Exploratory Data Analysis for Arabic News Stories dataset

By : Mohamed Essam Khalil Ibrahim

Email: mohamediki31@gmail.com // mohamedissam@ieee.org

Abstract:

Given of 11 Csv files of 11 news categories, each consists of 1000 record with columns [‘is’,‘title’, ’date’, ’author’, ‘story’, ‘topic’].

So the hole data consests of 11000 record before cleaning and 10998 after cleaning.

An Exploratory Data Analysis for Arabic has been done to get some calculations and relations between variables.

That done by fries cleaning the data and dropping unneeded columns and dropping the duplicates then adding some columns like [“Story_length”, “Story_words_number” , “title_length”, “title_words_number”] to apply examples length calculations and get some relations between titles and stories , also n-gram code applied to get top frequent n-grams for the hole data and for each class too.

Data example:

Data example for one topic (class)

Unnamed: 0	id	title	date	author	story	topic
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 والشباب والرياضة رسالة موسومة بـ"لماذا تحولت أحلام بي...	art-et-culture
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 بجائحة "كورونا"، أعلن مهرجان الرباط الدولي لسينما ال...	art-et-culture
2	f2d282a4054e11eb800f646e69d991ea	فيلم "بدون عنف" لهشام العسري ..."كعب الحذاء وواقع مؤلم للنساء	الجمعة 02 أكتوبر 2020 - 04:00	*عفيفة الحسينات	تشير مشاهدة فيلم قصير ضمن الثلاثية الأخيرة للمخرج المغربي هشام العسري إلى جملة من المرجعيات الثق...	art-et-culture
3	f3f46cac054e11eba403646e69d991ea	تنين ووهان" .. مريم أيت أحمد توقّع أولى" روايات الجائحة" بالمغرب	الجمعة 02 أكتوبر 2020 - 02:00	حاوزها: وائل بورشاشن	مِن قلب أيّام "الحَجْر"، رأتِ النّورَ الفصول الأولى مِن رواية مغربيّة تستلهم أحداثها مِن الجائ...	art-et-culture
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 الدّعم المخصّص لمشروعه الفني، المعلَن عنه في لائحة ...	art-et-culture
...	...	...	...	...	...	...
995	97e7b078055311eb972a646e69d991ea	مهَنّيّون: غياب توقيع رئيس الحكومة يوقف 200 مشروع سينمائي	الاثنين 18 نونبر 2019 - 00:25	هسبريس - وائل بورشاشن	تستمرّ الآثار الجانبية للانتقال الحكومي الأخير في الظّهور، هذه المرّة في القطاع السينمائي، بعد م...	art-et-culture
996	98f7723e055311ebb811646e69d991ea	ندوة تقارب "جهود السوسيين" في خدمة العربية	الأحد 17 نونبر 2019 - 11:15	الحسين حرّان	قال الدكتور المهدي السعيدي، في ندوة حول "جهود السوسيين في خدمة اللغة العربية وعلومها"، إن "أهمية ...	art-et-culture
997	9a29bc06055311ebbb05646e69d991ea	تتويج أشرطة في مهرجان سينما الذاكرة المشتركة	السبت 16 نونبر 2019 - 20:17	هسبريس من الرباط	أجمعت لجنة المسابقة الخاصة بالأفلام المغربية التي تناولت حقبة سنوات الرصاص، ضمن الدورة الثامنة ل...	art-et-culture
998	9b547968055311ebb870646e69d991ea	ريشة التشكيلية بثينة أزمي تتمرد على الظلم والعبودية تجاه النساء	السبت 16 نونبر 2019 - 10:00	هسبريس - كاميليا كريم	ألوان حيّة ولمسات تعبيرية تحْمل مآسي إنسانية لرصد تيمة النساء والعبودية، انسجمت في معرض تشكيلي ج...	art-et-culture
999	9c6b8f50055311eb9c5c646e69d991ea	مسرحية "بوحجبان" تبدأ جولة وطنية من مدينة خريبكة	السبت 16 نونبر 2019 - 04:20	هسبريس من الرباط	تقوم فرقة مسرح "سفر" بجولة وطنية لتقديم مسرحيتها "بوحجبان"، وهي من إخراج عزيز الخلوفي وتشخيص كل ...	art-et-culture

