

# HESPRESS STORIES

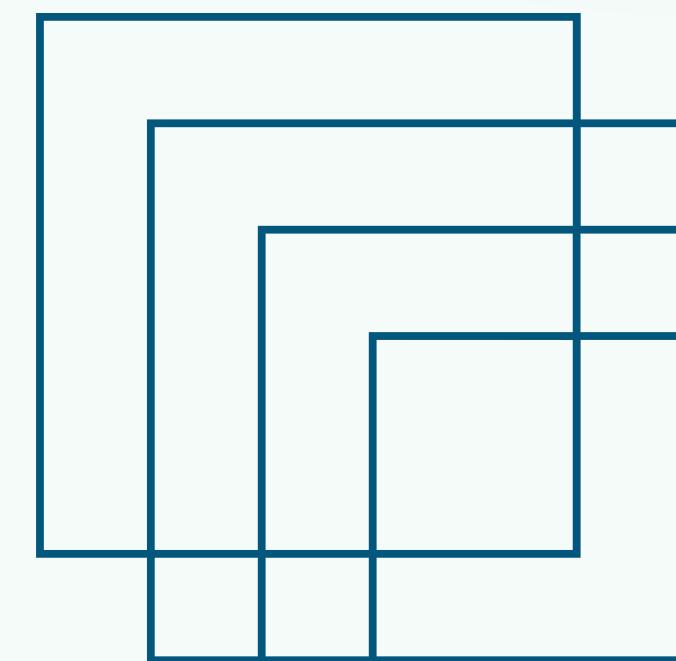


## EXPLORATORY DATA ANALYSIS REPORT

BY SOHAILA DIAB

# AGENDA

- 01 BUSINESS QUESTION
- 02 ABOUT THE DATA
- 03 FREQUENCY OF TOPICS
- 04 MOST FREQUENT WORDS
- 05 TOPICS WITH HIGHEST/LOWEST AVG WORD COUNT
- 06 MOST FREQUENT N-GRAMS PER TOPIC



# BUSINESS QUESTION

---

**Given news from the  
Morrocan online news  
website, Hespress,  
what is the topic of  
that story?**



# ABOUT THE DATA

---

The dataset consisted of  
**11,000 examples (stories)**  
and **6 columns:**  
**id, title, date, author, story, topic**

