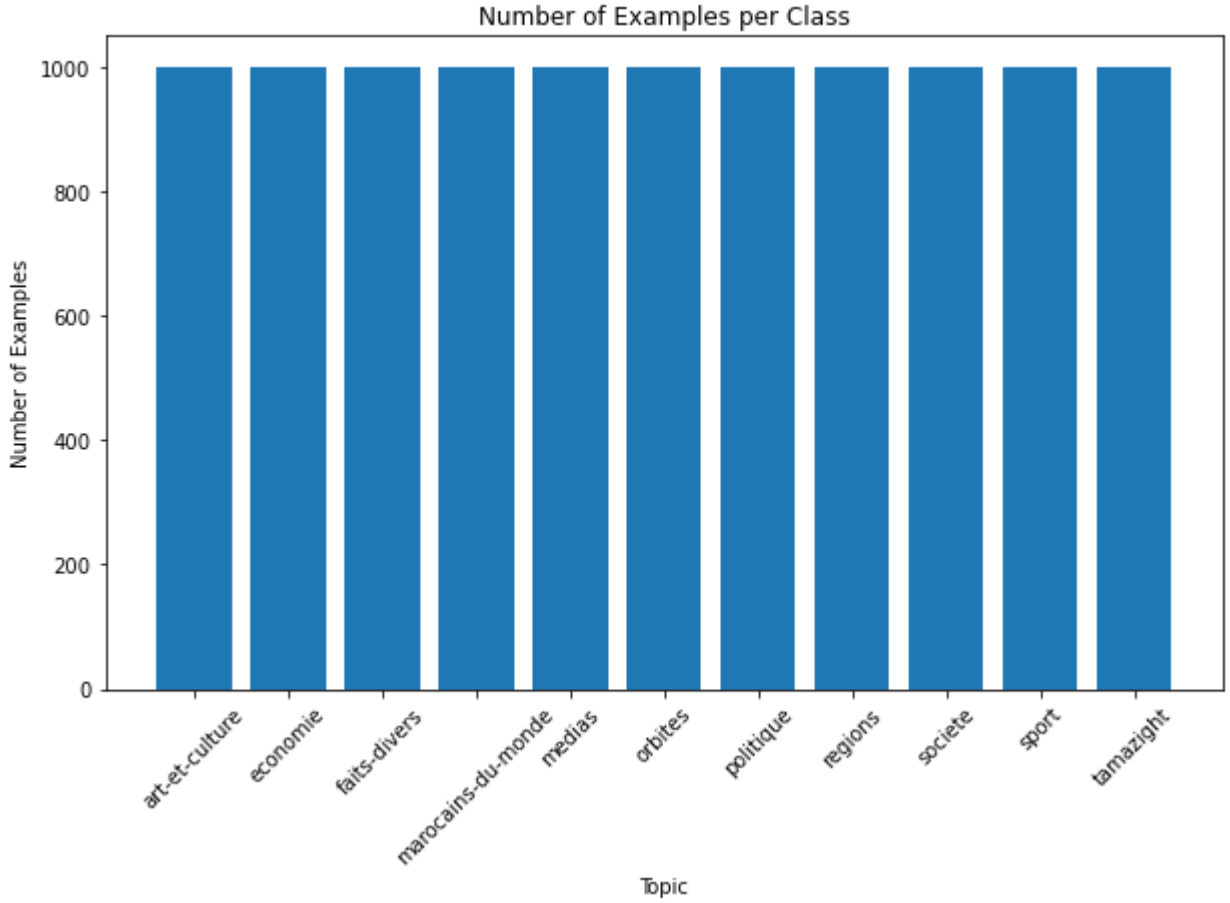
Out[8]:
```
Index(['Unnamed: 0', 'id', 'title', 'date', 'author', 'story', 'topic'], dtype='object')
```

