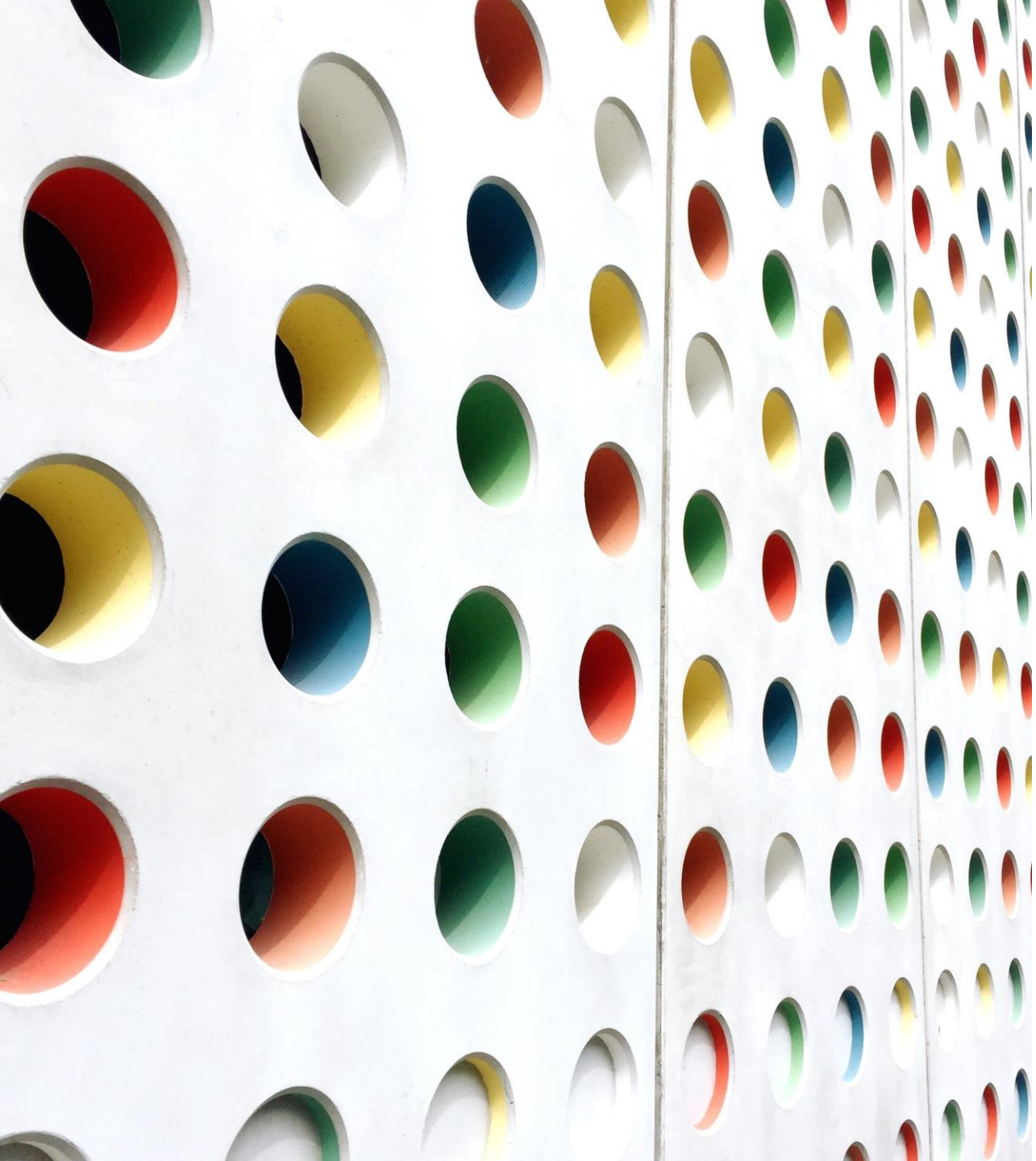


Topic Modeling

ABDULAZIZ & IBRAHIM

GOAL

Topic Modeling Newspapers' Articles