... and no missing data



# TRANSLATION OF TOPICS

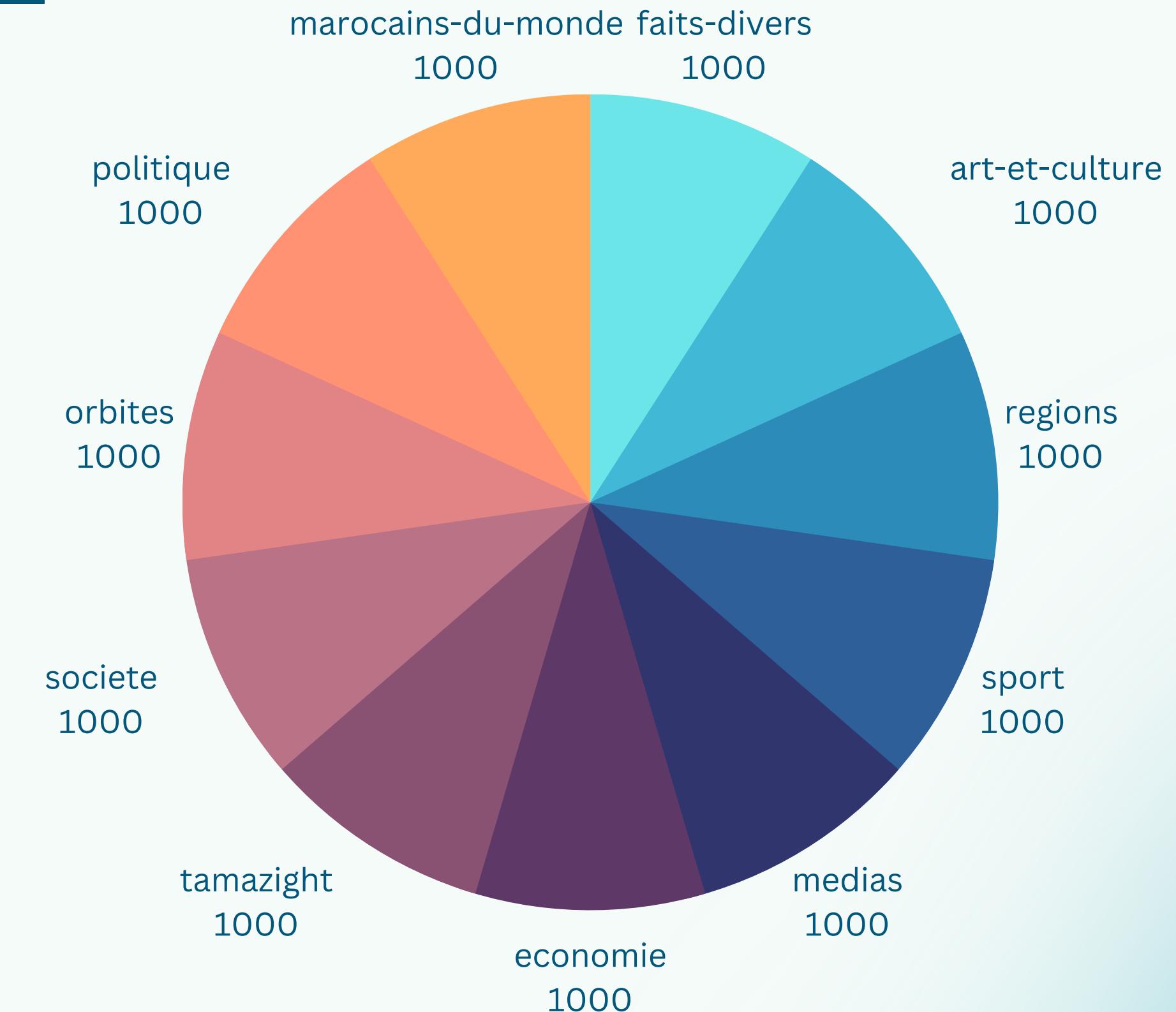
---

Since French  
is widely  
spoken in  
Morocco, the  
topics are in  
French

- **faits divers:** miscellaneous facts
- **art et culture:** art and culture
- **regions:** regions
- **sport:** sport
- **medias:** media
- **economie:** economy
- **tamazight:** tamazight (the language of the Amazigh people)
- **societe:** community/company
- **orbites:** orbits
- **politique:** policy
- **marocains du monde:** moroccans-of-the-world

