

Saudi Newspapers Articles

NLP Project



Introduction

The media occupies an important role in a democratic society, the press is not just a source of news, but a reference for current information and a tool for public criticism, it greatly influences the making of public opinion. Perceptions, reports etc., responsible for investigating and writing reports on prevailing global issues, political or public developments and on some miscellaneous areas; Such as sports, economics, internal issues, etc., and the press has many basic functions, and the nature of each newspaper differs from the other in terms of the personality of the writer.



Table of contents

01

Introduction

02

Tools

03

Preprocessing

04

EDA

05

Topic Modeling

06

Conclusion

Data set

This data set from Github contains a collection of 3000 rows , 3 columns descriptive newspaper articles extracted from various Saudi newspapers on the Internet .



Tools

Technologies



Python



Jupyter notebook



Google colab



Pandas , NumPy ,
Matplotlib , Seaborn



Sklearn , Nltk

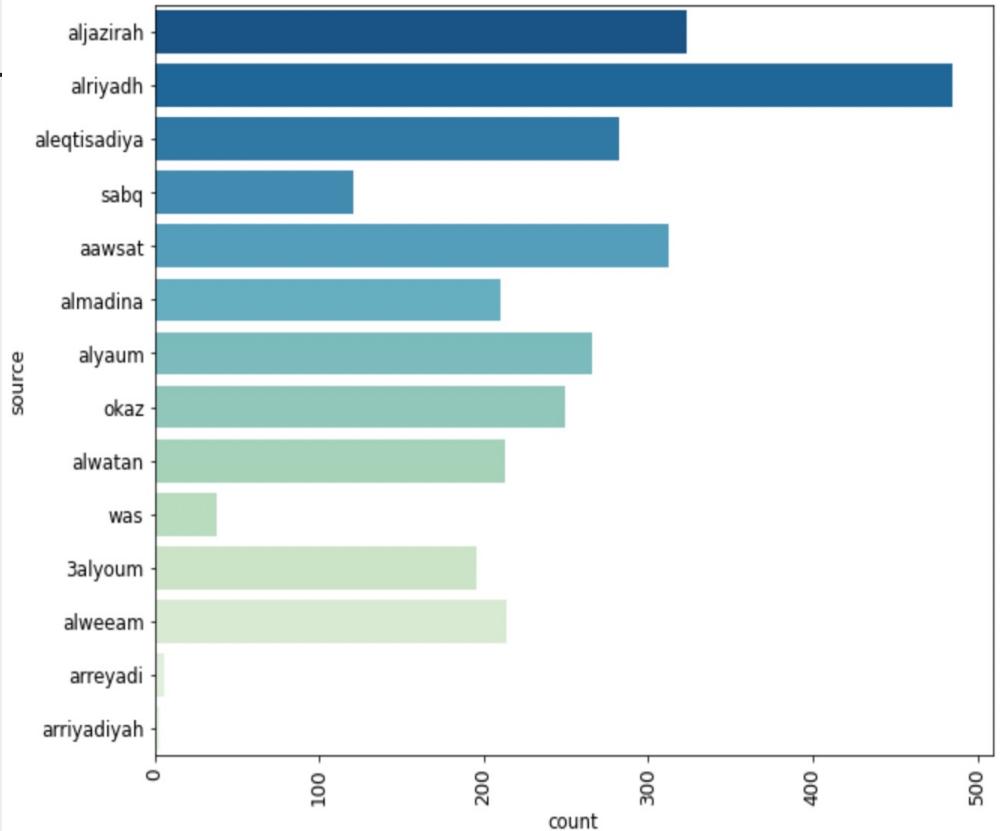


Arabic Libraries

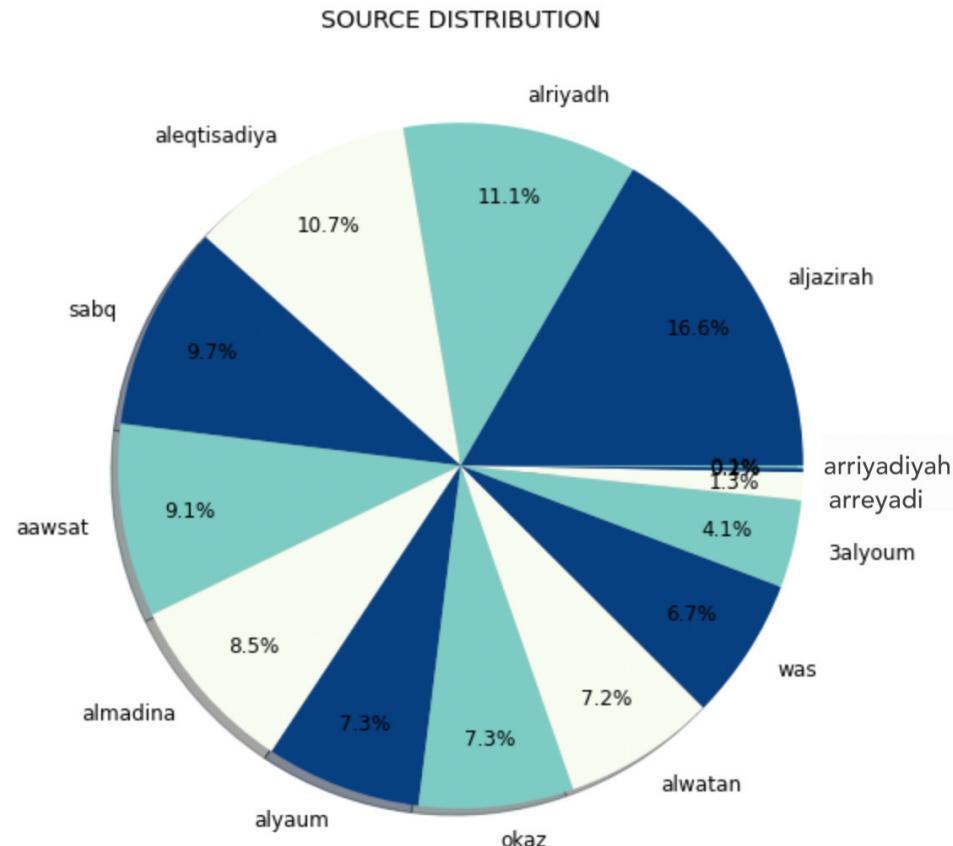


EDA

This figure shows us what is the most widely reported in Saudi news.



This figure shows us what is the percentage of each newspaper in publishing Saudi news.