In [9]:
```
print(combined_data['story'][0])
```

وجه "بيت الشعر في المغرب" إلى وزير الثقافة والشباب والرياضة رسالة موسومة بـ"لماذا تحولت أحلام بيضاء إلى كوابيس سوداء؟"، أشار من خلالها إلى أن "بيت الشعر تأسّس سنة 1996 بهدف ت
حقيق جُملة من الأهداف التي تروم جميعُها تعزيز مكانة الشّعر في المجتمع والحياة وترسيخ مكانته بين الناس كحاملٍ لقيم الحلم والخيال؛ فمن تلك الأهداف توطين الشعر المغربي في المقررات الدراسية وت
شجيع التلاميذ والطلبة على قراءته وتذوّق جمالياته، خاصّة في اللحظة التي ينتصرُ فيها لكل ما هو مدهش وإنساني".وجاء في الرسالة، التي توصلت هسبريس بنسخة منها، أن "بيت الشعر نجح في المغ
رب خلال مسيرته الطويلة التي تمتدّ على مدى ربع قرن في إقناع المنظمة العالمية للتربية والثقافة والعلوم (اليونيسكو) بإحداث يوم عالمي للشعر عن طريق المقترح الذي تقدّم به، والذي تبنته الحكومة ا
لمغربية في عهد الراحل عبد الرحمان اليوسفي، كما نجح في ضمان مكانة عالمية لجائزته الشعرية المعروفة "الأركانة" التي صار شعراء العالم يتطلعون للفوز بها، علاوة على انتظام منشوراته الشعرية
ومجلته الرصينة "البيت"، واستدامة برامجه الشعرية بالتعاون مع عدد من الشركاء الذين آمنوا بجدية مشروع مؤسستنا وجودة مستواه فنيا وجماليا".وأشارت الرسالة ذاتها إلى أنه "سعيا من بيت الشعر ف
ي المغرب إلى تقديم خِدمة جديدة للشعر المغربي، بربطه بالفضاء الإبداعي العام وخاصة الموسيقي والغنائي منه، واستعادةِ لحظات طيبة الذكر تعانق فيها شعرُنا المغربي (عبد الرفيع جواهري وإدريس الج
اي والخمار الكنوني وحسن المفتي وأحمد الطيب لعلج وعلي الحداني...) مع أوتار وألحان (عبد السلام عامر وعبد النبي الجيراري وحسن القديري...)، تقدم بمشروع فني تحت عنوان أحلام بيضا
ء".وأوضح "بيت الشعر" أن المشروع "عبارة عن مقطوعات غنائية جعلت من الشعر المغربي متنا لها، بهدف استعادة ماض جميل، أثرى خلاله الشعراء المغاربة السّجل الشعري للأغنية المغربية، فقد
منحُوها قصائد تحوّلت، من خلال أوتار مُلحنين مُقتدرين، إلى أغانٍ تتردّد على ألسنة وشفاهِ الناس في المناسبات والأفراح والأعراس، بل إنّ بعضها سيحظَى باهتمام مطربين مشارقة، أعادوا أداءها وتسجيلها
ا من جديد بأصواتهم".وورد في الرسالة ذاتها أنه "نجاحٌ ما كان له أنْ يتحقّق لولا أنّ هؤلاء الشعراء كتبوا قصائدَهم من داخل تجربتهم الإنسانية وفي أفق الرؤية التي امتلكوها تجاه اللغة والمجتمع والك
ون، وليس تحت الطلب أو إرغامات سوق الغناء".وحسب الرسالة ذاتها، فإنه "لأول مرّة في تاريخ برنامج الوزارة لدعم الأغنية المغربية تتقدّم هيأة متخصّصة كبيت الشعر في المغرب بمشروعٍ ذي رؤية
مندمجة، تربطُ الشعر المغربي بأفُقه الفنّي من أجل أن يحظى بمساحةٍ أوسع للانتشار عبر الأغنية؛ حيث يصيرُ بمقدور الجمهور المغربي أن يتعرّف على شعرائنا المغاربة، ليس من خلال دواوينهم الشعرية
أو عبر أمسياتهم الثقافية، بل من خلال الأغنية كلحظة فنية تتجمّع وتنصهر فيها عدّة أبعاد شعرية، ولحنية موسيقية، وطربية غنائية...".وقد سبق للفنانة صباح زيداني أن جعلت من الشعر المغربي أفقا
لتفكيرها واشتغالها، تضيف الرسالة ذاتها، "عندما قدمت بعضا من نصوصه الجميلة للشعراء بوجمعة العوفي وعبد الهادي السعيد، كما خاضت تجربة فنية مع الشاعر المغربي الكبير عبد الله زريقة بمعية ثُ
لة من الشعراء والموسيقيين الأجانب؛ وهو ما يجعل وجودَها ضمن هذا المشروع تثمينا لهذه الإرادة التي تلتقي فيها برغبة بيت الشعر في المغرب، في أن يكون شعرنا المغربي حاضرا في مختلف الحوام
ل التي تتيح له الذيوع والانتشار (مسرح، تشكيل، أغنية...)".وأكّد "بيت الشعر" أن "تشبيك الفنُون في ما بينها وتجسير الصلات بين مكوناتها أحد أهم مخرجات هذا المشروع، الذي يجمع شعراء مر
موقين، بملحنين مقتدرين وموزعين أكفاء وعازفين ماهرين، عِلاوة على الحضور الواعي والباذخ للفنانة صباح زيداني؛ جميع هؤلاء انخرطوا في هذا المشروع من أجل أنْ نعيد للكلمة الشعرية بريقها في لَ
حظة تعالقِها مع اللحن والموسيقى".واختتم "بيت الشعر" رسالته بالتساؤل: "لماذا تحولت الأحلام البيضاء إلى كوابيس سوداء؟ ولماذا خيّبت اللجنة التي شكلتموها من أجل دعم الأغنية المغربية والارتقاء بـ
مستواها ظنكم وظن الشعب المغربي الذي يموّل من ماله العام سياسة ثقافية فاشلة؟ وماذا ستفعلون لإنصاف بيت الشعر في المغرب وتحقيق المساواة وتكافؤ الفرص التي ينص عليها دستور المملكة؟ وكيف
ستردون من خلال إجراءات عملية وتدابير إدارية عن الضجة- الفضيحة التي خلفتها نتائج هذه السنة في مجال دعم الأغنية؟"، مضيفا "إننا نطعن في نتائج هذه الدورة من برنامج دعم الموسيقى والأغنيـ
ـة، وإننا ننتظر ما ستقومون به".

