# Arabic News Articles

Alanoud Alosaimi & Mashael Alfehaid

# WORKFLOW

- Introduction
- Tools
- Dataset
- EDA and Preprocess
- Topics modelling

# INTRODACTION

# TOOLS

**Numpy**


Jupyter

**Pandas**


**Seaborn**

Matplotlib


**Scikit-Learn**

# DATASET

Extract the data from json file

Data Source: GitHub, Link

Before Cleaning: (31,030 x 6)

After Cleaning: (25,617 x 1)

# EDA


Removed Null
Removed Duplicated
Removed unneeded columns

# PREPROCESS

Removed punctuations and numbers.

Removed Arabic tatweel.

('السلام' to 'الســــــــلام')

Removed Arabic tashakeel.

('الْعَرَبِيّةُ' to 'العربيه')

Removed Arabic Stop words

(adding more than 200 words).

Removed repeating char.

('وكان' to 'ووووكان')

Removed Non Arabic latters.

Normalize Arabic text.

('يئ' to 'ي')

('أآإ' to 'ا')

Stemming and tokenization

# EMBEDDING

- Count Vectorizer
- TF-IDF Vectorizer

# TOPICS MODELING

**Latent Dirichlet Allocation (LDA)**

- Count Vectorizer
- TF-IDF Vectorizer

**Latent Semantic Analysis (LSA)**

- Count Vectorizer
- TF-IDF Vectorizer

**Non-Negative Matrix Factorization (NMF)**

- Count Vectorizer
- TF-IDF Vectorizer

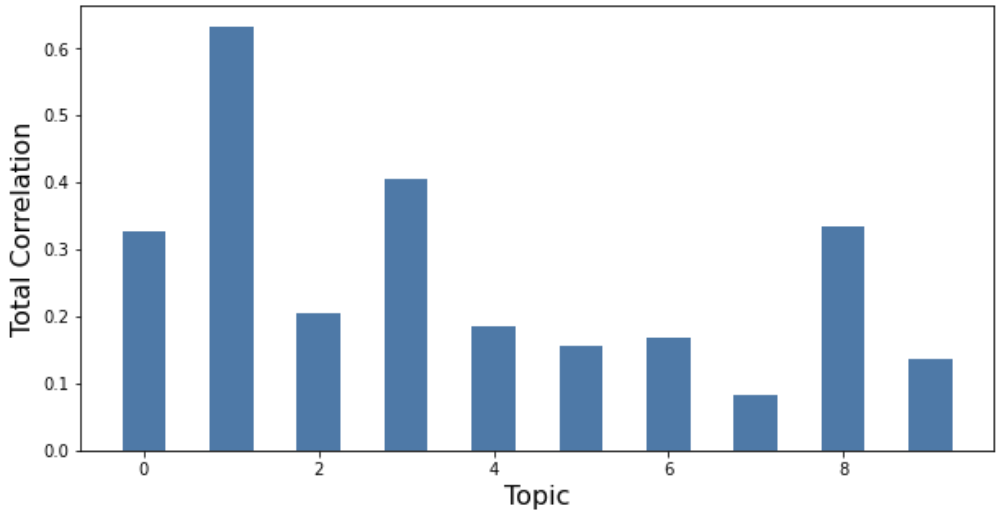
**CorEx**

- Count Vectorizer
- TF-IDF Vectorizer

**CorEx with Anchors**

- TF-IDF Vectorizer

# CorEx | TF-IDF Vectorizer




- داعش, قتل,تنظيم, هجوم, سوري, ارهاب, تركي, مقاتل, كردستان, مقتل ( إرهاب )
- فريق, كره, اعب, ناد, لاعب, مدرب, معسكر, لكر, اتحاد, موسم ( رياضة )
- جامعه ,طلاب ,تعليم ,انتخابيه ,جامع ,بلديه ,مدارس ,قبول ,مرشح ,ناخب ( تعليم )
- اسعار, سوق, مايه, نفط, اسهم, تراجع, مليار, موشر, دولار, ميه ( مالية )
- وزراء, حكومه, عدن, حوث, يمن, مقاومه, لاغاث, يمنيه, تحالف, انسانيه ( سياسية )
- مرض ,علاج , مستشف , طبيه, صحيه , طفله , مريض, امراض, فيروس, صحه ( صحة )
- حفل, مهرج, عاليا, عيد, فعال, مسرح, تهان, عروض, فطر, زوار ( ترفيه )
- امطار, رياح, سطحيه, حراره, طقس, افقيه, هطول, كمساع, لارصاد, غزير ( طقس )
- فكر, كتاب , دين, لماذ, شيء, مسلم, كثير, كاتب, نفس, لان ( ثقافة )
- شهداء, شهيد, طوار, عسير, مسجد, استشهاد, تفجير, تعاز, امن, يتغمد ( حرب )

# FUTURE WORK

Classification Model

Find similar articles

# Thanks for your listening