# FREQUENCY OF TOPICS

The dataset is  
balanced, with  
1000 stories  
for each topic



# MOST FREQUENT WORDS PER TOPIC

# faits divers

# art et culture

# الحاجة دعم من الوزاراة الأولى المغربية

# regions

# sport

# كرة القدم الجديدة المباريات

# medias

# economie

# المجلد الرابع - المعاشرة الضريبية

# tamazight

# societe

# orbites

# politique

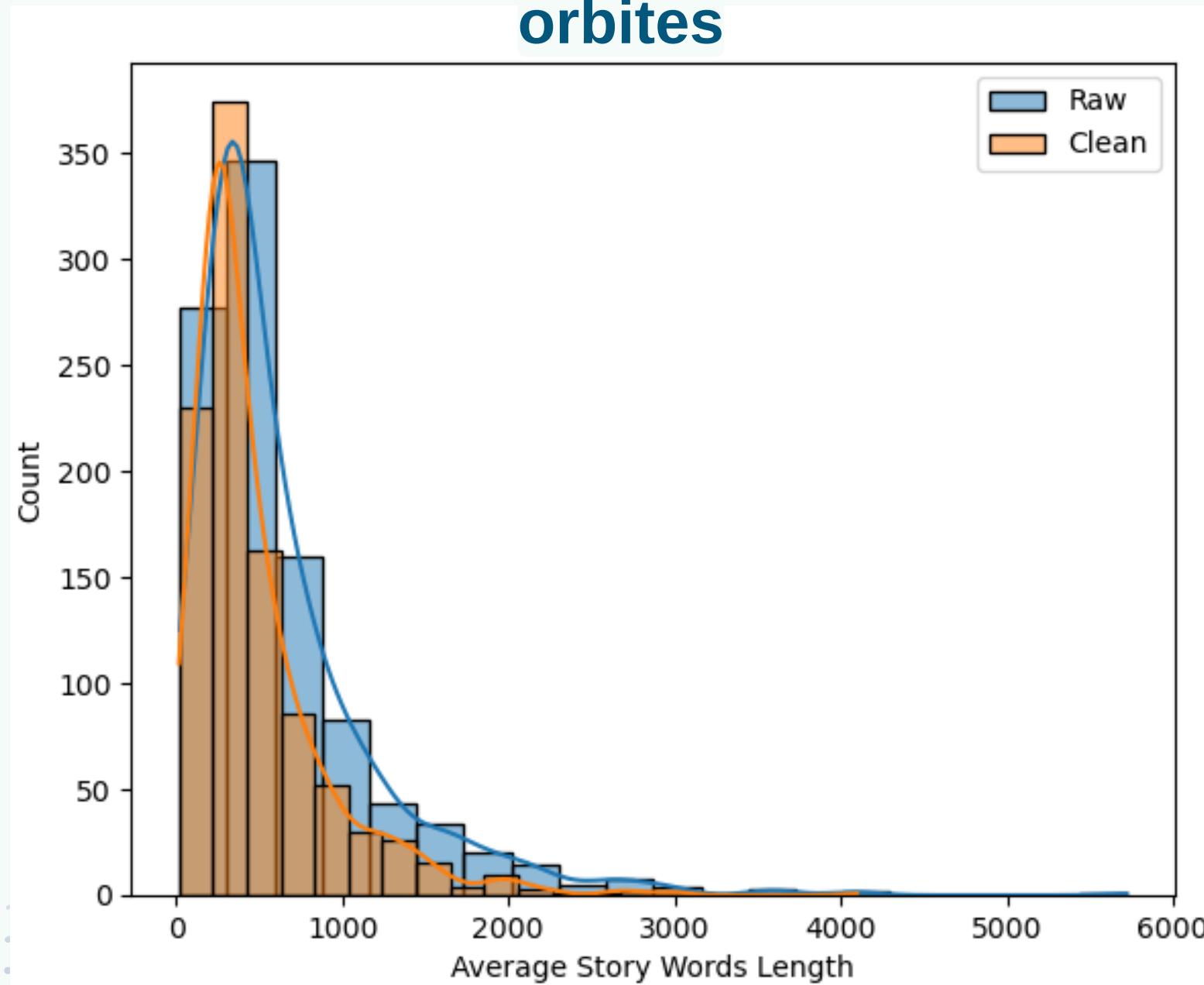