exploratory data analysis

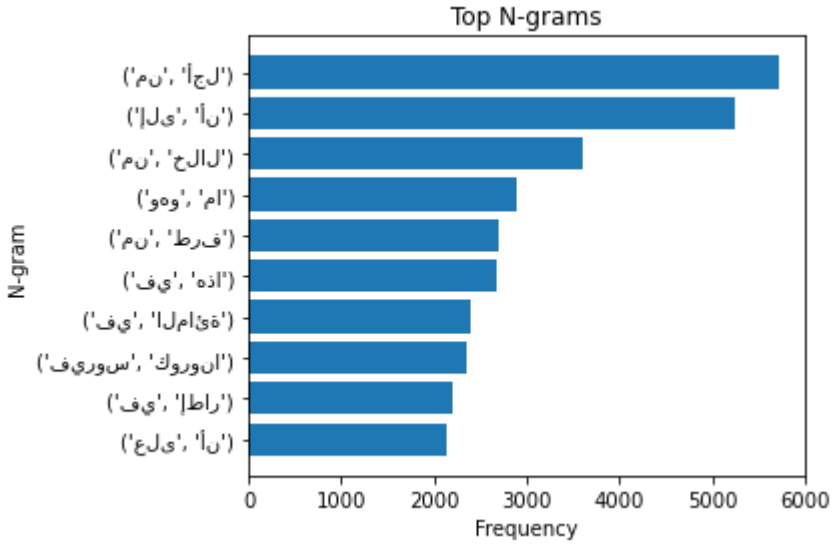
In [19]:
```
# Create a custom-sized figure
plt.figure(figsize=(10, 6))  # Change the values (width, height) as needed

# Plot number of examples for each class
class_counts = combined_data['topic'].value_counts()
plt.bar(class_counts.index, class_counts.values)
plt.xlabel('Topic')
plt.ylabel('Number of Examples')
plt.title('Number of Examples per Class')
plt.xticks(rotation=45)

# Display the plot
plt.show()
```



In [11]:
```python
def preprocess_text(text):
    # Tokenize the text into words
    tokens = word_tokenize(text)
     # Define a list of Arabic punctuation characters
    arabic_punctuation = ['``', '-' ,',','؛',"''",".",'،']
    # Combine Arabic and ASCII punctuation characters
    all_punctuation = arabic_punctuation + list(string.punctuation)
    # Remove unwanted punctuation marks
    cleaned_tokens = [token for token in tokens if token not in all_punctuation]

    return cleaned_tokens

def get_ngrams(text, n):
    cleaned_tokens = preprocess_text(text)
    return list(ngrams(cleaned_tokens, n))

def plot_top_ngrams(data, n, top_n=10):
    ngram_counter = Counter()

    for text in data:
        ngrams_list = get_ngrams(text, n)
        ngram_counter.update(ngrams_list)

    top_ngrams = ngram_counter.most_common(top_n)

    ngram_labels, ngram_counts = zip(*top_ngrams)
    print("ngram: ",ngram_labels)

    # Convert ngram_counts tuple to list
    ngram_counts = list(ngram_counts)

    # Plotting the n-grams
    plt.barh(range(len(ngram_labels)), ngram_counts)
    plt.yticks(range(len(ngram_labels)), ngram_labels)
    plt.xlabel('Frequency')
    plt.ylabel('N-gram')
    plt.title('Top N-grams')
    plt.gca().invert_yaxis()
    plt.tight_layout()
    plt.show()
```

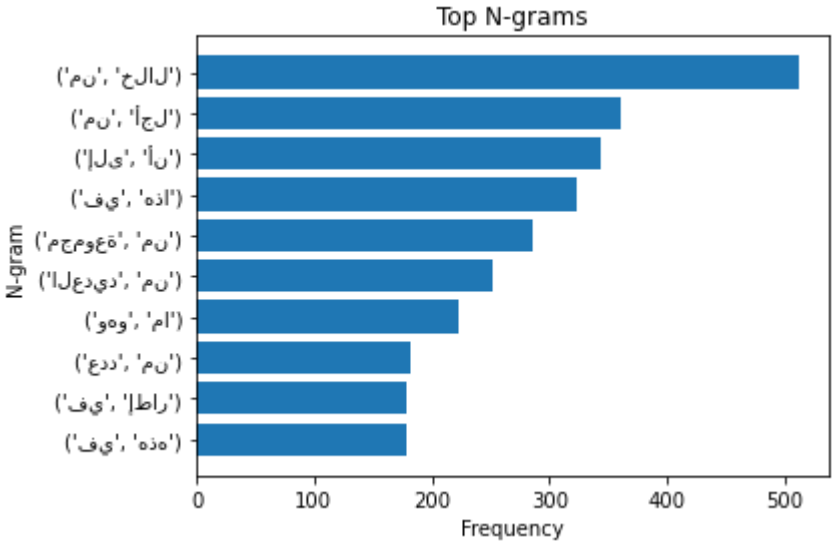
In [12]:
```python
# plot top frequent n-grams generally
n = 2
top_n = 10
print("-"*20,"top frequent n-grams generally for all classes", "-"*20)
print('*'*100)
plot_top_ngrams(combined_data['story'], n, top_n)
```

```
-------------------- top frequent n-grams generally for all classes --------------------
****************************************************************************************************
ngram:  (('أجل' ,'من'), ('أن' ,'إلى'), ('خلال' ,'من'), ('ما' ,'وهو'), ('طرف' ,'من'), ('هذا' ,'في'), ('المائة' ,'في'), ('كورونا' ,'فيروس'), ('إطار' ,'ي'), ('أن' ,'على'))
```



