Arabic Poem

Ohood Soliman
Ajwad Almajnuni
Mashael Asiri

أنا الذي نظر الأعمى إلى أدبي
وأَسمّعت كلماتي من به صمم

أبو الطيب المتنبي

01 Introduction

02 Methodology

03 EDA

04 Tools

05 Preprocessing

06 Modeling

07 Conclusion

Introduction

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.

About Data : This dataset contains 1,831,770 observations and 8 columns

Methodology

EDA

Preprocessing

Modeling

EDA

01

Check null values

Duplicates

02

Rename columns

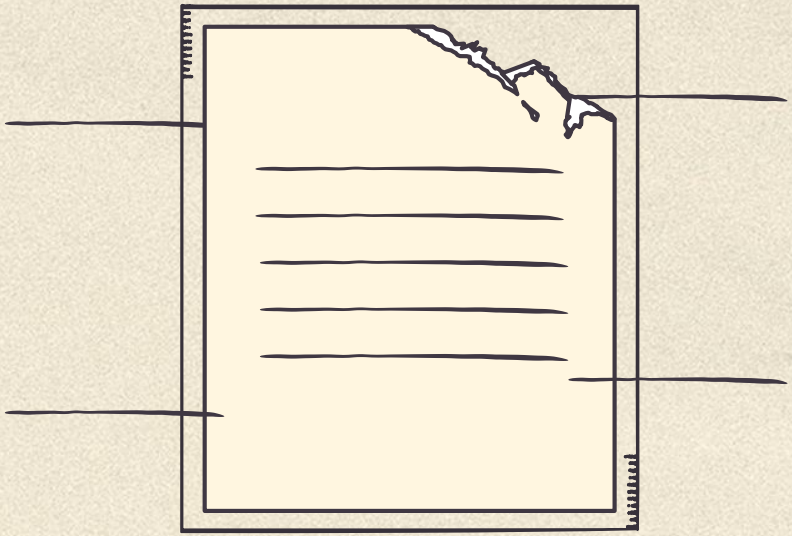
03