# المجلسين

# marocains du monde

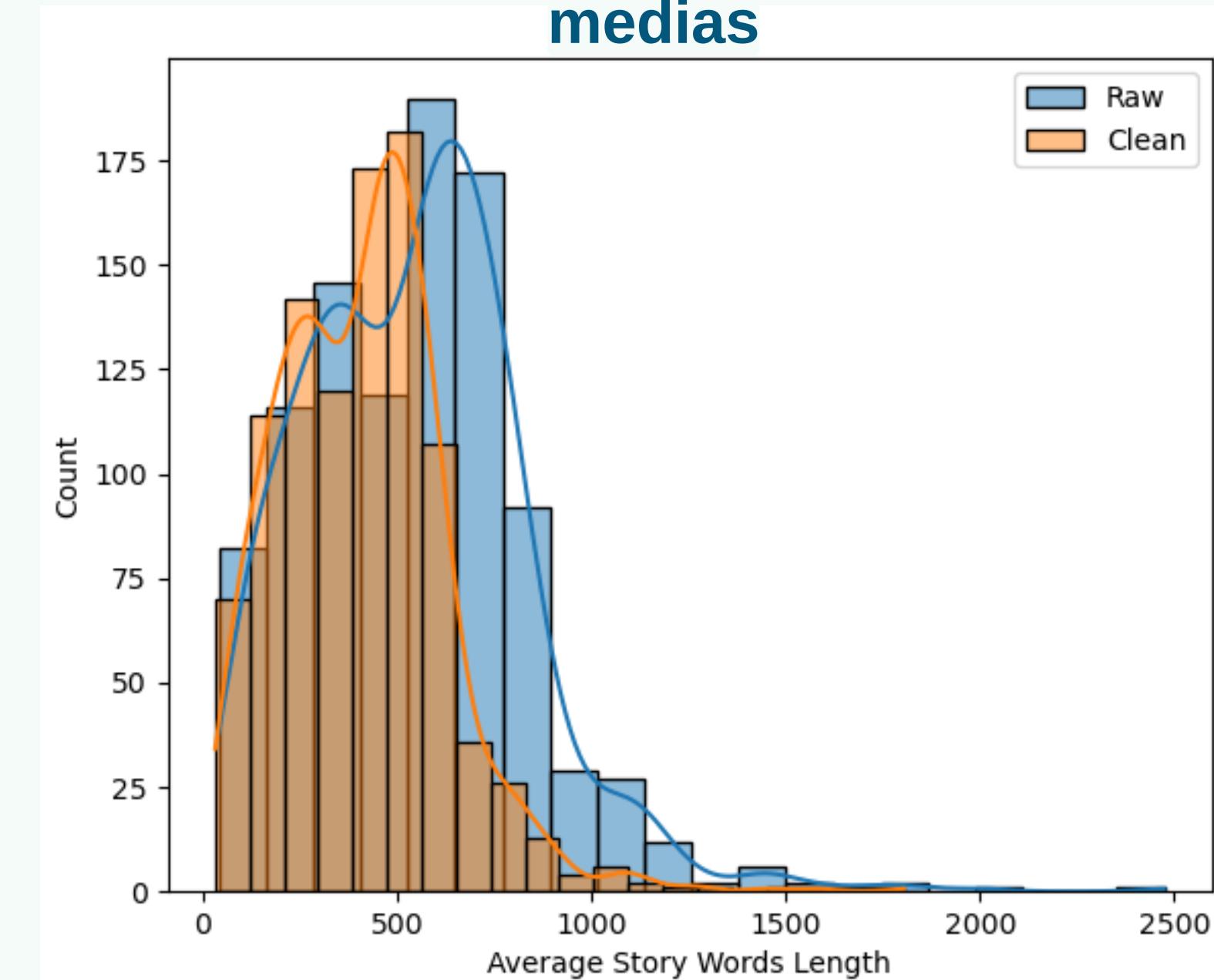
# المغرب

# TOPICS WITH HIGHEST AVG WORD COUNT

orbites



medias



**Raw:** 643 words on avg.

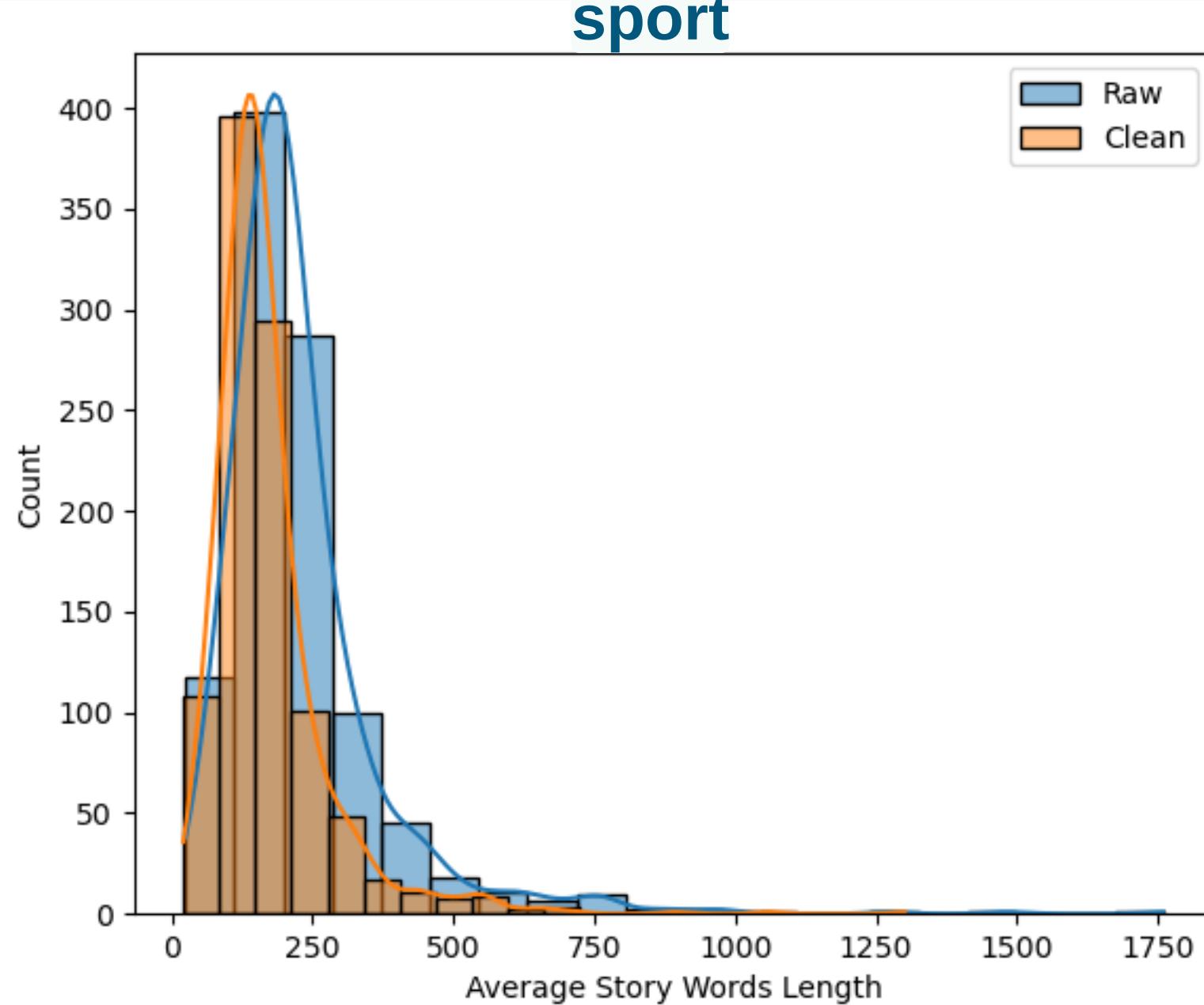
**Clean:** 482 words on avg.