In [13]:
```
# plot top frequent n-grams for each class separately
n = 2
top_n = 10
for topic in combined_data['topic'].unique():
    print("-"*30,f"topic: {topic}", "-"*30)
    print('*'*100)
    plot_top_ngrams(combined_data[combined_data['topic'] == topic]['story'], n, top_n)
    print('*'*100)
```

```
------------------------------ topic: art-et-culture ------------------------------
****************************************************************************************************
ngram:  (('خلال' ,'من'), ('أجل' ,'من'), ('أن' ,'إلى'), ('هذا' ,'في'), ('من' ,'مجموعة'), ('من' ,'العديد'), ('ما' ,'وهو'), ('من' ,'عدد'), ('في' ,'إطار''), ('هذه' ,'في'))
```

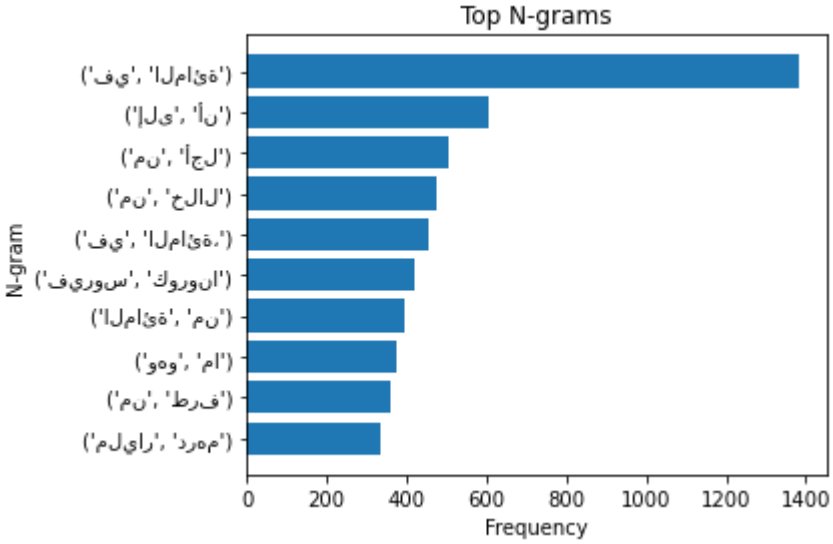


****

---------- topic: economie ----------

****

ngram: (('المائة' ,'في'), ('أن' ,'إلى'), ('أجل' ,'من'), ('خلال' ,'من'), ('المائة' ,'في،'), ('فيروس' ,'كورونا'), ('من' ,'المائة'), ('ما' ,'وهو'), ('م' ,'طرف' ,'ن'), ('درهم' ,'مليار'))

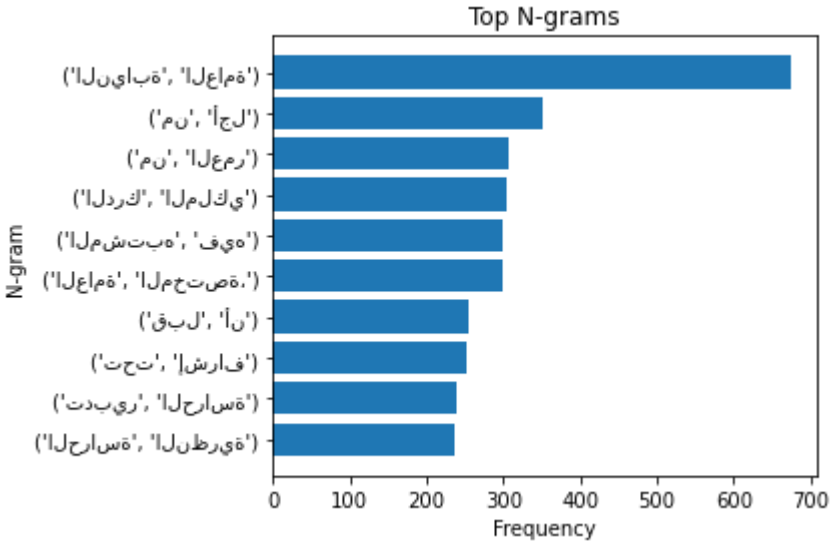


****

---------- topic: faits-divers ----------

****

ngram: (('العامة' ,'النيابة'), ('أجل' ,'من'), ('العمر' ,'من'), ('الملكي' ,'الدرك'), ('فيه' ,'المشتبه'), ('المختصة' ,'العامة،'), ('أن' ,'قبل'), ('إشر' ,'تحت' ,'اف'), ('الحراسة' ,'تدبير'), ('النظرية' ,'الحراسة'))