Visualization

Tools

Matplotlib
wordcloud

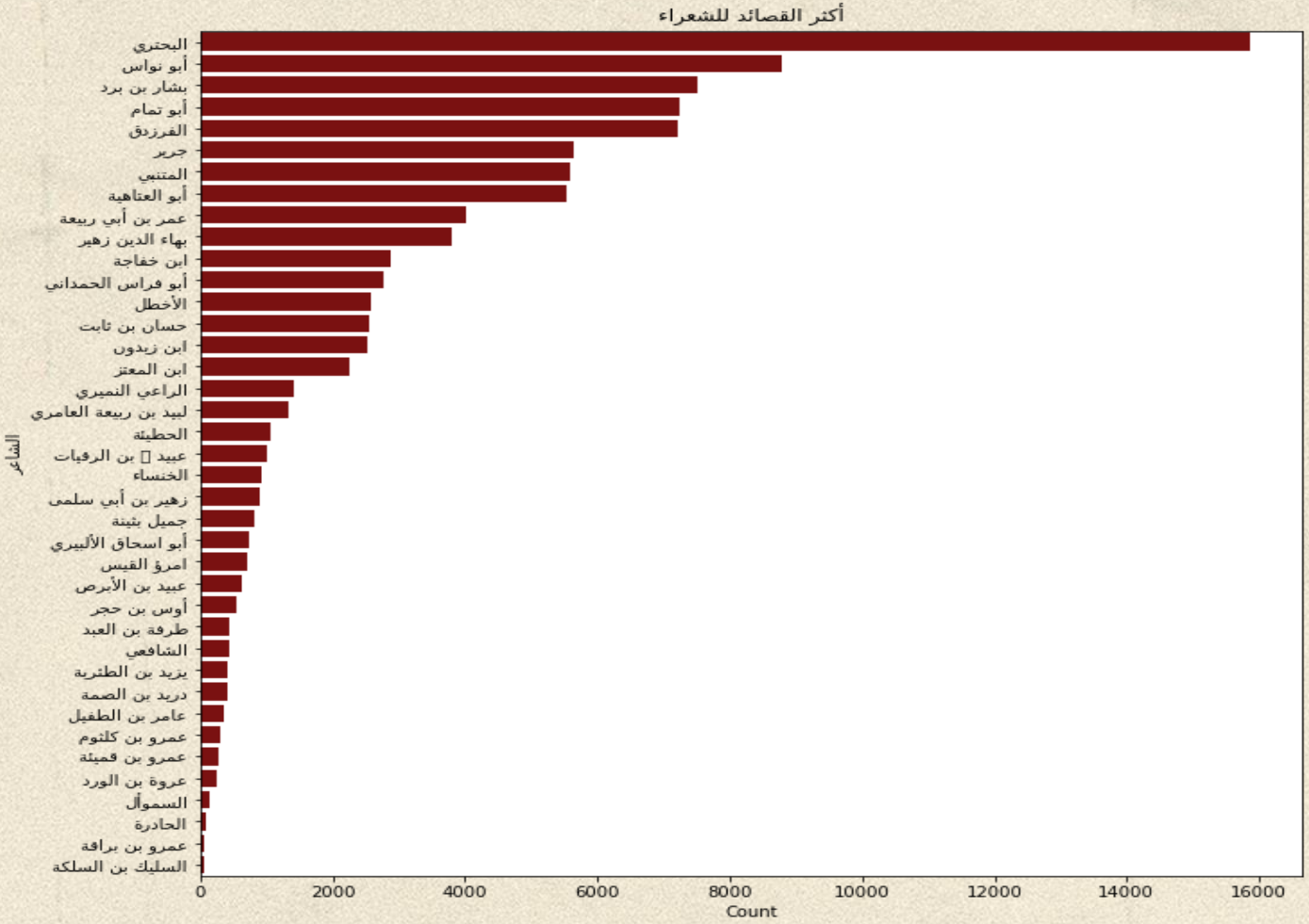
NLTK
Pyarabic



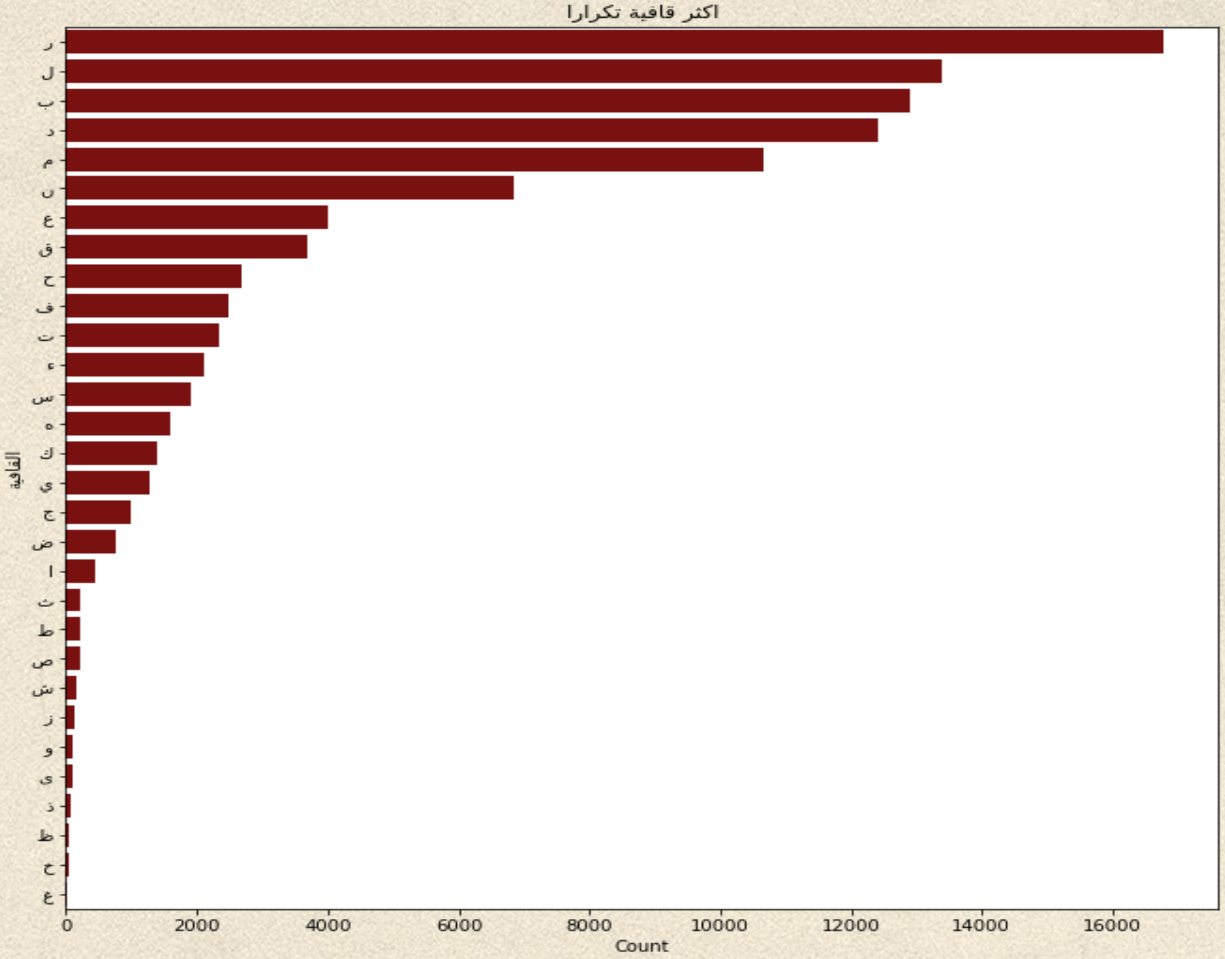
Python
Jupyter
notebook

NumPy
Pandas
SKLearn

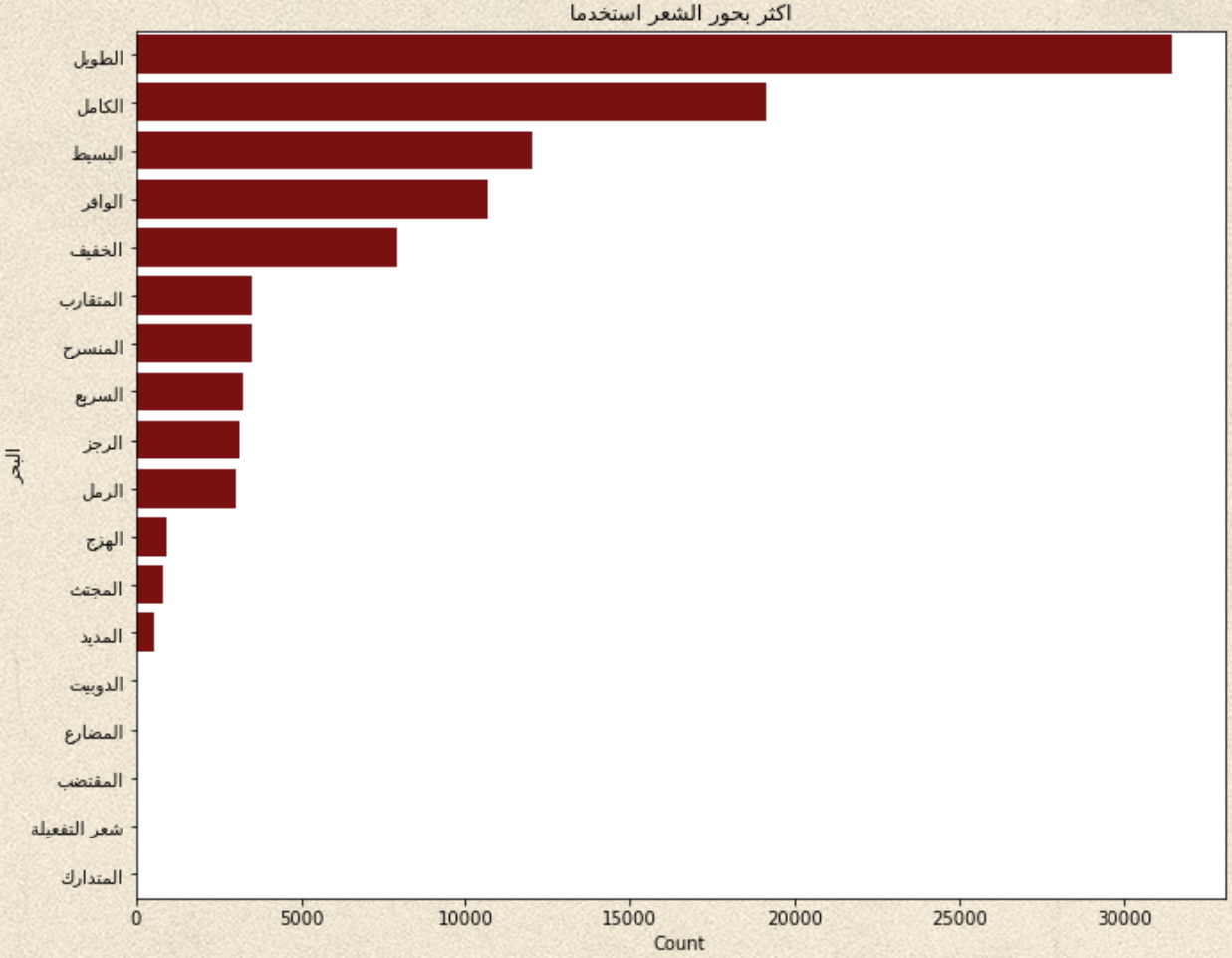
اكثر القصائد كانت للشاعر ...