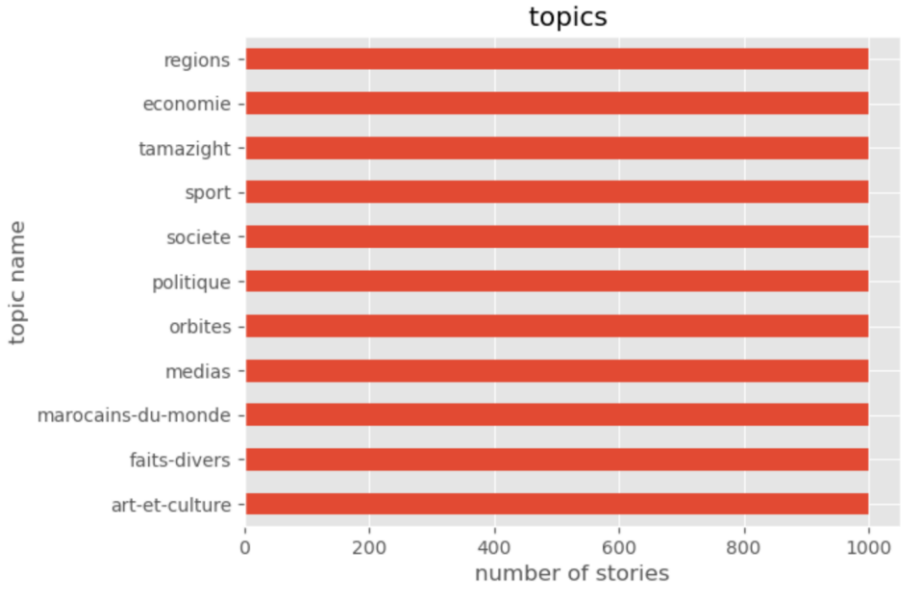
Data example for the hole data

	title	date	author	story	topic	story_length	title_length	story_word_count	title_word_count
0	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 والشباب والرياضة رسالة موسومة بـ"لماذا تحولت أحلام بي...	art-et-culture	3868	46	622	8
1	مهرجان "سينما المؤلّف" يستحضر روح ثريا جبران	الجمعة 02 أكتوبر 2020 - 07:26	هسبريس من الرباط	في ظلّ استمرار حالة الطوارئ الصحية المرتبطة بجائحة "كورونا"، أعلن مهرجان الرباط الدولي لسينما ال...	art-et-culture	2464	44	381	7
2	فيلم "بدون عنف" لهشام العسري "..كعب الحذاء وواقع مؤلم للنساء	الجمعة 02 أكتوبر 2020 - 04:00	*عفيفة الحسينات	تشير مشاهدة فيلم قصير ضمن الثلاثية الأخيرة للمخرج المغربي هشام العسري إلى جملة من المرجعيات التي...	art-et-culture	3056	61	516	10
3	تنين ووهان" .. مريم أيت أحمد توقّع أولى "روايات الجائحة" بالمغرب	الجمعة 02 أكتوبر 2020 - 02:00	حاورها: وائل بورشاشن	مِن قلب أيّام "الحَجْر"، رأتِ النّور الفصول الأولى مِن رواية مغربيّة تستلهم أحداثها مِن الجائ...	art-et-culture	4921	66	771	11
4	مسكر يتخلّى عن دعم "الوزارة" بسبب "الجمهور"	الخميس 01 أكتوبر 2020 - 19:40	هسبريس من الرباط	أعلن الفنان المغربيّ سعيد مسكر تخليه عن مبلغ الدّعم المخصّص لمشروعه الفني، المعلَن عنه في لائحة ...	art-et-culture	1238	43	200	7
...	...	...	...	...	...	...	...	...	...
10993	نشطاء أمازيغ يدافعون عن "الحقّ" في استقبال إسرائيليّين بالمغرب	الثلاثاء 10 شتنبر 2013 - 02:00	هسبريس ـ ميمون أم العيد	دافع ناشطون أمازيغ استقبلوا أخيرا وفدا من الطلبة الباحثين الإسرائيليين عن حقهم فيما قاموا به، مب...	tamazight	7773	62	1255	9
10994	شاعرة أمازيغية تعتصم بالمطار لرفض استمارتها بـ"تيفيناغ"	الاثنين 09 شتنبر 2013 - 08:20	هسبريس ـ عبد المغيث جبران	خاضت الشاعرة الأمازيغية ملكية مزان اعتصاما لمدة ست ساعات، قبل أيام قليلة، في مطار محمد الخامس با...	tamazight	1486	55	241	7
