ARABIC NEWS ARTICLES

PREPARED BY:
ALANOUD ALOSAIMI
MASHAEL ALFEHAID

ABSTRACT

In this project, we have a large collection of Arabic news articles which is a great resource that contains natural language text from various newspapers and that can be used in different tasks such as Text Classification and Word Embedding.

DESIGN

Our goal is to analyze a series of news articles to identify the domain that each article belongs to.

DATA

The dataset was extract from JSON files which is a dataset containing over 30K News articles from 13 different newspaper.

ALGOHRITHMS

Embedding:

- Count Vectorizer
- TF-IDF Vectorizer

Topic Modelling:

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)
- Non-Negative Matrix Factorization (NMF)
- CorEx
- CorEx with Anchors

TOOLS

- Data manipulation and cleaning: Pandas and Numpy.
- Data Storing: Sqlite3.
- Text preprocessing: NLTK,, gensim, scikit-learn , pyarabic.
- Visualization: matplotlib, seaborn.