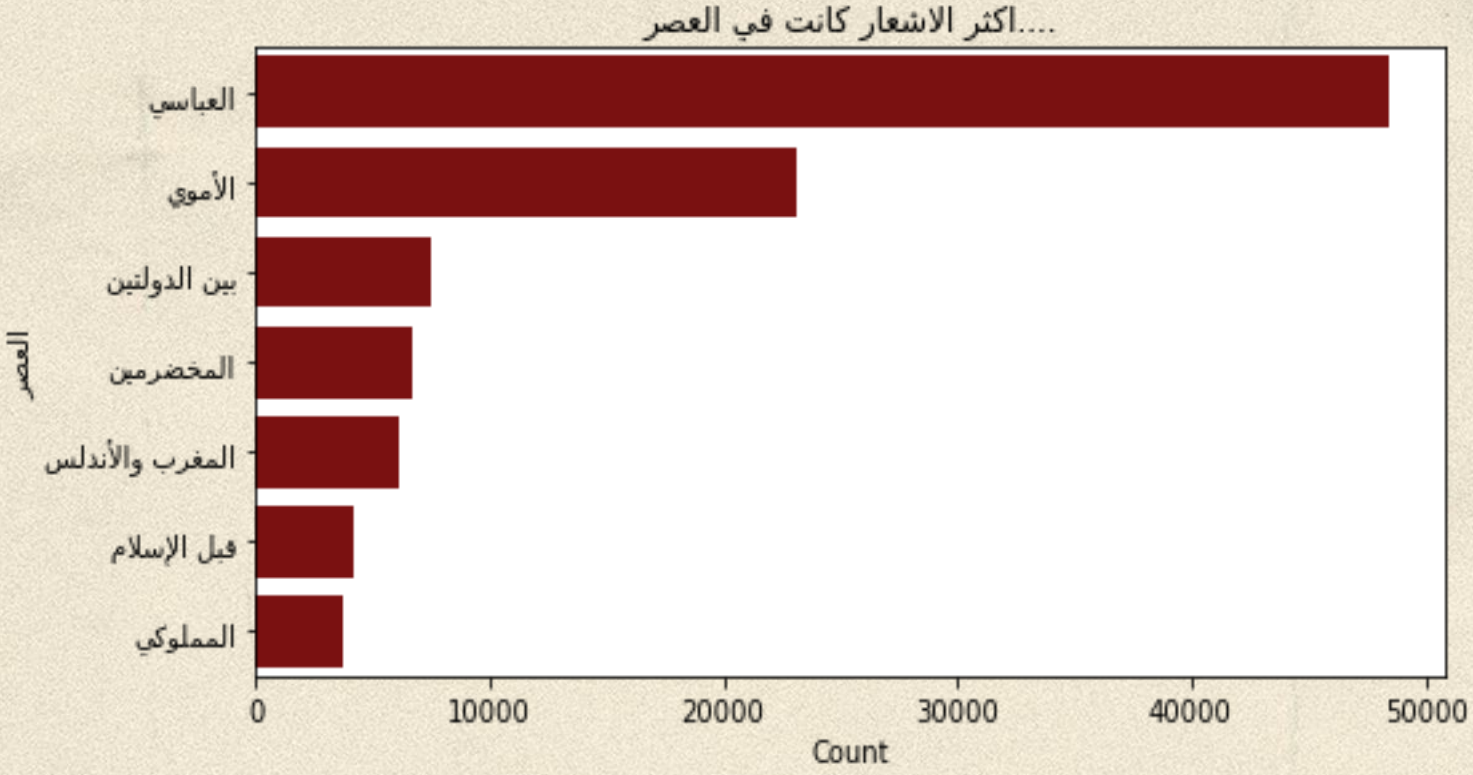
القافية الأكثر تكرارا



اكثر بحور الشعر استخداما



كانت اكثر الاشعار في العصر ....