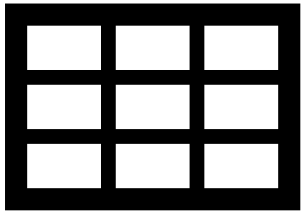
DATA

www.alriyadh.com

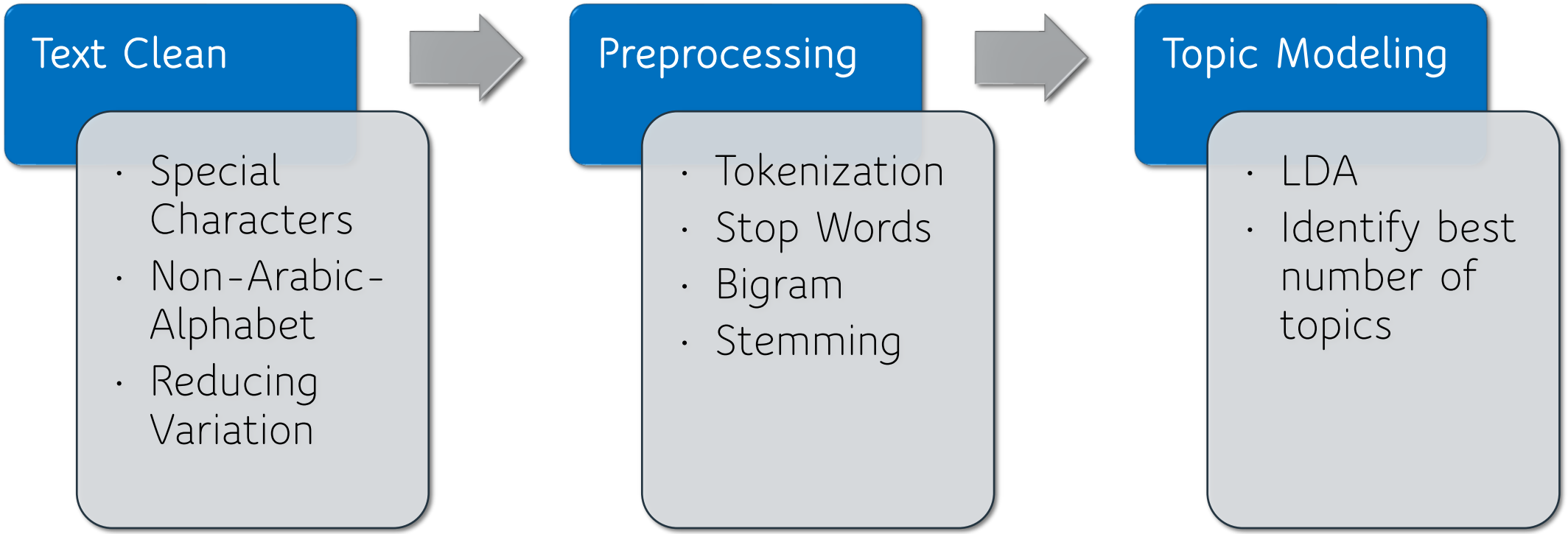
Web Scraping ~1GB

16 Features

~300K Rows (Articles)