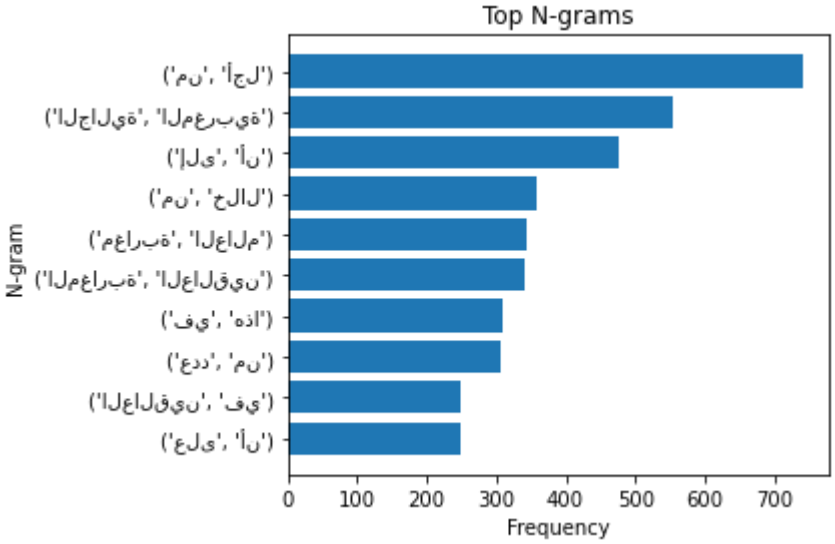


****

---------- topic: marocains-du-monde ----------

****

ngram: (('أجل' ,'من'), ('المغربية' ,'الجالية'), ('أن' ,'إلى'), ('خلال' ,'من'), ('العالم' ,'مغاربة'), ('العالقين' ,'المغاربة'), ('هذا' ,'في'), ('م' ,'عدد'), ('في' ,'العالقين'), ('أن' ,'على'))

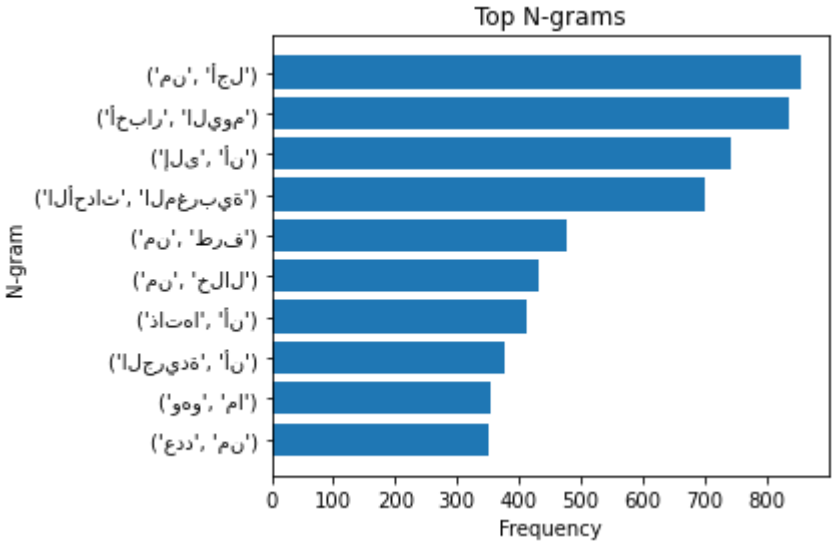


****

---------- topic: medias ----------

****

ngram: (('أجل' ,'من'), ('اليوم' ,'أخبار'), ('أن' ,'إلى'), ('المغربية' ,'الأحداث'), ('طرف' ,'من'), ('خلال' ,'من'), ('أن' ,'ذاتها'), ('أن' ,'الجريدة'), ('ما' ,'وهو'), ('من' ,'عدد'))