10995	وفد إسرائيلي يزور المغرب ويلتقي نشطاء أمازيغ بعدد من المدن	الثلاثاء 03 شتنبر 2013 - 16:24	هسبريس - ماجدة أيت لكتاوي	أدانت المنسقية الوطنية للمبادرة الطلابية ضد التطبيع والعدوان "السماح لوفد صهيوني مكون من طلبة وأ...	tamazight	2113	58	328	10
10996	نقاش أمازيغيّ مؤنّث بطنجة يذكّر بكّون الحقوق تُنتزع ولا تُعطى	السبت 17 غشت 2013 - 10:30	هسبريس من طنجة	طالبت الناشطة الأمازيغية مريم الدمناتي بضرورة فتح الحدود المغربية الجزائرية، مؤكدة أن نضال الأما...	tamazight	3612	63	597	10
10997	أمازيغ يقتحمون مقر البرلمان الليبي "مطالبين بـ"دسترة حقوقهم	الثلاثاء 13 غشت 2013 - 21:00	محمد الناجم من *طرابلس	اقتحم المئات من المتظاهرين المنحدرين من الأقليات الأمازيغية بليبيا، اليوم الثلاثاء، مبنى المؤتمر...	tamazight	1985	59	287	8

10998 rows × 9 columns

Data analysis:

First insight is getting number of examples per class , and as the collector collected 1000 example per class , after cleaning still close to 1000 each.



n-grams:

Second insight is getting top frequent n-grams generally for the hole data.

By applying unigram method, a Prepositions and pronouns are appeared as shown.

Word sum

```
[(('في',), 128055),
 (('من',), 103058),
 (('على',), 60243),
 (('أن',), 54954),
 (('إلى',), 53530),
 (('التي',), 32928),
 (('عن',), 25071),
 (('ما',), 19151),
 (('الذي',), 18975),
 (('مع',), 17688),
 (('هذا',), 15253),
 (('هذه',), 13964),
 (('لا',), 12216),
 (('أو',), 11766),
 (('خلال',), 10800),
 (('بين',), 10734),
 (('بعد',), 10146),
 (('المغرب',), 7711),
 (('كل',), 7522),
 (('كان',), 7264)]
```

So that step not helping so much, so unigram is not recommended in such data.