**Raw:** 545 words on avg.

**Clean:** 412 words on avg.

# TOPICS WITH LOWEST AVG WORD COUNT

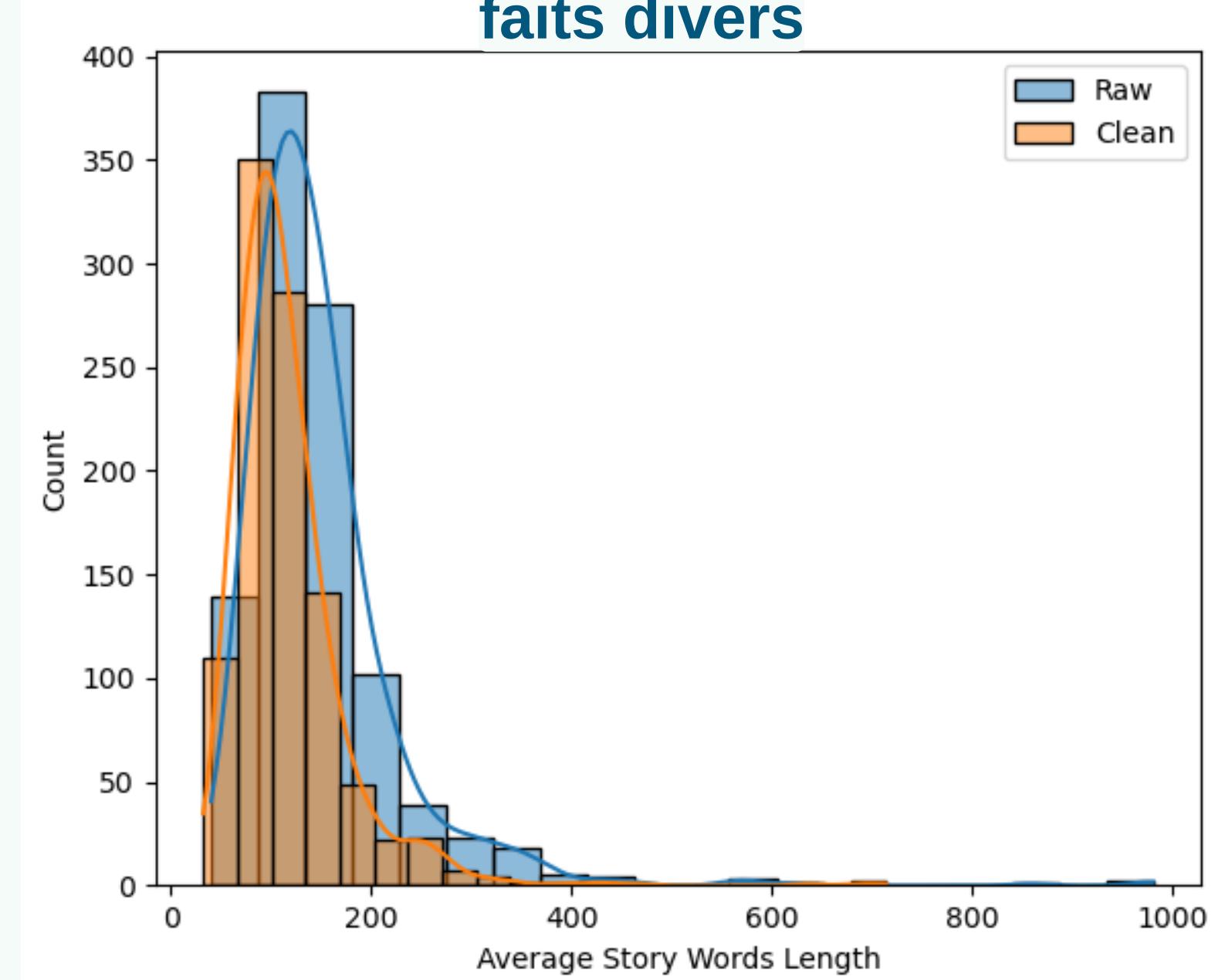
sport



Raw: 227 words on avg.

Clean: 172 words on avg.

faits divers



Raw: 149 words on avg.

Clean: 117 words on avg.

# TOP 5 UNIGRAMS PER TOPIC

## faits-divers:

- العامة.1
- عناصر.2
- النيابة.3
- المشتبه.4
- القضائية.5

## art-et-culture:

- المغربي.1
- المغربية.2
- أن.3
- المغرب.4
- محمد.5

## regions:

- حالة.1
- كورونا.2
- حالات.3
- إقليم.4
- السلطات.5

## sport:

- القدم.1
- الفريق.2
- المغربي.3
- لكرة.4
- الموسم.5

## medias:

- المغربية.1
- المغرب.2
- المساء.3
- الجريدة.4
- ذاته.5

## economie:

- المائة.1
- المغرب.2
- كورونا.3
- الحكومة.4
- القطاع.5

## tamazight:

- الأمازيغية.1
- الأمازيغي.2
- اللغة.3
- المغرب.4
- المغربية.5

## societe:

- حالة.1
- كورونا.2
- الصحة.3
- وزارة.4
- التعليم.5

## orbites:

- الله.1
- المغرب.2
- كورونا.3
- المغربية.4
- العالم.5

## politique:

- الحكومة.1
- المغرب.2
- رئيس.3
- المغربية.4
- السياسي.5

## marocains-du-monde:

- المغربية.1
- المغاربة.2
- المغرب.3
- الجالية.4
- المغربي.5

# TOP 5 BIGRAMS PER TOPIC

## faits-divers:

- النهاية, العامة.
- العامة, المختصة.
- الحراسة, النظرية.
- الدرك, الملكي.
- تدبير, الحراسة.

## art-et-culture:

- وزارة, الثقافة.
- الثقافة, والشباب.
- والشباب, والرياضة.
- فيروس, كورونا.
- التواصل, الاجتماعي.

## regions:

- فيروس, كورونا.
- كورونا, المستجد.
- حالة, بإقليم.
- دار, البيضاء.
- فيروس, كورونا.