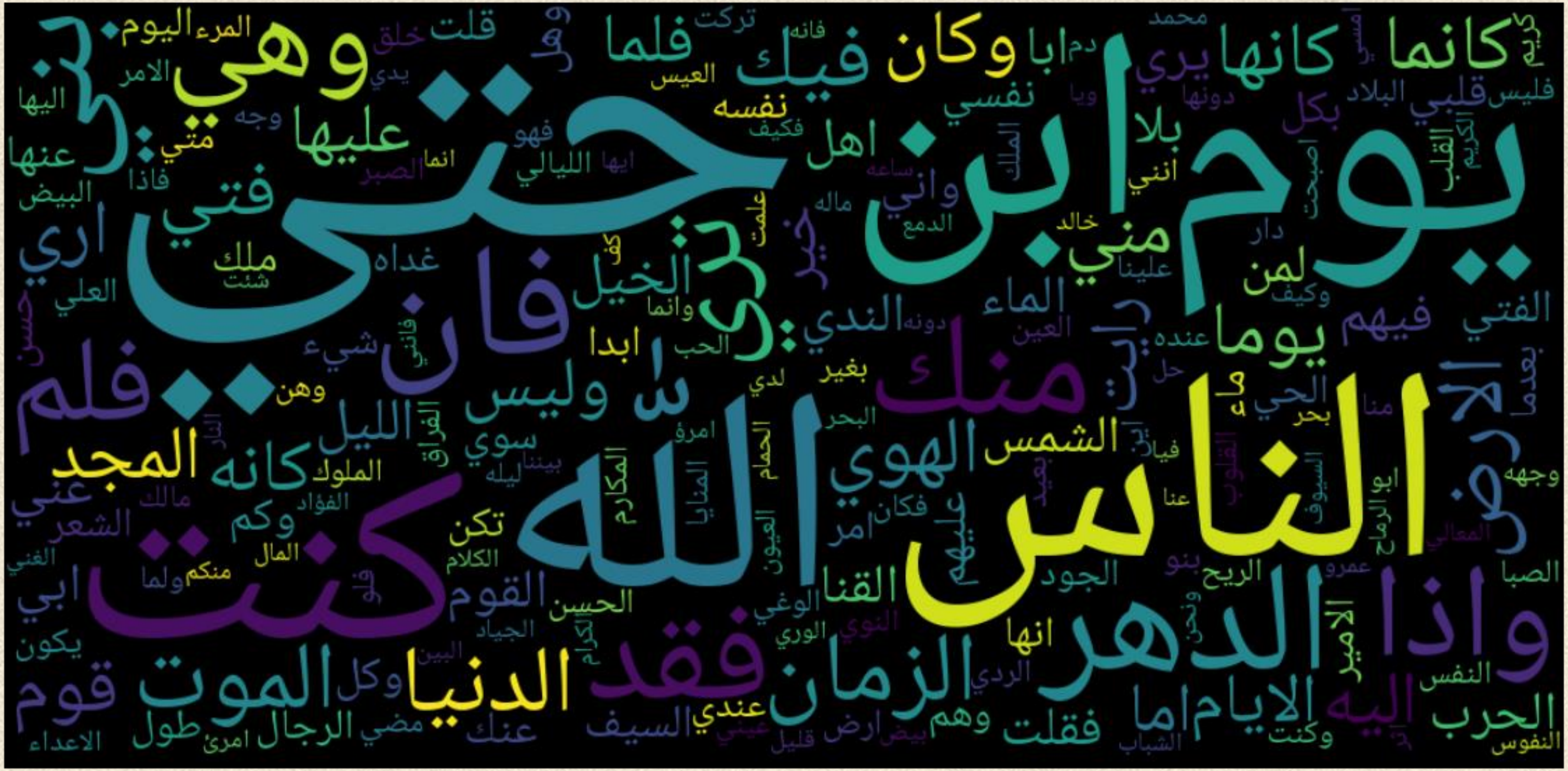
NLP Preprocessing

- **Remove**
  - *normalize Arabic*
  - repeated letters remove
  - punctuations
  - special character
  - non-arabic-alphabet

NLP Preprocessing

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

Wordcloud


حتى
يوم
ابن
الناس
الله
كنت
الدهر
واذا
منك
فان
فلم
فقد
الزمان
الموت
الدنيا
قوم
وهي
فلما
فيك
وكان
كانها
كانما
يري
عليها
فتي
المجد
كانه
الليل
وليس
الهوى
الشمس
الندى
يوما
مني
لمن
فيهم
الفتي
الحرب
الايام
اما
فقلت
السيف
عندي
القوم
القنا
الجود
انها
اهل
نفسي
بلا
ابا
قلت
اليوم
المرء
عنها
اري
ملك
الخيل
ابدا
الماء
سوي
تكن
وكم
ابي
عني
الشعر
يكون
طول
الرجال
عنك
وهم
ارض
النفس
الصبا
وجهه
النفوس