But it can help by ignoring it when dealing with the data

By applying the 4-grams method, that appears as shown.

```
[(('هسبريس', 'لجريدة', 'تصريح', 'في'), 614),
 (('الإلكترونية', 'هسبريس', 'لجريدة', 'تصريح،'), 529),
 (('أن', 'الإلكترونية،', 'هسبريس', 'لجريدة'), 381),
 (('والمغاربة', 'الإفريقي', 'والتعاون', 'الخارجية'), 247),
 (('الإفريقي', 'والتعاون', 'الخارجية', 'الشؤون'), 245),
 (('العامة', 'النيابة', 'إشراف', 'تحت'), 244),
 (('المقيمين', 'والمغاربة', 'الإفريقي', 'والتعاون'), 243),
 (('أن', 'لهسبريس،', 'تصريح', 'في'), 225),
 (('والتعاون', 'الخارجية', 'الشؤون', 'وزير'), 222),
 (('المهني', 'والتكوين', 'الوطنية', 'التربية'), 216),
 (('الأمازيغية', 'للثقافة', 'الملكي', 'المعهد'), 214),
 (('والتعليم', 'المهني', 'والتكوين', 'الوطنية'), 197),
 (('العالي', 'والتعليم', 'المهني', 'والتكوين'), 196),
 (('والبحث', 'العالي', 'والتعليم', 'المهني'), 190),
 (('بالخارج', 'المقيمين', 'والمغاربة', 'الإفريقي،'), 161),
 (('الماضية', 'ساعة', 'الـ24', 'خلال،'), 156),
 (('الأمازيغية', 'للثقافة', 'الملكي', 'المعهد،'), 151),
 (('المختصة', 'العامة', 'النيابة', 'إشراف،'), 151),
 (('العثماني', 'الدين', 'سعد', 'الحكومة،'), 146),
 (('أن', 'بها', 'ورد', 'التي'), 135)]
```

So that leads to the sources of the news, data and some commonly used sentences in Arabic news

That will be helpful as those sentences could be ignored when applying classification.

getting top frequent n-grams generally for some classes after dropping prepositions and pronouns

first class is art and by applying unigram method we get most common words used in art topics as shown in the table.

word	Word count
المغربي	852
محمد	793
الثقافة	532
الفنان	529
الفيلم	486
فيلم	433
مجموعة	379
الفنية	362
العمل	356
السينما	312
الثقافي	300
الكتاب	298
الفنانين	297
الكاتب	297
المخرج	268

by applying 4-grams method, we get most common sentences used in art topics as shown in the table.

word	Word count
('وزارة', 'الثقافة', 'والشباب', 'والرياضة')	**39**
('محمد', 'السادس', 'للفنّ', 'الحديث')	23
('وزير', 'الثقافة', 'والشباب', 'والرياضة')	19
('النقابة', 'المغربية', 'لمهنيي', 'الفنون')	19
('المهرجان', 'الدولي', 'للفيلم', 'بمراكش')	18
('متحف', 'محمد', 'السادس', 'للفنّ')	16

The second class is economy and by applying unigram method we get most common words used in economies topics as shown.

(('مليار',), 704),

(('درهم',), 701),

(('سنة',), 697),

(('الاقتصاد',), 645),

(('بشكل',), 620),

(('القطاع',), 616),

(('بنسبة',), 607),

(('المالية',), 581),

(('الحكومة',), 581),

(('العام',), 580),

(('كورونا',), 575),

(('المقاولات',), 531),

(('فيروس',), 522),

(('مليون',), 502),

(('السنة',), 495),

(('الوطني',), 483),

(('قطاع',), 483),

(('الاقتصادية',), 455)]

By applying 4-grams

(('الاقتصاد', 'والمالية', 'وإصلاح', 'الإدارة'), 89)

(('الاقتصاد', 'والمالية', 'وإصلاح', 'الإدارة،'), 83)

(('وزير', 'الاقتصاد', 'والمالية', 'وإصلاح'), 73)

(('في', 'المائة', 'من', 'الناتج'), 68)

(('الاتحاد', 'العام', 'لمقاولات', 'المغرب'), 61)

(('وزارة', 'الاقتصاد', 'والمالية', 'وإصلاح'), 58)

(('في', 'الفترة', 'نفسها', 'من'), 54)

(('المائة', 'من', 'الناتج', 'الداخلي'), 52)