## sport:

- كرة, القدم.
- فيروس, كورونا.
- الدولي, المغربي.
- كرة, القدم.
- البالغ, العمر.

## medias:

- الأحداث, المغربية.
- المنبر, ذاته.
- النهاية, العامة.
- فيروس, كورونا.
- محمد, السادس.

## economie:

- فيروس, كورونا.
- الحجر, الصحي.
- كورونا, المستجد.
- جائحة, كورونا.
- السنة, الجارية.

## tamazight:

- اللغة, الأمازيغية.
- القانون, التنظيمي.
- للتقالفة, الأمازيغية.
- التابع, الرسمي.
- الملكي, للتقالفة.

## societe:

- وزارة, الصحة.
- التربية, الوطنية.
- فيروس, كورونا.
- فيروس, كورونا.
- الحجر, الصحي.

## orbites:

- فيروس, كورونا.
- محمد, السادس.
- الولايات, المتحدة.
- الملك, محمد.
- الحجر, الصحي.

## politique:

- رئيس, الحكومة.
- مجلس, النواب.
- العدالة, والتنمية.
- الأمين, العام.
- محمد, السادس.

## marocains-du-monde:

- الجالية, المغربية.
- مغاربة, العالم.
- المقيمين, بالخارج.
- المغاربة, العالقين.
- أفراد, الجالية.

We can see that COVID is very popular amongst most topics

# TOP 5 TRIGRAMS PER TOPIC

## faits-divers:

- النيابة, العامة, المختصة.
- تدبير, الحراسة, النظرية.
- إشراف, النيابة, العامة.
- رهن, إشارة, البحث.
- الحراسة, النظرية, رهن.

## art-et-culture:

- الثقافة, والشباب, والرياضة.
- (وزارة, الثقافة, والشباب).
- المركز, السينمائي, المغربي.
- جريدة, هسبريس, الإلكترونية.
- تصريح, جريدة, هسبريس.

## regions:

- بفيروس, كورونا, المستجد.
- المديرية, الجهوية, للصحة.
- جديدة, بفيروس, كورونا.
- فيروس, كورونا, المستجد.
- بني, ملال, خنيفرة.

## sport:

- المملكة, المغربية, لكرة.
- المغربية, لكرة, القدم.
- الجامعة, الملكية, المغربية.
- فيروس, كورونا, المستجد.
- الدفاع, الحسني, الجديدي.

## medias:

- الملك, محمد, السادس.
- (حالة, الطوارئ, الصحية).
- (قراءة, رصيف, صحفة).
- المجلس, الوطني, للصحافة).
- سعد, الدين, العثماني.

## economie:

- فيروس, كورونا, المستجد.
- والمالية, وإصلاح, الإدارة.
- الاقتصاد, والمالية, وإصلاح.
- العام, لمقاولات, المغرب.
- أزمة, فيروس, كورونا.

## tamazight:

- الملكي, للثقافة, الأمازيغية.
- التابع, الرسمي, للأمازيغية.
- المعهد, الملكي, للثقافة.
- مشروع, القانون, التنظيمي.
- تفعيل, التابع, الرسمي.

## societe:

- وزارة, التربية, الوطنية.
- جريدة, هسبريس, الإلكترونية.
- بفيروس, كورونا, المستجد.
- تصريح, جريدة, هسبريس.
- العالى, والبحث, العلمي.

## politique:

- الملك, محمد, السادس.
- سعد, الدين, العثماني.
- حزب, العدالة, والتنمية.
- الأمين, العام, لحزب.
- التجمع, الوطني, للأحرار.

## orbites:

- الملك, محمد, السادس.
- صلى, الله, وسلم.
- فيروس, كورونا, المستجد.
- الولايات, المتحدة, الأمريكية.
- النبي, صلى, الله.

## marocains-du-monde:

- أفراد, الجالية, المغربية.
- الجالية, المغربية, المقيمة.
- بالمغاربة, المقيمين, بالخارج.
- الملك, محمد, السادس.
- المكلفة, بالمغاربة, المقيمين.

Trigrams in this example are more meaningful than unigrams and bigrams, and can give an idea about what topic it is

