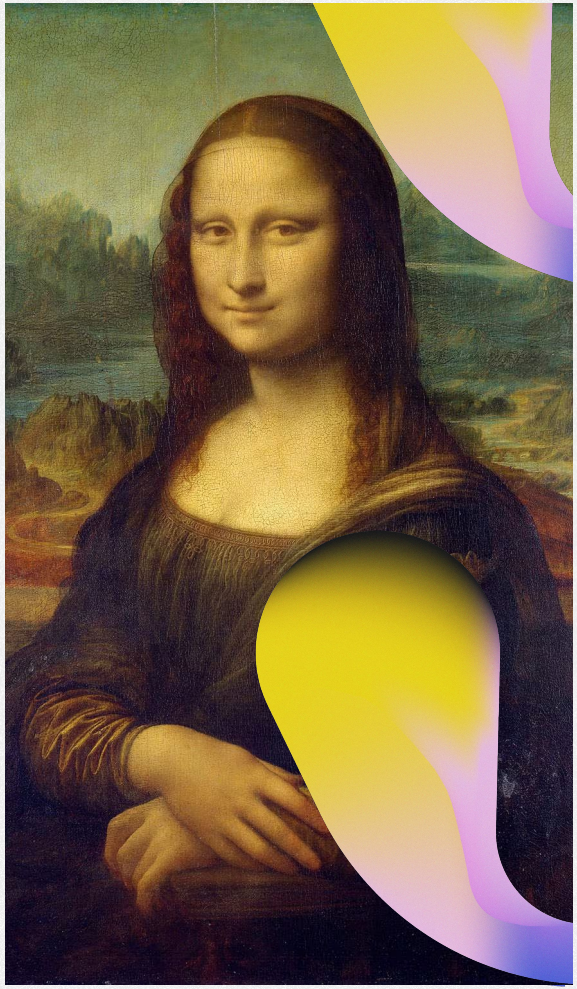
The Evolution of
Themes in Artworks

Annie Lou



Problem Statement

Bridging my passions for art and data science...

- What are the common themes of artworks across history?
- How do the themes of artworks change over time?

Dataset

- Dataset containing 124170 artworks, scraped from wikiart
- Artwork information contains artwork title, artist, style, date, and link to artwork image
- Date ranges from 3050 BC to 2023 AD
- The dataset contains artwork titles with multiple languages

Approach

Use different machine learning methods to analyze trends in *Artwork Title* over time using the *Date* column

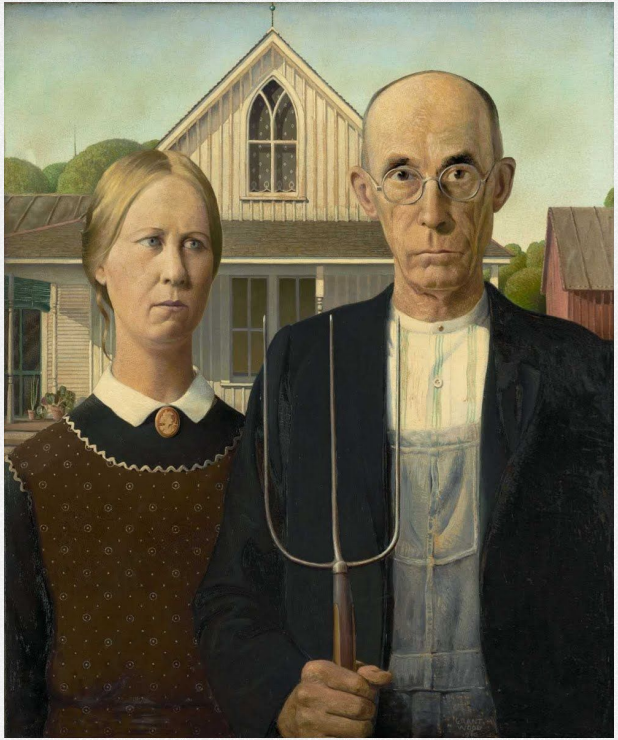
Assumption

The title of an artwork captures the content and theme of the artwork.