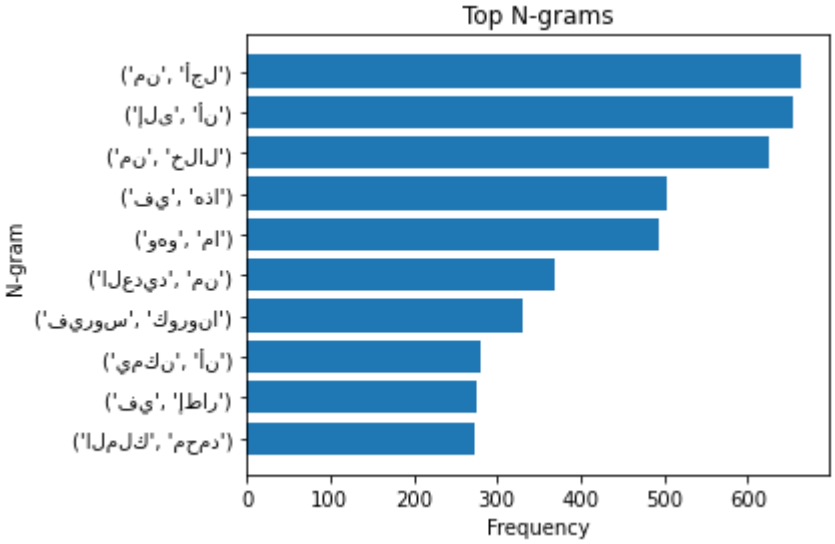


****

---------- topic: orbites ----------

****

ngram: (('أجل' ,'من'), ('أن' ,'إلى'), ('خلال' ,'من'), ('هذا' ,'في'), ('ما' ,'وهو'), ('من' ,'العديد'), ('كورونا' ,'فيروس'), ('أن' ,'يمكن'), ('ف' ,'إطار' ,'ي'), ('محمد' ,'الملك'))

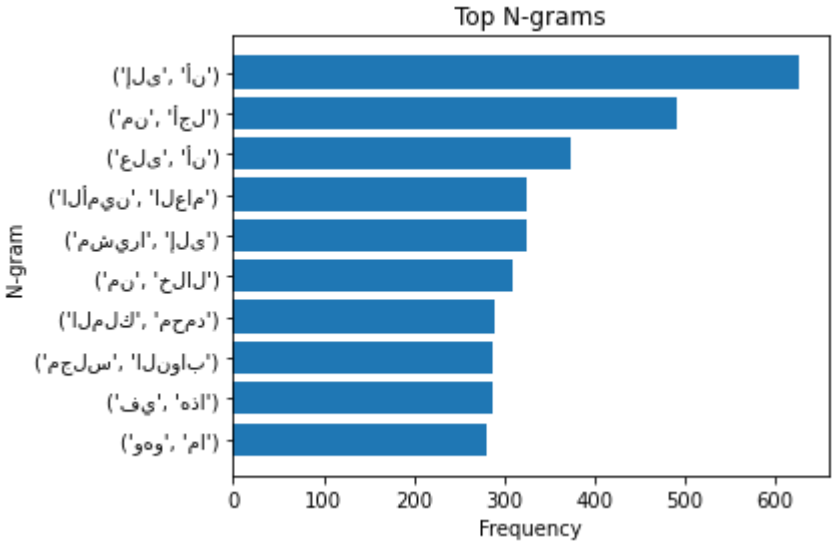


****

---------- topic: politique ----------

****

ngram: (('أن' ,'إلى'), ('أجل' ,'من'), ('أن' ,'على'), ('العام' ,'الأمين'), ('إلى' ,'مشيرا'), ('من' ,'خلال'), ('محمد' ,'الملك'), ('النواب' ,'مجلس'), ('هذا' ,'في'), ('ما' ,'وهو'))

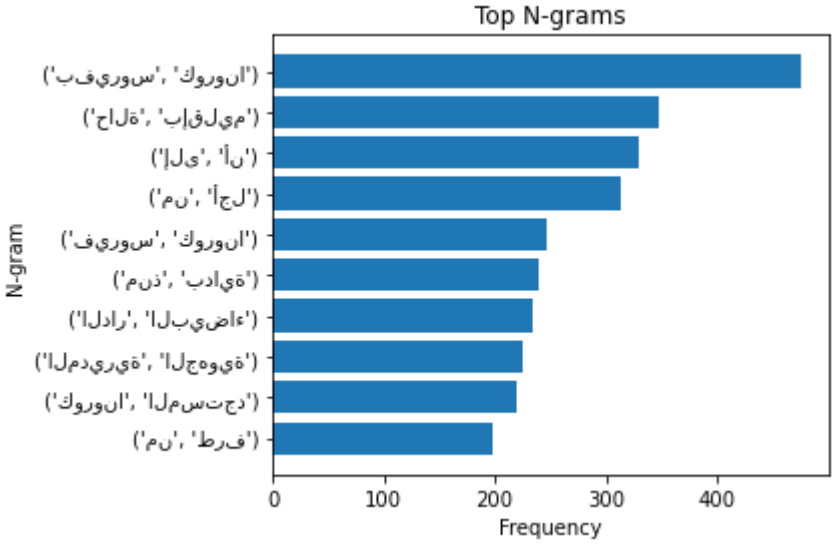


****

---------- topic: regions ----------

****

ngram: (('كورونا' ,'بفيروس'), ('بإقليم' ,'حالة'), ('أن' ,'إلى'), ('أجل' ,'من'), ('كورونا' ,'فيروس'), ('بداية' ,'منذ'), ('البيضاء' ,'الدار'), ('المديرية' ,'الجهوية''), ('المستجد' ,'كورونا'), ('طرف' ,'من'))