(('الصناعة', 'والتجارة', 'والاقتصاد', 'الأخضر'), 48)

(('الفصل', 'الأول', 'من', 'السنة'), 41)

(('مليار', 'درهم', 'على', 'شكل'), 40)

(('درهم', 'في', 'الفترة', 'نفسها'), 37)

(('والصيد', 'البحري', 'والتنمية', 'القروية'), 37)

(('البحري', 'والتنمية', 'القروية', 'والمياه'), 37)

(('محمد', 'بنشعبون،', 'وزير', 'الاقتصاد'), 36)

(('الفلاحة', 'والصيد', 'البحري', 'والتنمية'), 36)

(('بنشعبون،', 'وزير', 'الاقتصاد', 'والمالية'), 35)

(('أزمة', 'فيروس', 'كورونا', 'المستجد'), 35)

(('على', 'شكل', 'تسبيقات', 'لمدة'), 34)

The third class is sports and by applying unigram method we get most common words used in sports topics as shown.

(('لكرة',), 1047),

(('الفريق',), 981),

(('المغربي',), 938),

(('الموسم',), 824),

(('القدم',), 808),

(('الدوري',), 801),

(('اللاعب',), 759),

(('نادي',), 666),

(('فريق',), 546),

(('مباراة',), 479),

(('الدولي',), 467),

(('القدم،',), 449),

(('الرياضي',), 426),

(('النادي',), 417),

By applying 4-grams

(('الجامعة', 'الملكية', 'المغربية', 'لكرة'), 114),

(('الملكية', 'المغربية', 'لكرة', 'القدم'), 83),

(('الملكية', 'المغربية', 'لكرة', 'القدم،'), 80),

(('للجامعة', 'الملكية', 'المغربية', 'لكرة'), 44),

(('الاتحاد', 'الدولي', 'لكرة', 'القدم'), 38),

(('العصبة', 'الوطنية', 'لكرة', 'القدم'), 36),

(('الدولي', 'المغربي', 'أشرف', 'حكيمي،'), 26),

(('على', 'موقع', 'التواصل', 'الاجتماعي'), 25),

(('خوض', 'تجربة', 'احترافية', 'جديدة'), 24),

(('اللاعب', 'البالغ', 'من', 'العمر'), 23),

(('الوطنية', 'لكرة', 'القدم', 'الاحترافية'), 22),

The fourth class is media and by applying unigram method we get most common words used in medias topics as shown.

(('الجريدة',), 944),

(('رئيس',), 813),

(('المغربية',), 810),

(('الوطنية',), 782),

(('""المساء',), 781),

(('الأحداث',), 647)]

