Poem Comprehensive Dataset (PCD)

Abstract:

The purpose of this project is to apply the most important natural language processing techniques and starting with implementing all text preprocessing in "Arabic" then using clssification model to classify each Poet name and his poetries and we will using topic modeling to cluster each peorties depend on poetry types or poetry rhymes. We worked with data provided by Kaggle (https://hci-lab.github.io/ArabicPoetry-1-Private/).

Design:

This project originates from the Data Science Bootcamp (T5) to apply the most important natural language processing techniques .

Data:

The dataset contains 1,831,770 observations and 8 columns.

Algorithm:

- Cleaning the data from duplicate, nulls.

- Data Preprocessing
- Modeling
  K-means
  Hierarchical clustering with different linkages
  Ward , Single , complete , average

Visualization:

- bar plot
- Word cloud

Tools:

- Python and Jupyter Notebook.
- Numpy and Pandas for data manipulation.
- Matplotlib and Seaborn for plotting visuialization.
- Sklearn.
- nltk , pyarabic