Girl With a Pearl Ring
Johannes Vermeer, (1665)

Success Metrics

- Interpretability: the results should enable us to identify clear topics within each time period
- Diversity: topics identified should cover a variety of themes in art history
- Historical Consistence: topics should align with historical trends and events



American Gothic
Grant Wood (1930)

Preprocessing

- **Tokenize:** turn a title into smaller units (words)
- **Lemmatize:** change words into their original forms (i.e. women -> woman)
- **Stopword removal:** remove high-frequent words that contain little meaning (i.e. “the”) in multiple languages

Methods

01

Word Frequency

02

Topic Modeling

03

Dimensionality Reduction + Clustering

01

Word Frequency

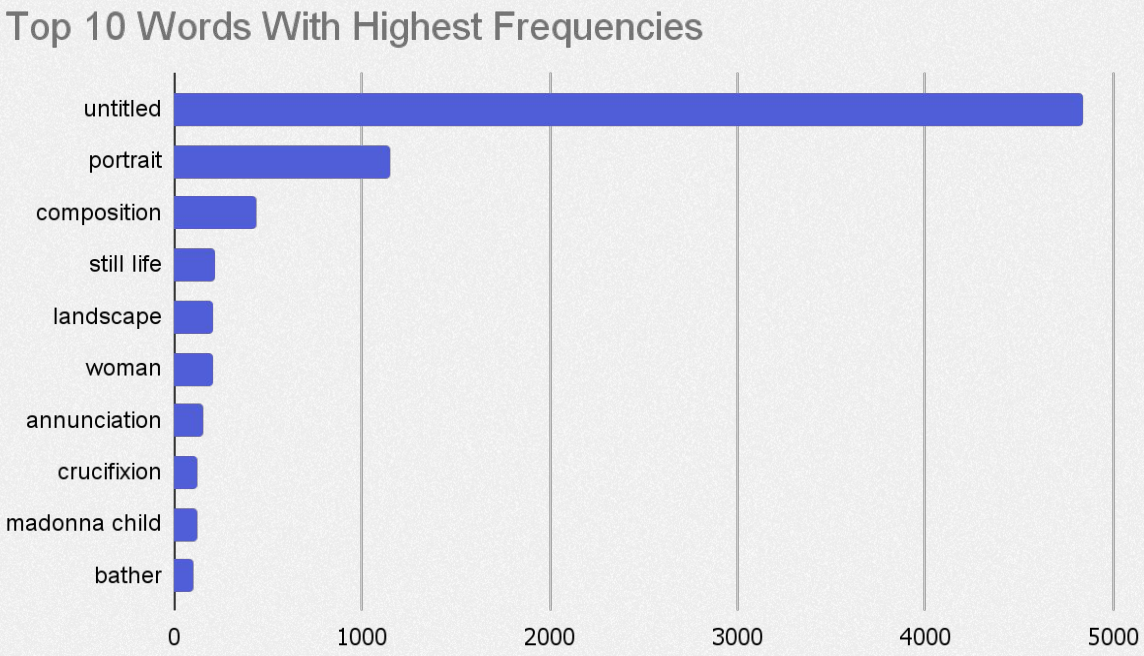
1. Obtained word frequency of each word occurred in artwork titles
2. Found top 10 words with highest frequency across history
3. Analyzed frequency trends of selected words

1. Word Frequency

STEP 1:

Find top 10 words with highest frequency across history

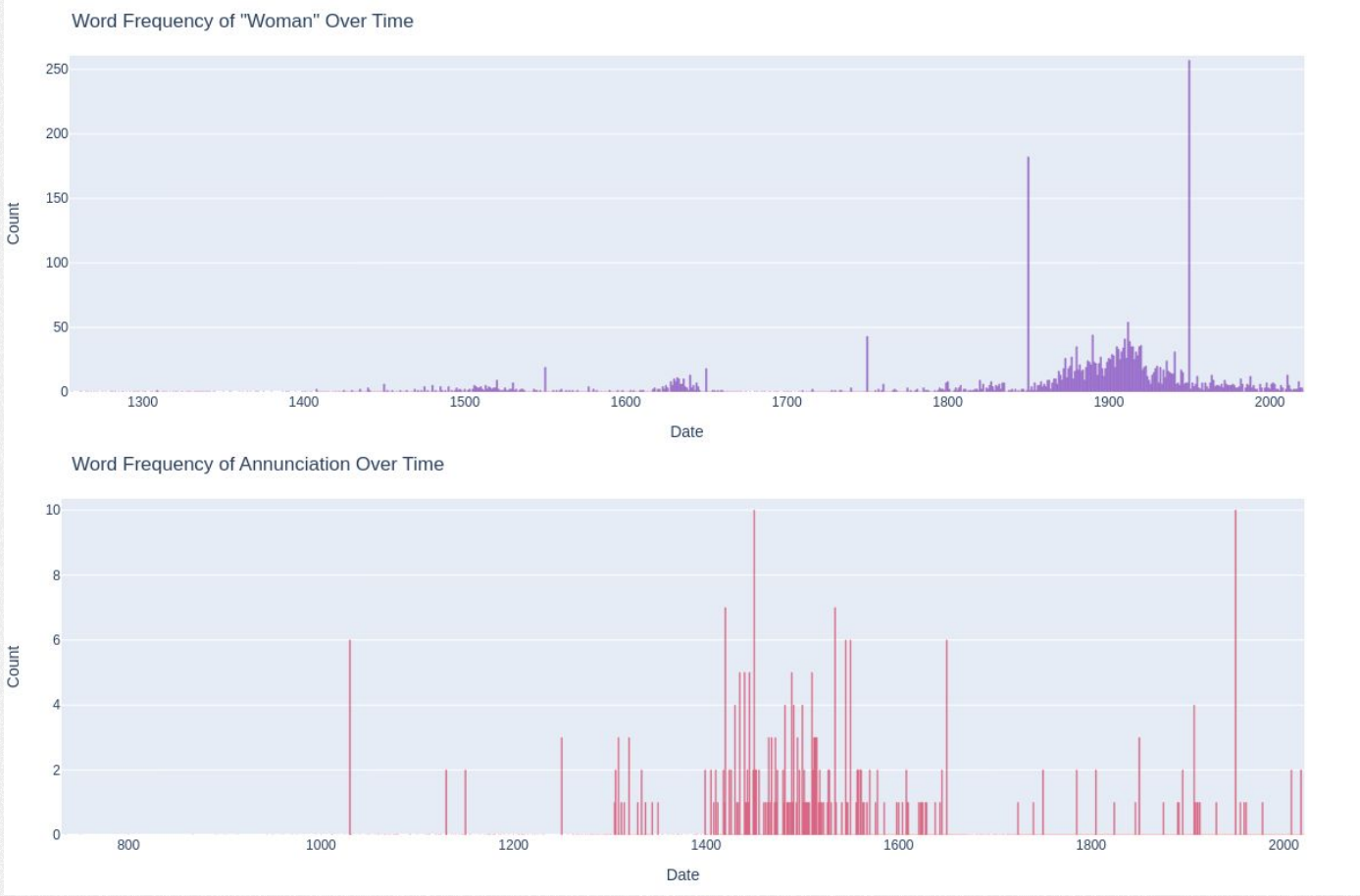
Note: The text has been lemmatized, so the results account for different forms of a word. For example, "woman" in the results could represent both "woman" and "women" from the original title.