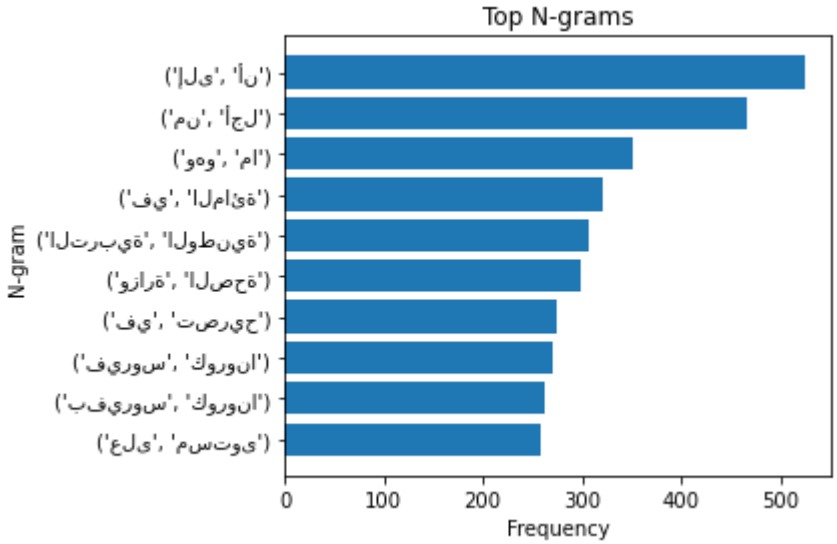


****

---------- topic: societe ----------

****

ngram: (('أن' ,'إلى'), ('أجل' ,'من'), ('ما' ,'وهو'), ('المائة' ,'في'), ('الوطنية' ,'التربية'), ('الصحة' ,'وزارة'), ('تصريح' ,'في'), ('كوروز' ,'فيروس'), ('كورونا' ,'بفيروس'), ('مستوى' ,'على'))

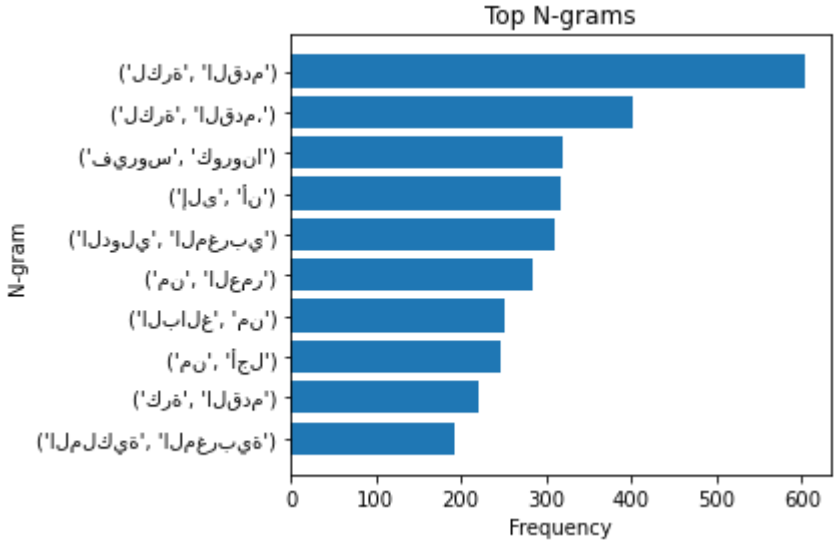


****

---------- topic: sport ----------

****

ngram: (('القدم' ,'لكرة'), ('القدم' ,'لكرة،'), ('كورونا' ,'فيروس'), ('أن' ,'إلى'), ('المغربي' ,'الدولي'), ('العمر' ,'من'), ('من' ,'البالغ'), ('أجل' ,'من'), ('كرة' ,'القدم'), ('المغربية' ,'الملكية'))