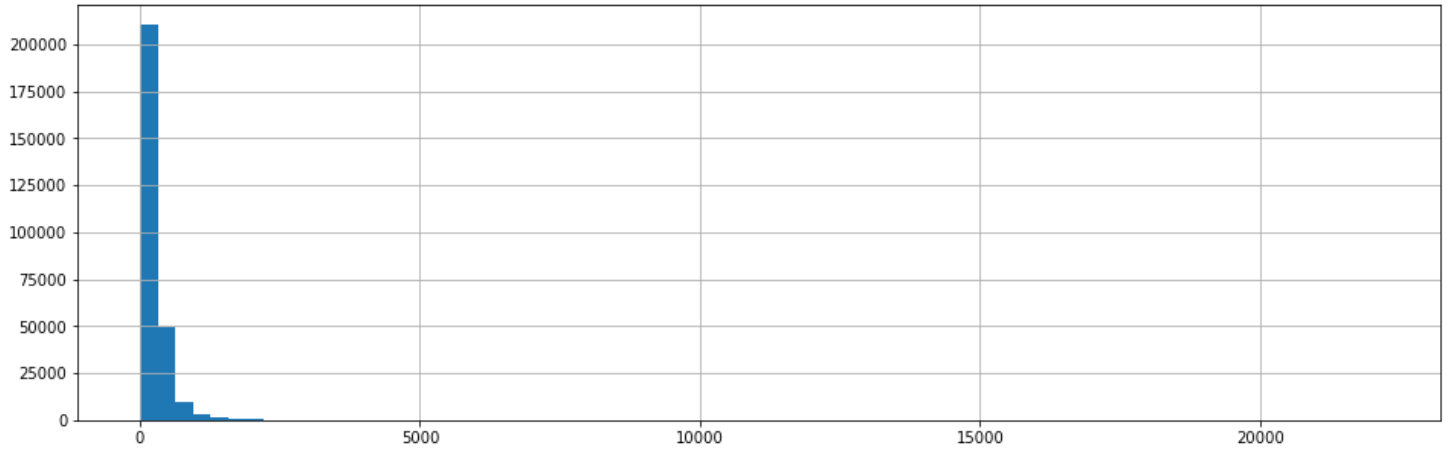


PIPELINE

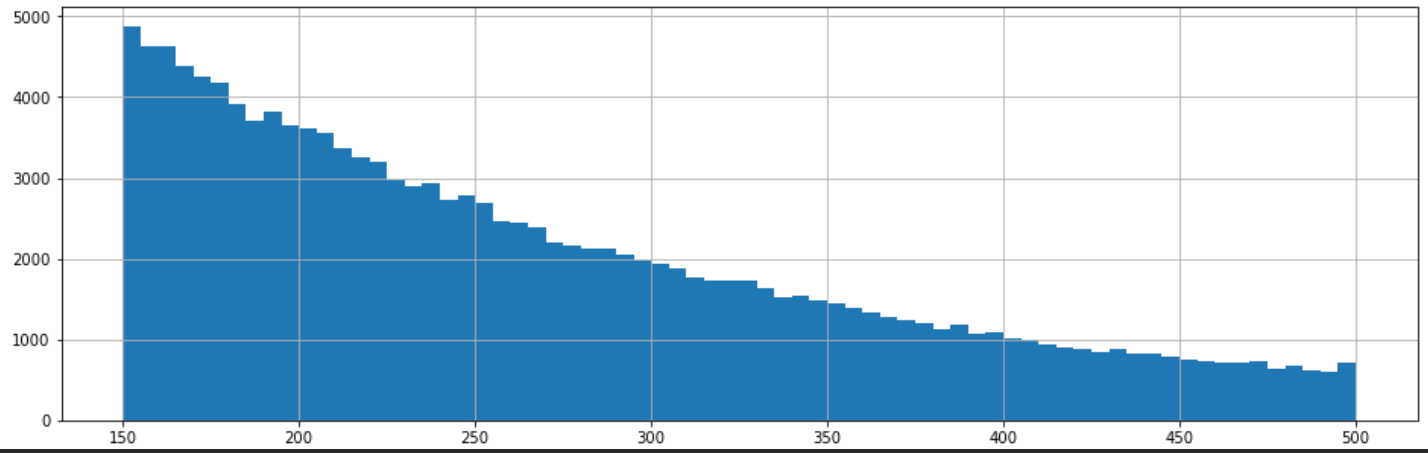

Text Clean
Special Characters
Non-Arabic-Alphabet
Reducing Variation
Preprocessing
Tokenization
Stop Words
Bigram
Stemming
Topic Modeling
LDA
Identify best number of topics

CLEANING (Useless Articles)

Original articles lengths



By taking articles with words 150-500



CLEANING (Special Characters)

\nرعى وزير التعليم\xa0د. أحمد بن محمد العيسى رئيس المؤتمر العام لمكتب التربية العربي لدول الخليج صباح أمس بمقر وزارة التعليم بالرياض احتفالية جائزة مكتب التربية العربي لدول الخليج للعام 1438-1439هـ في دورتها التاسعة والتي فازت بها مؤسسة الملك عبدالعزيز ورجاله للموهبة والإبداع "موهبة"

رعى وزير التعليم د. أحمد بن محمد العيسى رئيس المؤتمر العام لمكتب التربية العربي لدول الخليج صباح أمس بمقر وزارة التعليم بالرياض احتفالية جائزة مكتب التربية العربي لدول الخليج للعام 1438-1439هـ في دورتها التاسعة والتي فازت بها مؤسسة الملك عبدالعزيز ورجاله للموهبة والإبداع "موهبة"

CLEANING (Non-Arabic-Alphabet)

رعى وزير التعليم د. أحمد بن محمد العيسى رئيس المؤتمر العام لمكتب
التربية العربي لدول الخليج صباح أمس بمقر وزارة التعليم بالرياض
احتفالية جائزة مكتب التربية العربي لدول الخليج للعام 1438
-1439هـ في دورتها التاسعة والتي فازت بها مؤسسة الملك عبدالعزيز
ورجاله للموهبة والإبداع "موهبة"