ML algrathim

Supervised modeling

- Logistic Regression
- Random Forest
- XGBoost
- Decision Tree

Unsupervised modeling

- *k*-means
- Hierarchical clustering

Logistic Regression

Training

Training Logistic Regression with best parameters1 **F1 score= 0.7456413443067098**

Test Logistic Regression1 with best parameters1 **F1 score= 0.6387911437172631**

Test

Decision Tree

Training

Training Decision Tree with best parameters1
F1 score= 0.30552456974887593

Test Decision Tree with best parameters1
F1 score= 0.2918717311288094

Test

Random Forest

Training

Training Random Forest with best parameters1
F1 score= 0.34913956079465364

Test Random Forest with best parameters1
F1 score= 0.3137159705219611

Test

XGBoost

Training

Training XGboost with best parameters1
F1 score= 0.6858045498280895

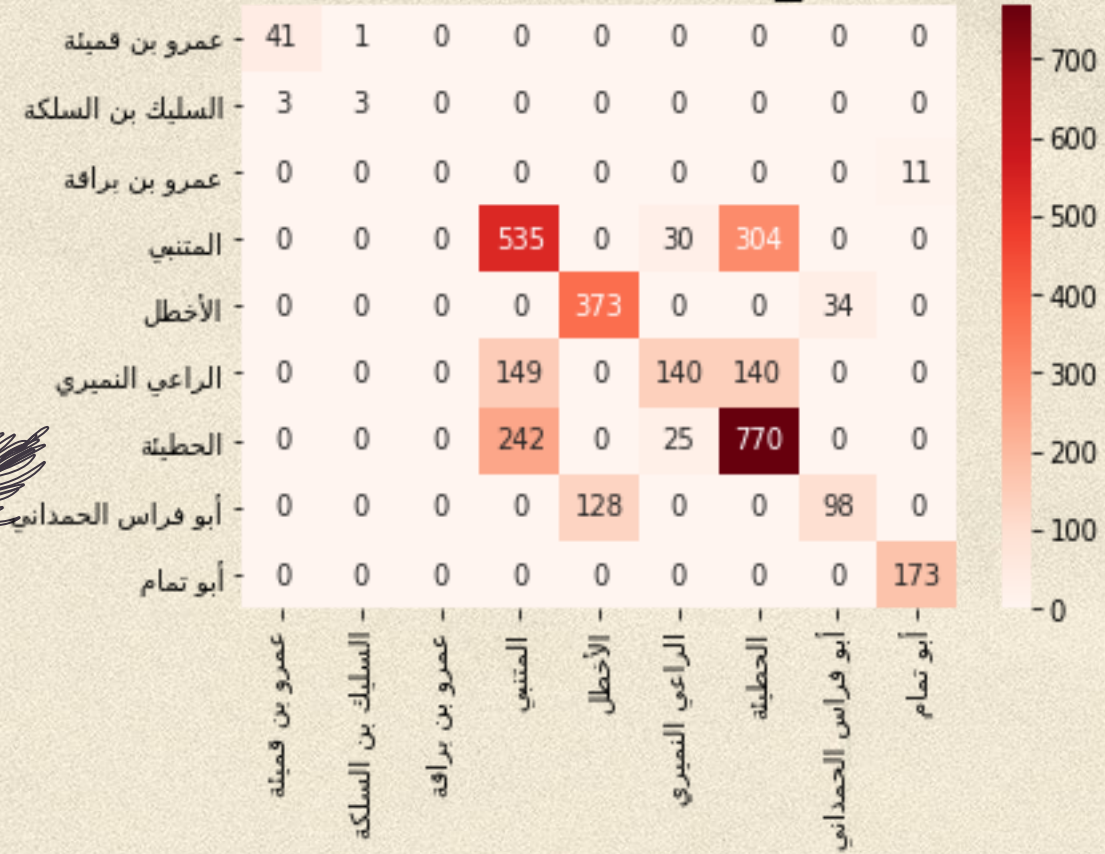
Test XGboost with best parameters1
F1 score= 0.6529368761859843