By applying 4-grams

(('بعض', 'الجرائد', 'الورقية', 'الخاصة'), 76),

(('مواد', 'بعض', 'الجرائد', 'الورقية'), 54),

(('الجرائد', 'الورقية', 'الخاصة', 'بيوم'), 52),

(('نستهلها', 'من', '""المساء"،', 'التي'), 49),

(('الحكومة،', 'سعد', 'الدين', 'العثماني،'), 48),

(('المديرية', 'العامة', 'للأمن', 'الوطني'), 45),

(('""المساء"،', 'التي', 'ورد', 'بها'), 43),

(('نشرت', '""أخبار', 'اليوم""', 'أن'), 43),

(('من', '""الأحداث', 'المغربية""', 'التي'), 43),

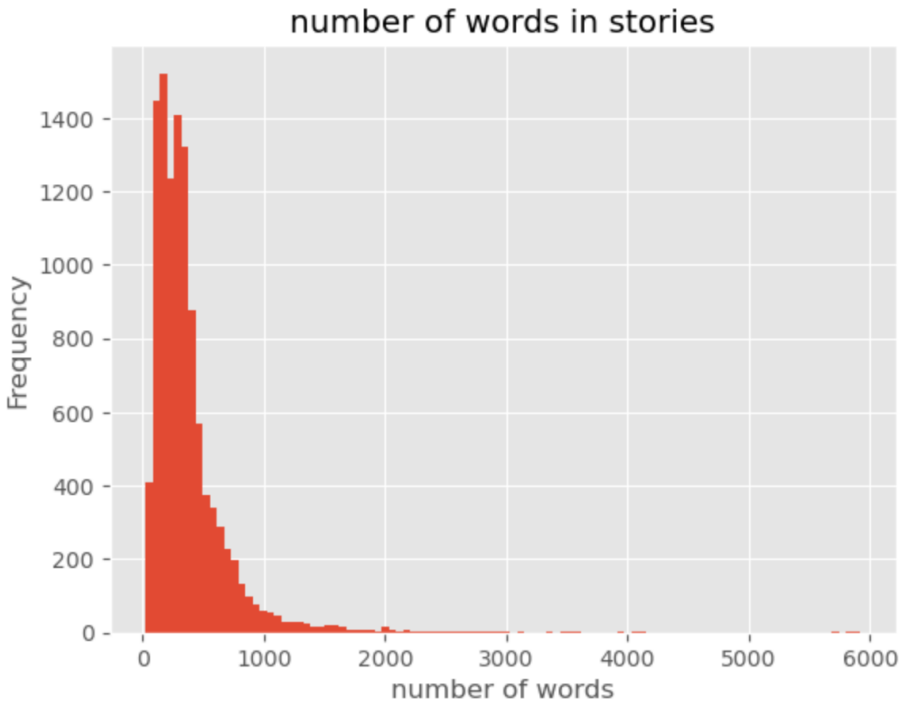
(('نشرت', '""الأحداث', 'المغربية""', 'أن'), 42),

(('رئيس', 'الحكومة،', 'سعد', 'الدين'), 42),

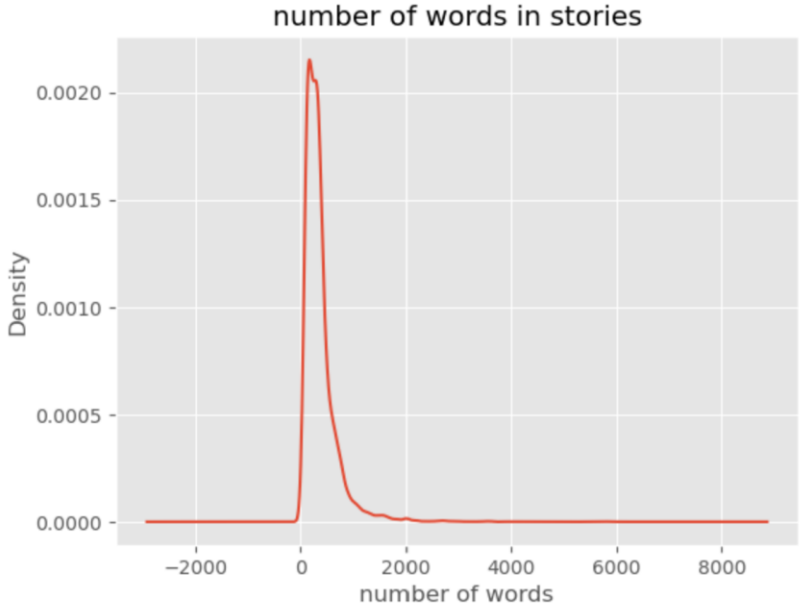
(('إلى', 'مصادر', 'الجريدة،', 'فإن'), 42),