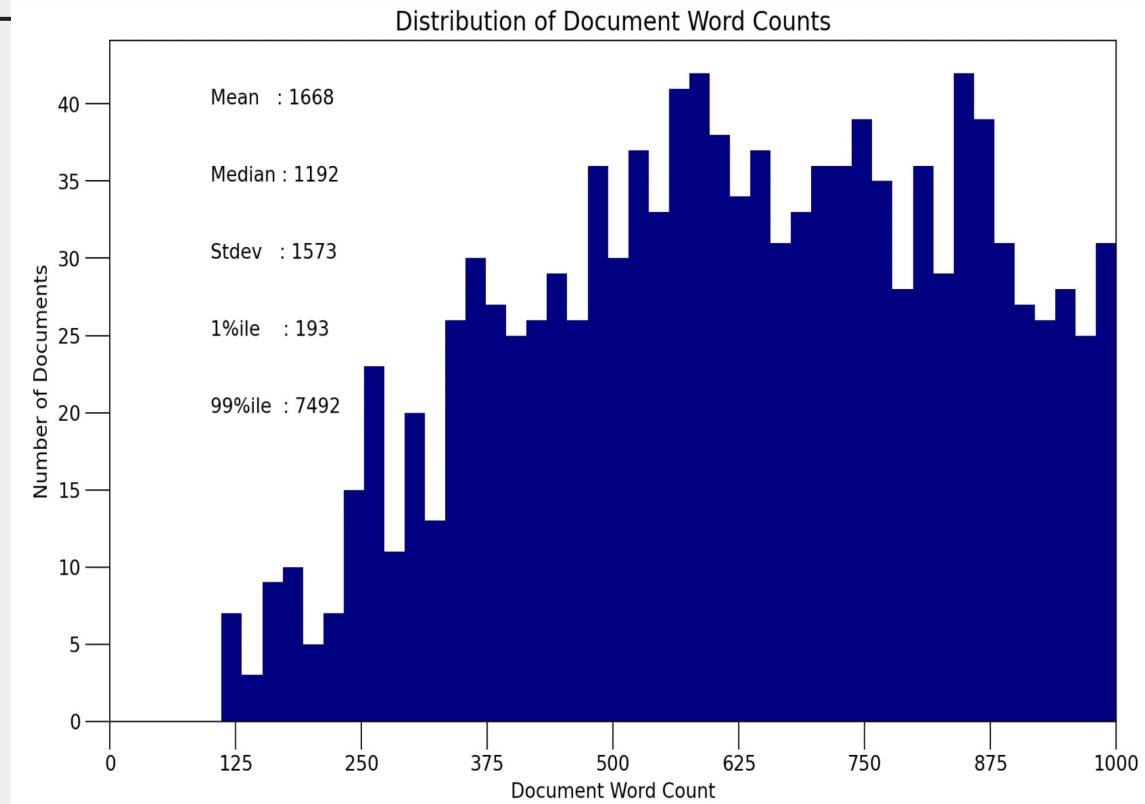


Word Cloud of all the words



Word cloud of the most frequent words

The number of words in each document is greater than the number of documents



NLP Preprocessing

Remove :

- English letters
- English number
- Special characters
- Arabic punctuation



NLP Preprocessing

- **Stemming**
- Remove Arabic stop words
- **TF-IDF
vectorizer**



Topic Modeling



NMF (Non-Negative Matrix Factorization)

After reviewing 8 of the topics, it became clear to us that the best topic is 5 and contains different documents



Public politics
Saudi politics
Sports



Public politics
Saudi politics
Sports
Economy

The best choice

Public politics
Saudi politics
Sports
Economy
Entertainment

topic extraction

Topic 0

واشنطن، مجلس، العربيه، الحكومه، الدولى، الارهابيه، الدول، العراقيه، القوات، المعارضه، التعاون، الكردستاني، مصر، الاميركي

Topic 1

دفاع، جمهوريه، السعوديه، الشهداء، الفيصل، الوطن، سمو، حفظه، تركي، مصر، فهد، النائب، المصابين، ومعالي، التفجير، الارهابي

Topic 2

ني، سان، لاعبين، معسكر، الاسباني، التدريبات، الجديد، يونايتد، الكره، ملعب، الجهاز، الشباب، الرياضي، وكان، الملعب، مباريات

Topic 3

لار، الاسعار، الطاقه، الاتصالات، الشركه، ادنى، سوق، الذهب، ارتفاعا، تداولات، لاوقيه، المدين، الماليه، عام، برنت، برميل، المؤشر

Topic 4

منطقه، امانه، برنامج، العيد، الانتخابات، الجهات، الشرقيه، الجامعه، طريق، ضمن، الفعاليات، عمل، مشروع، سوق، مكه، السعوديه

Topic 5

عسكريه، الجيش، العسكريه، الحكومه، الشرعيه، لحج، المخلوع، الجويه، تحرير، والمقاومة، مواقع، لاغائه، ميليشيات، عبد، السيطره

Topic Modeling



LDA (Latent Dirichlet Allocation)

We used the grid search function to give us a suggestion for the best topic, and the choice was topic 5, but the result was bad

Topic 0

النادي، ايران، داعش، الرئيس، ولی، مدير، اخري، الحرمين، المدينه، بشكل، الامن، الشريفين، محافظه، المتحده، الدولى

Topic 1

الدائرى، افعل، رياح، شاسعه، الاشكال، الغرف، الابواب، سر، الكل، ورفيق، وايلاف، لشققاته، اثيوبي، زل، ارقاوي، هيلان

Topic 2

زهمنه، والوريد، بدءى، الحشا، طعمته، بكى، كتمته، باتشينو، وتنكر، ذعر، التزمته، قباني، حرمته، شوق، حبك، والحزن

Topic Modeling



LSA (Latent Semantic Analysis)

Topic 0:

علي الي الله الملکه العام السعودیه محمد عبدالعزیز الامیر رئيس

Topic 1:

عبدالعزیز الامیر السمو الملکي ولی الحرمین الله الشریفین خادم سلمان

Topic 2:

الفريق النادي الاتحاد الاعب الموسم الاعبين القدم المدرب نادي لكره

Topic 3:

بنسبة المائة اسهم النفط سهم مليار اسعار السوق نقطه

Topic 4:

البلديه العمل الانتخابيه المنوره الوزاره الجنه المدينه التعليم الصحیه على

Topic 5:

عدن المقاومه الحوثين اسهم الحوثي الشعبيه اليمن بنسبة اسهم اليمنيه

Topic 6:

الله عسير اسهم داعش الطوارئ قوات اسهم الارهابي بنسبة الامن

Topic 7:

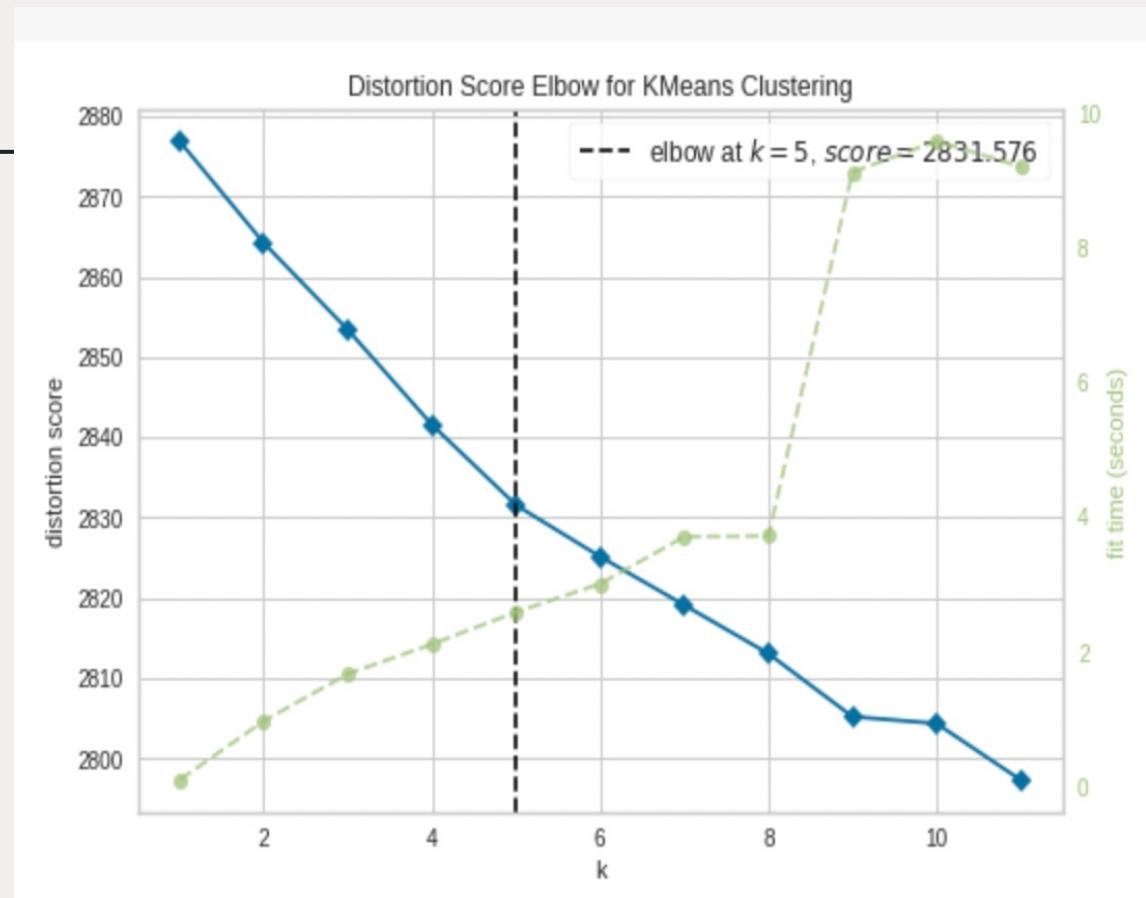
اسهم بنسبة اسهم لتأمين السعودیه سهم الاکثر نقطه قطاع شركات

We also used LSA and
the result was useless

K-MEANS



The initial score to the inertia model was 7 on the ELBOW graph, but after we cleaned the data again, the result was better 5 on the figure



RESULT

5 topics were extracted from the articles :

- Public politics
- Saudi politics
- Sports
- Economy
- Entertainment



conclusion

- We used 3 types of models and compared the results with the best model, NMF
In the future
- we will work on developing the model and doing more processing to show the best results



Thank you for listening



YARA ALDOSSARI

REHAB ALZAIDI

AMAL ALTHAQAFI

MUZOON ALSHAHRANI

NOURA ALOTABI