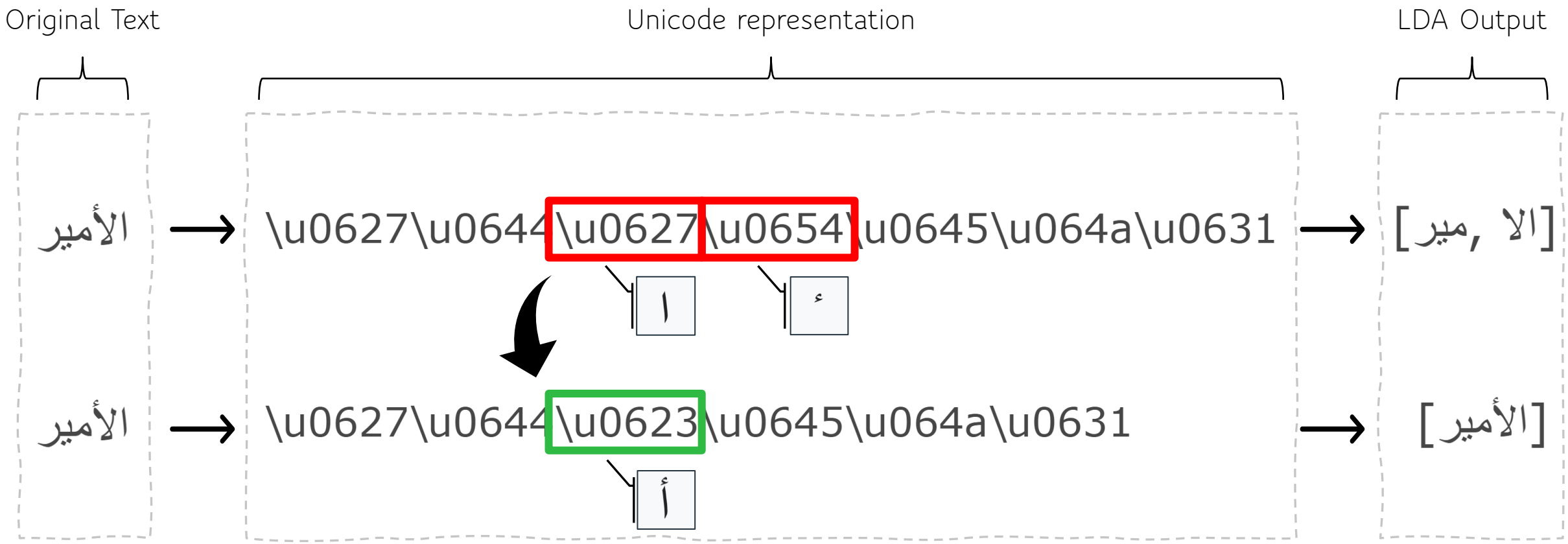
رعى وزير التعليم د أحمد بن محمد العيسى رئيس المؤتمر العام لمكتب
التربية العربي لدول الخليج صباح أمس بمقر وزارة التعليم بالرياض
احتفالية جائزة مكتب التربية العربي لدول الخليج للعام هـ في دورتها
التاسعة والتي فازت بها مؤسسة الملك عبدالعزيز ورجاله للموهبة
والإبداع موهبة

CLEANING (Reducing Variation)

رعى وزير التعليم د أحمد بن محمد العيسى رئيس المؤتمر العام لمكتب
التربية العربي لدول الخليج صباح أمس بمقر وزارة التعليم بالرياض
احتفالية جائزة مكتب التربية العربي لدول الخليج للعام هـ في دورتها
التاسعة والتي فازت بها مؤسسة الملك عبدالعزيز ورجاله للموهبة
والإبداع موهبة

رعى وزير التعليم د احمد بن محمد العيسى رئيس المؤتمر العام لمكتب
التربيه العربي لدول الخليج صباح امس بمقر وزاره التعليم بالرياض
احتفاليه جائزه مكتب التربيه العربي لدول الخليج للعام هـ في دورتها
التاسعه والتي فازت بها مؤسسه الملك عبدالعزيز ورجاله للموهبه
والابداع موهبه

Challenge #1


Original Text
Unicode representation
LDA Output
الأمير
\u0627\u0644\u0627\u0654\u0645\u064a\u0631
[الا, مير]
ا
ء
الأمير
\u0627\u0644\u0623\u0645\u064a\u0631
[الأمير]
أ