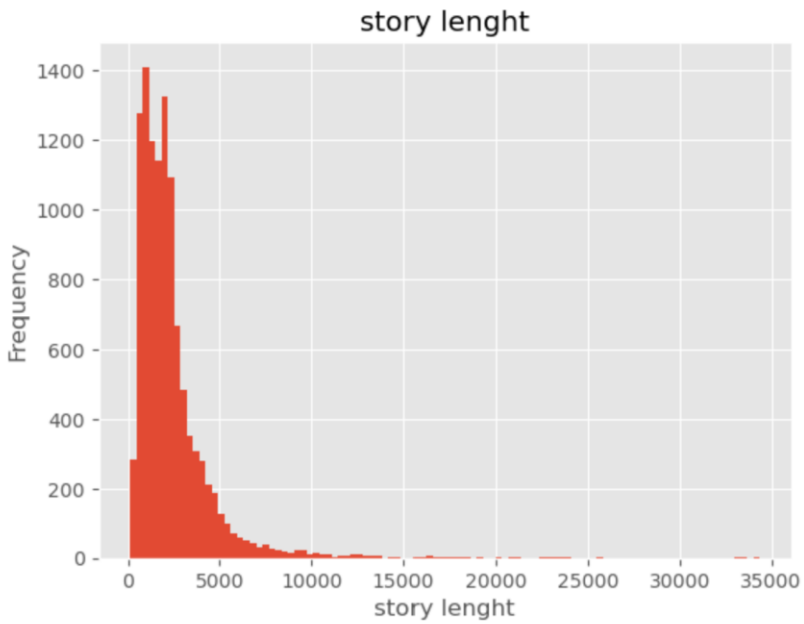
(('الشركة', 'الوطنية', 'للإذاعة', 'والتلفزة'), 40),

Third insight is lengths of examples in words and letters and some relations.



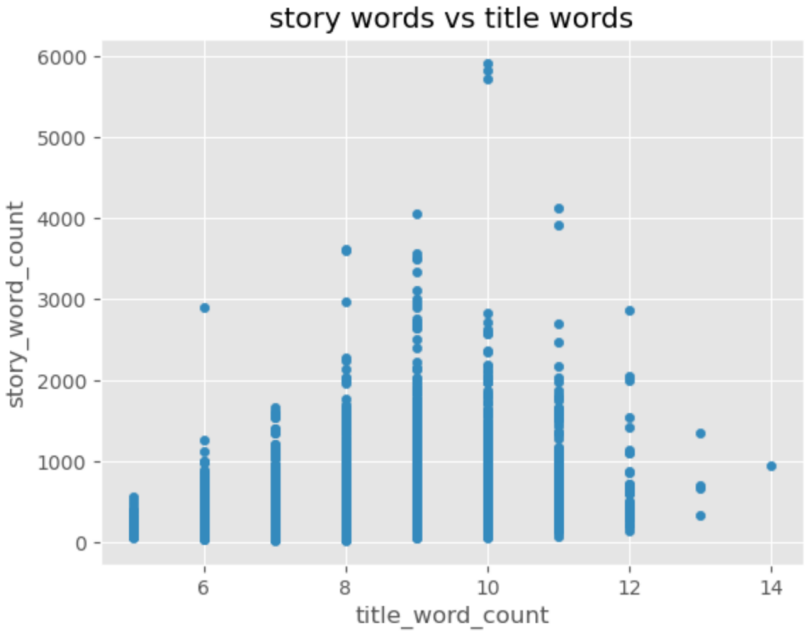
As shown, most of stories are in range 100 to 500 word





As shown most of stories are in range 1000 to 3000 letters

The next plot is the relation between the title length.



As shown stories that have biggest word number have ana average title word number (from 8 to 12-word title)

Fourth insight is getting top 10 authors in the data.