Test

Comparison all classification algorithms scores

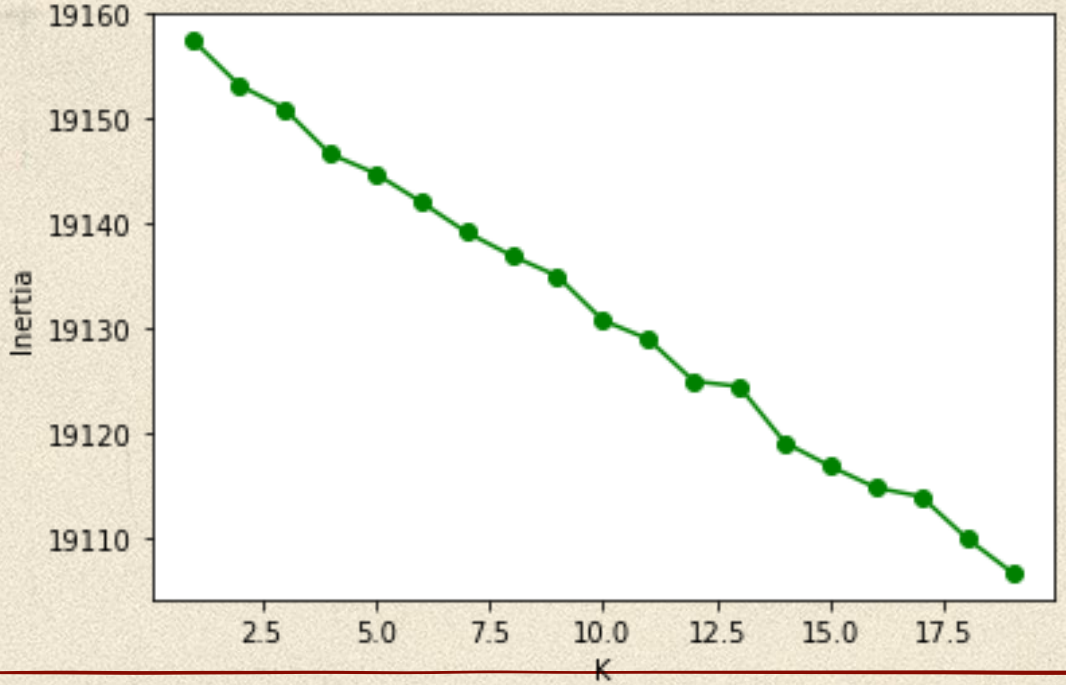
Model	Accuracy
Logistic Regression Algorithm	0.64
Decision Tree Algorithm	0.37
Random Forst Algorithm	0.41
Xgboost Algorithm	0.67

