[AI Team] assignment - WideBot' Summer Internship 2023

Task 2 - EDA:

- Using this dataset, you are supposed to prepare an Exploratory Data Analysis (EDA) report

in PDF format that shows at least 3 different insights about this data (number of examples per class, top frequent n-grams generally and per class, lengths of examples in words and letters, ...) . Feel free to show your insights in a good format (description, tables, charts, ...).

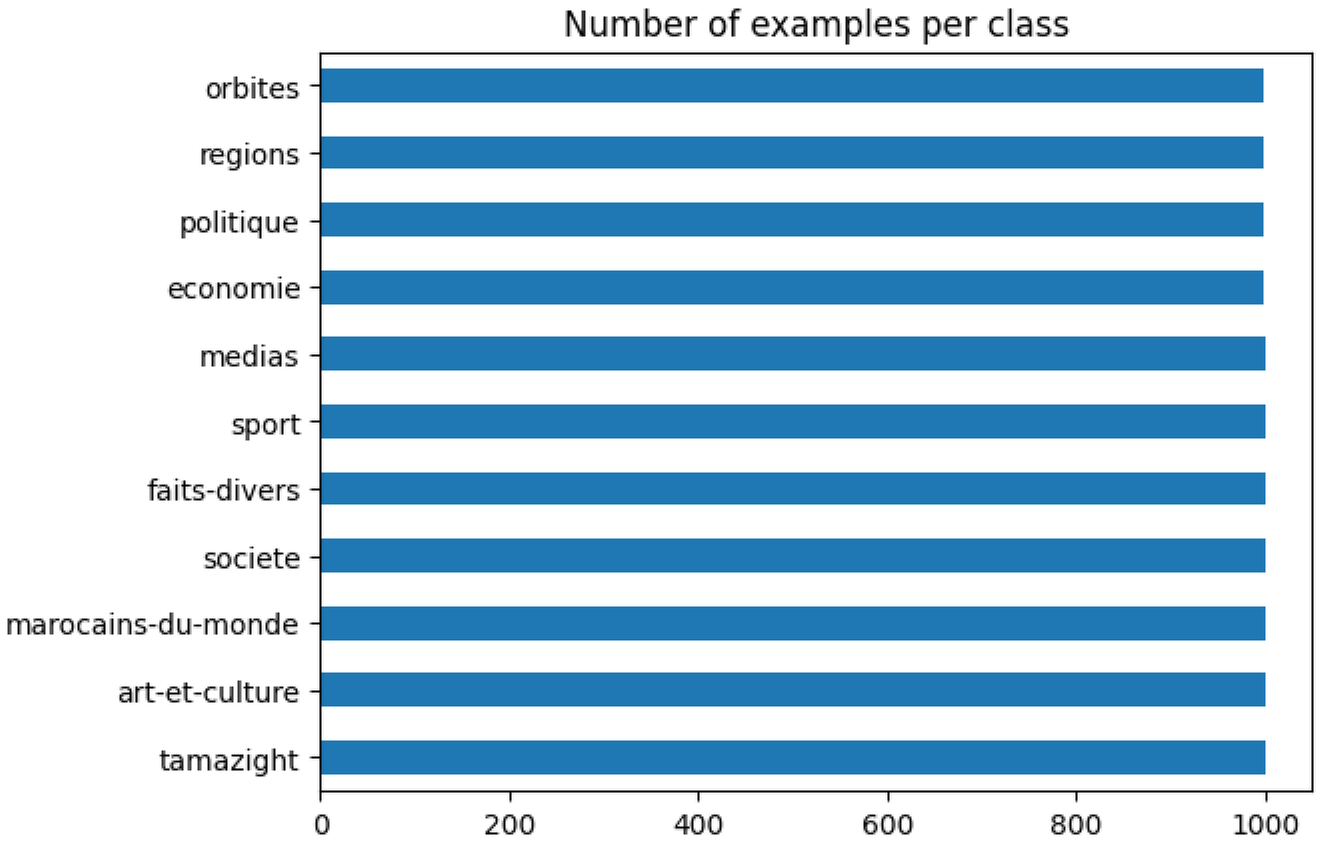
- Note that you can prepare the PDF manually after generating the analytics in any other format (CSV, xls, doc), ...).
- Note: Use stories data only (not comments).