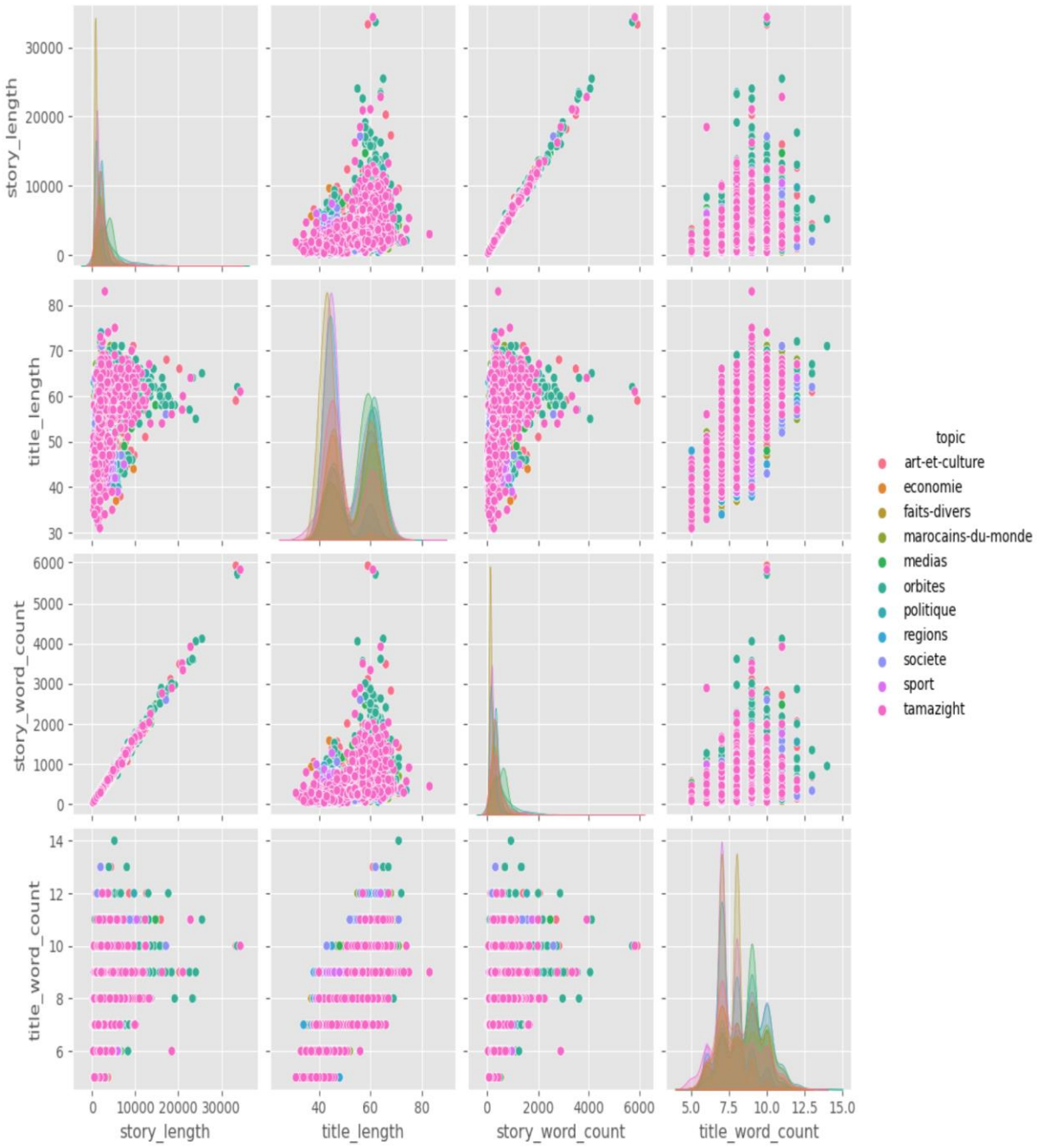
The author that has biggest number of stories is “هسبير من الرباط”

With a number near to 3900.

The next insight in the relation between topic name and story word number

And as shown Tamazight , orbites and art are the most topics that uses big number of words in a story.

Finally this is a matrix of relations between [story length, title length , story word sum , title word sum] and the topic



Resources:

Data : Hespress | Kaggle

Notebook draft : notebook1ef13cdf5b | Kaggle