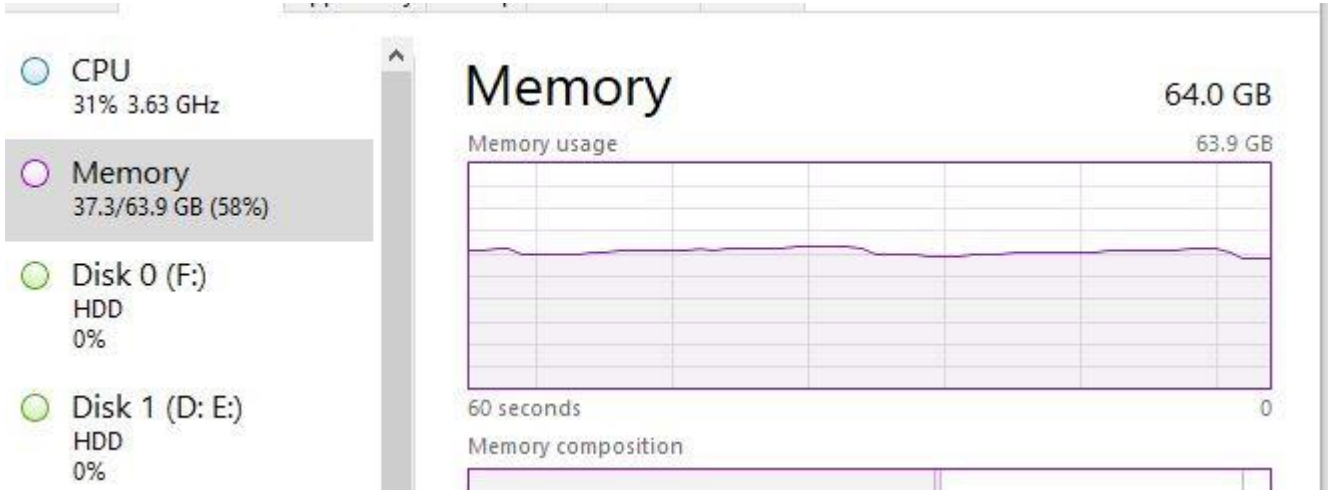
Challenge #2

- Arabic NLTK Stop Words are **NOT enough**
- We had to add stop words e.g..

'هـ', 'د', 'م', 'الى', 'ان', 'اذ', 'لهذه', 'قال', 'وقال', 'اكد', 'عدد', 'بعدد', 'وعدد'
'والتي', 'بن', 'بنت', 'وقد', 'ا', 'عبر', 'خلال', 'او', 'الا', 'وان', 'اي', 'بان', 'كان'
'كانت', 'تم', 'الف', 'مليون', 'وفي', 'وقد', 'اكثر', 'اقل', 'انه', 'وانه', 'قالت', 'وقالت', 'وتم'

Challenge #3 Memory issues

- LDA takes a lot of memory.
- Faced issues using Google cloud DataProc cluster for pySpark.
- We used high spec PC for processing.