XGboost Mosel to predict each Poet_name with his peoms



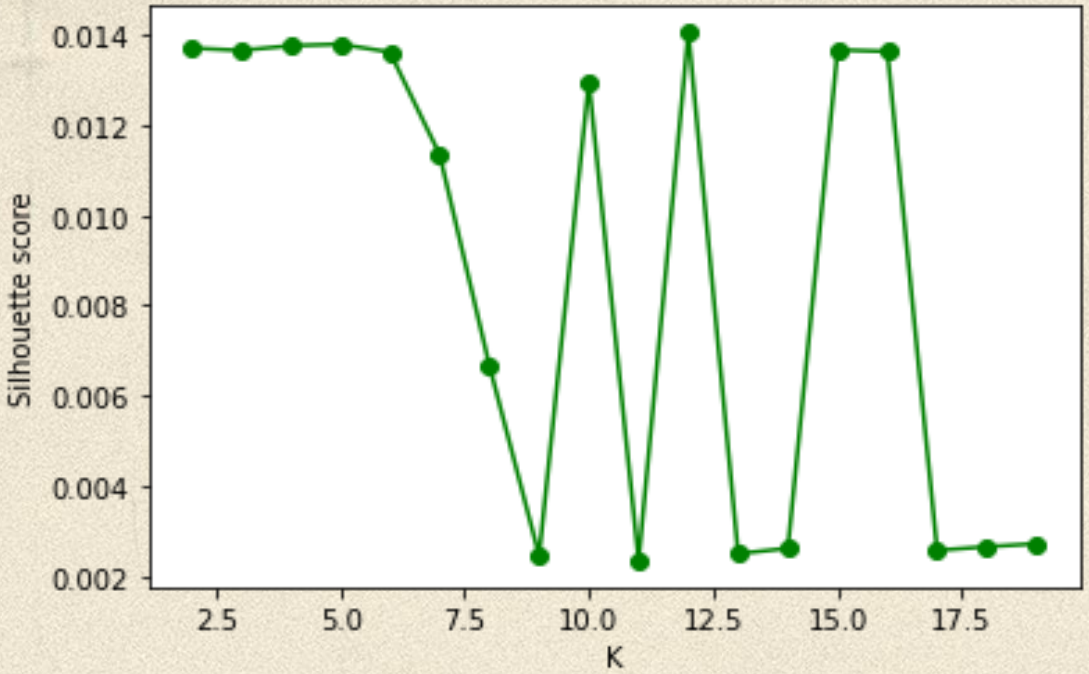
Unsupervised Modeling

- **Kmeans**



Modeling

- **Kmeans**

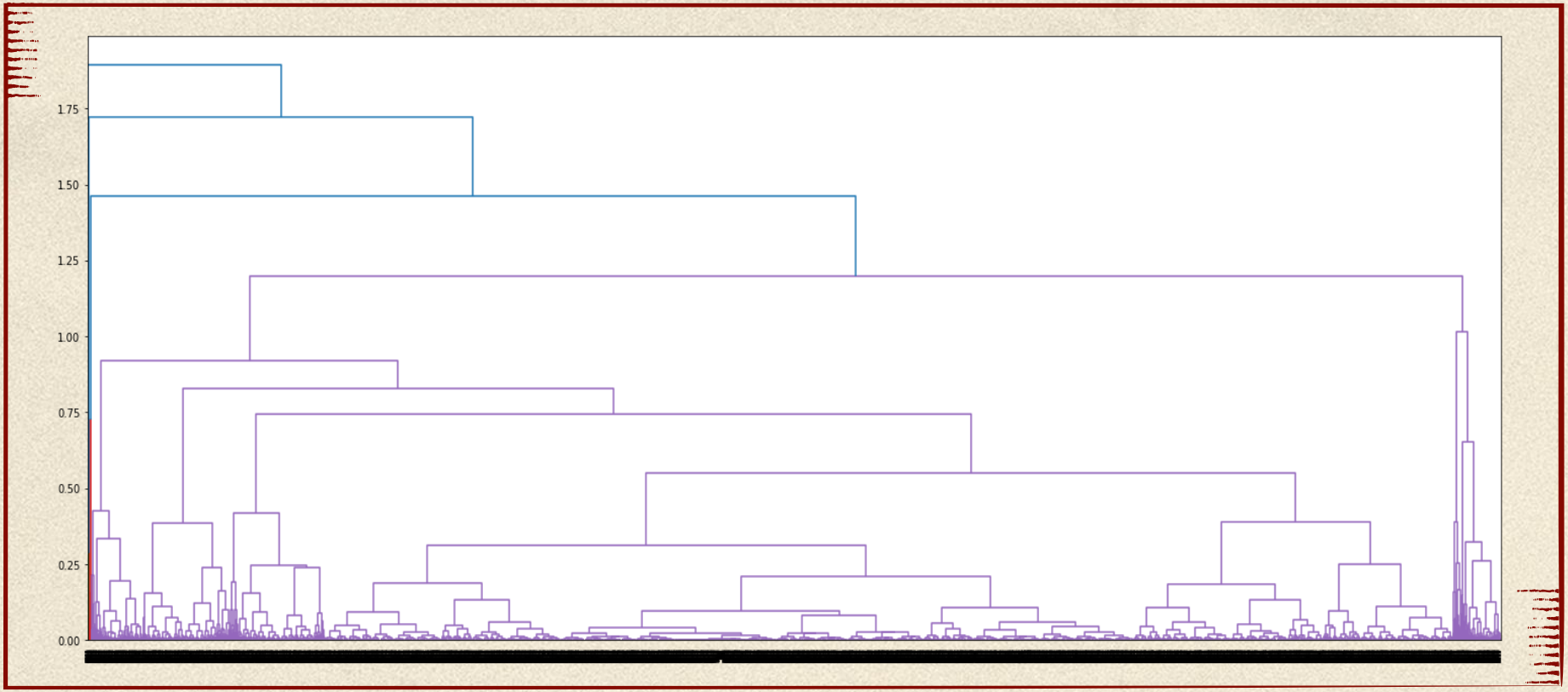


Modeling

- **Hierarchical clustering with different linkages**

Ward	Complete	average
2.83s	2.60s	2.71s

Ward



Conclusion

In attempts to predict the best model , if you enter the poem, the name of the poet will be predicted .
We made several models to determine the best model. .
The best result is **XGBoost .**

THANK YOU