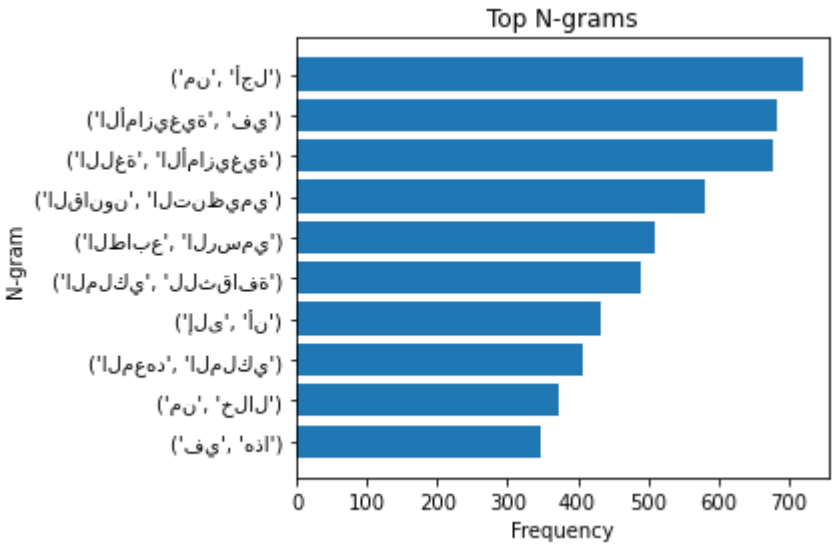


****

---------- topic: tamazight ----------

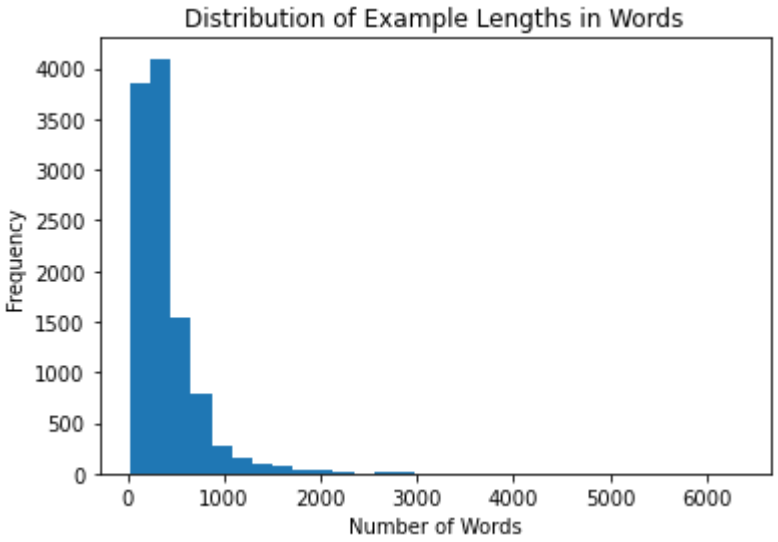
****

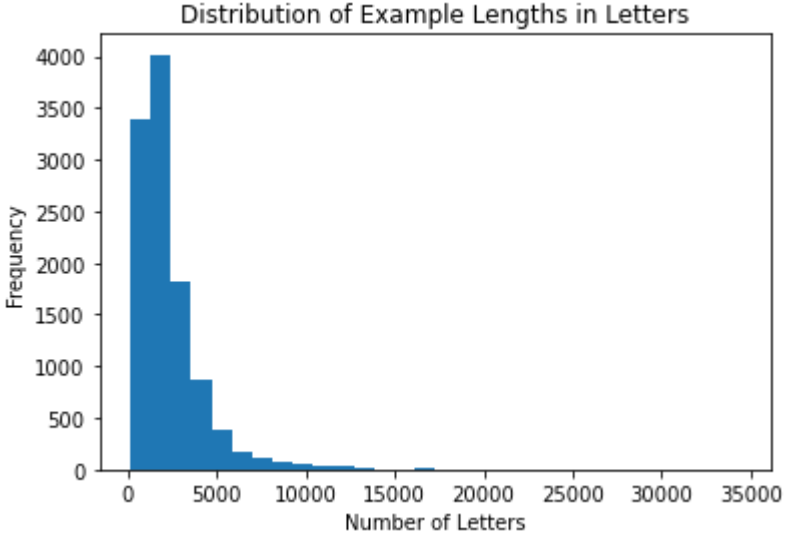
ngram: (('أجل' ,'من'), ('في' ,'الأمازيغية'), ('الأمازيغية' ,'اللغة'), ('التنظيمي' ,'القانون'), ('الرسمي' ,'الطابع'), ('للثقافة' ,'الملكي'), ('أن' ,'إلى'), ('المعهد' ,'الملكي'), ('خلال' ,'من'), ('هذا' ,'في'))



****

In [14]:
```
def plot_length_distribution(data, measure):
    if measure == 'words':
        lengths = data.apply(lambda x: len(word_tokenize(x)))
        plt.xlabel('Number of Words')
        plt.title('Distribution of Example Lengths in Words')
    elif measure == 'letters':
        lengths = data.apply(len)
        plt.xlabel('Number of Letters')
        plt.title('Distribution of Example Lengths in Letters')
    else:
        raise ValueError("Invalid measure. Use 'words' or 'letters'.")

    plt.hist(lengths, bins=30)
    plt.ylabel('Frequency')
    plt.show()

# Plot distribution of example lengths in words
plot_length_distribution(combined_data['story'], measure='words')

# Plot distribution of example lengths in letters
plot_length_distribution(combined_data['story'], measure='letters')
```


Distribution of Example Lengths in Words
Frequency
Number of Words
0
500
1000
1500
2000
2500
3000
3500
4000
0
1000
2000
3000
4000
5000
6000


Distribution of Example Lengths in Letters
Frequency
Number of Letters
0
500
1000
1500
2000
2500
3000
3500
4000
0
5000
10000
15000
20000
25000
30000
35000