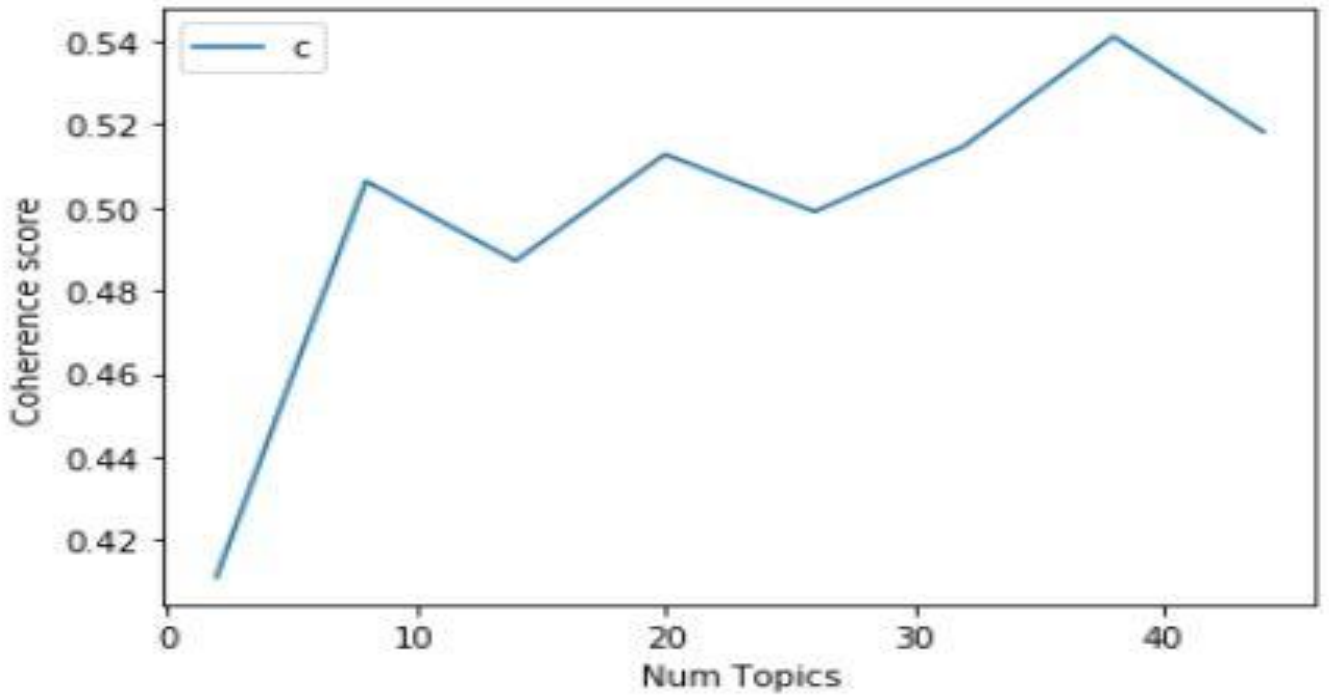
Topic Modeling (LDA)

- Meaningless topics at first.
- Uncleaned stopwords were a topic.
- difficulty distinguishing topics with shared words.

Count Vectorizer	TF-IDF
Higher probability topics/document	Higher composition of topics/document

Topic Modeling (Number of Topics)

Coherence was used to find optimal number of topics.



Topic Modeling

	No Stemming	Stemming
Perplexity	-13.00	-7.24
Coherence	0.578	0.462

Topics (Without Stemming)

قطر	دولي	مجلس الوزراء	انظمه	؟؟	الامن	مناطق	طاقه	عسكرية	
10	9	8	7	6	5	4	3	1	0
قطر	سورية	عبدالعزيز	نظام	اذا	الامن	سموه	الطاقة	اليمن	خدمات
القطرية	رييس	مجمد	لجنة	قايلا	الشرطة	امير	الصين	القوات	صحية
عام	البلاد	العهد_ولي	العمل	عندما	الداخلية	المنطقة	ترامب	الجيش	وزارة
العالم	الحكومة	مجلس	قرار	الامر	الأمنية	عبدالعزيز	ارامكو	العسكرية	معلومات

فعاليات	النقل العام	الشرق الاوسط	طاقة	مشاريع	تعليم عالي	صحة عامه	بلدية	تعليم	رياضه
30	29	27	26	20	19	18	17	15	13- 11
معرض	مطار	العراق	النفط	مشروع	الجامعة	السلامة	البلدية	التعليم	نقطة
مهرجان	السيارة	فلسطين	الانتاج	تنفيذ	الجايزه	المرور	الامانه	المدارس	ملعب
فعاليات	الرحلات	الاحتلال	السودان	طريق	الدكتور	التدخين	مخالفة	الطلاب	الهلال
العديد	الدولي	الاردن	اوبك		البرامج	النوم	النظافة	التعليمية	النصر

Topics (With Stemming)

0	1	2	3	4	9	12	15	16	18
شرع	ملك	بنك	ريس	لعب	سعر	حكم	بحر	خدر	سلم
دين	وطن	صرف	صحف	ندي	نفط	نظم	نقل	كشف	الل
نفذ	سلم	مول	امر	فرق	خفض	قرر	قطر	خطر	سجد
كرم	شرف	سوق	شرط	قدم	رفع	عمل	طار	امن	دين

19	20	23	26	29	30	32	34	35	39
كتب	قوت	صحة	سكن	عرب	نقط	طفل	بلغ	سكر	وطن
ارخ	نظم	طبه	عقر	دول	ثلث	اسر	نسب	شخص	هدف
ترث	يمن	شفى	وزر	رهب	رصد	اجتماعيه	شهر	علج	دعم
ثقف	امن	علج	اسك	سلم	فوز	جمع	سجل	مرض	وظف

Conclusion

- LDA achieved good results.
- Stemming have better metric but affect interpretability.
- Over 1 million Arabic words have been grouped into topics.

Ask us