number of examples per class

In [90]:

```
print(df['topic'].value_counts())
df['topic'].value_counts().plot(kind='barh')
plt.title('Number of examples per class')
plt.show()
```

```
tamazight             1000
art-et-culture        1000
marocains-du-monde    1000
societe               1000
faits-divers          1000
sport                 1000
medias                1000
economie               999
politique              999
regions                999
orbites                999
Name: topic, dtype: int64
```



lengths of examples in words and letters

In [40]:

```
# 1. Total words in the dataframe
total_words = df['word_count'].sum()
print("Words count in the dataframe:", total_words)

# 2. Total letters in the dataframe
total_letters = df['letter_count'].sum()
print("Letters count in the dataframe:", total_letters)

# 3. Total words by each class
words_by_class = df.groupby('topic')['word_count'].sum()
print("\nWords count per class:")
print(words_by_class)

# 4. Total letters by each class
letters_by_class = df.groupby('topic')['letter_count'].sum()
print("\nLetters count per class:")
print(letters_by_class)
```

```
Words count in the dataframe: 3193442
Letters count in the dataframe: 18176260

Words count per class:
topic
art-et-culture        335340
economie              265836
faits-divers          121037
marocains-du-monde    293131
medias                430330
orbites               496053
politique             267358
regions               179464
societe               259165
sport                 180461
tamazight             365267
Name: word_count, dtype: int64

Letters count per class:
topic
art-et-culture        1868525
economie              1533354
faits-divers           673237
marocains-du-monde    1686830
medias                2437992
orbites               2805323
politique             1548699
regions               1025435
societe               1488042
sport                  993393
tamazight             2115430
Name: letter_count, dtype: int64
```

top frequent n-grams generally and per class

In [35]:

```
# Get the most frequent n-gram generally
# Concatenate the most_frequent_2gram and most_frequent_3gram columns into a single column
df['most_frequent_ngram'] = df['most_frequent_2gram'] + ', ' + df['most_frequent_3gram']
# Split the n-gram into a list
ngrams_list = df['most_frequent_ngram'].str.split(', ')
# Count the frequency of each n-gram
ngram_counts = Counter([tuple(ngram) for ngram in ngrams_list])
# Get the most frequent n-gram
most_frequent_ngram = ' , '.join(ngram_counts.most_common(1)[0][0])
# Print the most frequent n-gram
print("Most frequent (2,3)n-gram:", most_frequent_ngram)
```

```python
# Get the most frequent n-gram by class
# Group the dataframe by class
grouped = df.groupby('topic')
# Get the most frequent 2-gram for each class
most_frequent_2gram_by_class = grouped.apply(lambda x: get_most_frequent_ngram(x.iloc[0], 2))
# Get the most frequent 3-gram for each class
most_frequent_3gram_by_class = grouped.apply(lambda x: get_most_frequent_ngram(x.iloc[0], 3))
# Combine the most frequent 2-gram and 3-gram for each class into a single column
most_frequent_ngram_by_class = most_frequent_2gram_by_class + ' , ' + most_frequent_3gram_by_class
# Print the most frequent n-gram for each class
print("\nMost frequent (2,3)n-gram by class:")
for c, ngram in most_frequent_ngram_by_class.items():
    print(f"Class {c}: {ngram}")
```

```
Most frequent (2,3)n-gram: وزارة الصحة , كشفت وزارة الصحة

Most frequent (2,3)n-gram by class:
Class art-et-culture: رشيد شباري , فئة تلاميذ الثانوي
Class economie: الميزان التجاري , ستشرع الحكومة تطبيق
Class faits-divers: ضاية افرط , ضاية افرط النجوم
Class marocains-du-monde: محمد السادس , بمناسبة الذكري الواحدة
Class medias: الاحداث المغربية , العام للامم المتحدة
Class orbites: الارهاب خلال , الديمقراطيات الغربية اهتزت
Class politique: محمد الخامس , الملك محمد السادس
Class regions: المجمع التربوي , المجمع التربوي غوستاف
Class societe: الدخول المدرسي , الدخول المدرسي المقبل
Class sport: يونس عبد , يونس عبد الحميد
Class tamazight: المسيرة الاحتجاجية , احتجاجا اقصاء المبدعين
```