Saudi Newspapers Articles
Topic Extraction

Rawabi Alharbi

INTRODUCTION

- Newspapers Articles
- Problem statement

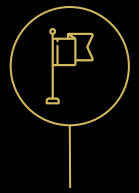
INTRODUCTION

- **Dataset**
  - Downloaded from Github
  - **31,030** Arabic newspaper articles

Methodology

Preprocessing

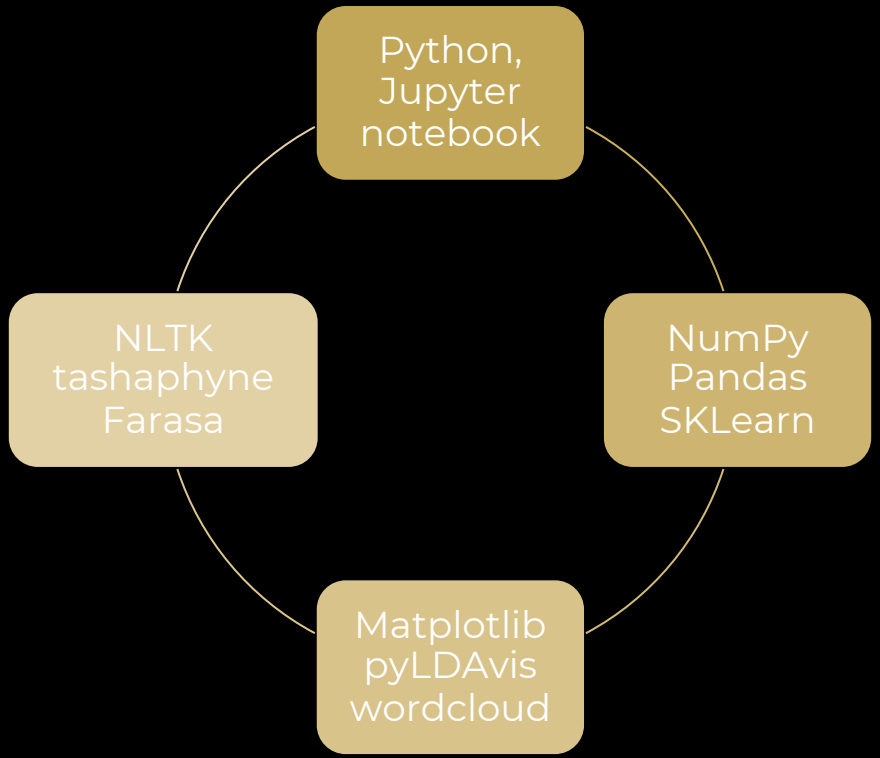
EDA



Topic Modeling

EDA

Tools


Python,
Jupyter
notebook
NumPy
Pandas
SKLearn
Matplotlib
pyLDAvis
wordcloud
NLTK
tashaphyne
Farasa

EDA

Word Cloud of the most frequent words

NLP Preprocessing

- **Remove**
  - English letters
  - English Numbers
  - Special characters
  - Arabic Punctuations
  - Newline characters

NLP Preprocessing

- **Stemming**
  - ISRIStemmer ❌
  - ArabicLightStemmer ❌
  - FarasaStemmer

NLP Preprocessing

- **TF-IDF Vectorizer**
  - Remove Arabic stop words

Topic Modeling - NMF

Tuning

Topic Modeling - LDA

EDA – After Topic Modeling

Word Cloud of the most frequent words in Topic 1

الرياضة

EDA – After Topic Modeling

Word Cloud of the most frequent words in Topic 2

السياسة السعودية

EDA – After Topic Modeling

Word Cloud of the most frequent words in Topic 3

السياسة العامة

EDA – After Topic Modeling

Word Cloud of the most frequent words in Topic 4

الاقتصاد

EDA – After Topic Modeling

Word Cloud of the most frequent words in Topic 5

السياحة/الترفيه

Result

5 topics were extracted from the articles

- Sports
- Saudi Politics
- General Politics
- Economy
- Entertainment

Conclusion

The model works with good performance to extract the most meaningful topics from the Saudi newspapers.

Future work

- Apply a good stemming & lemmatization
- Apply clustering


Thank
You!