1. Word Frequency, continued

STEP 2:

Analyze word frequency over time.

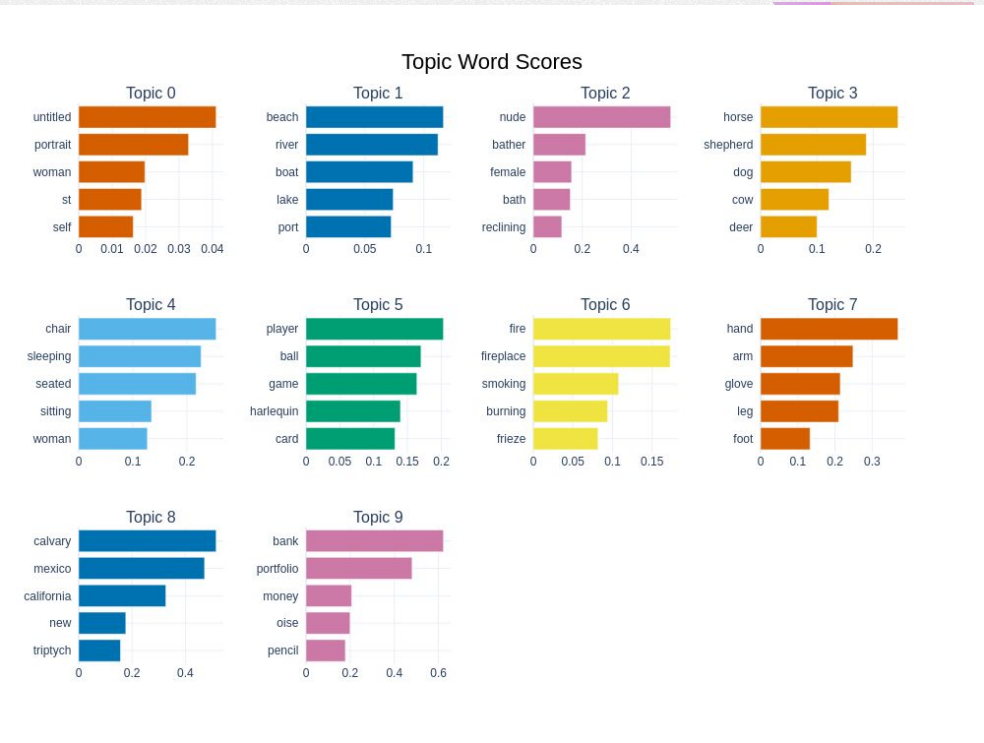


Word frequency analysis offers insights into the prevalence of words in artwork titles over time. However, it does not capture all semantically related terms (i.e. Mona Lisa is also a woman).

This is why we need topic modeling.

2. Topic Modeling

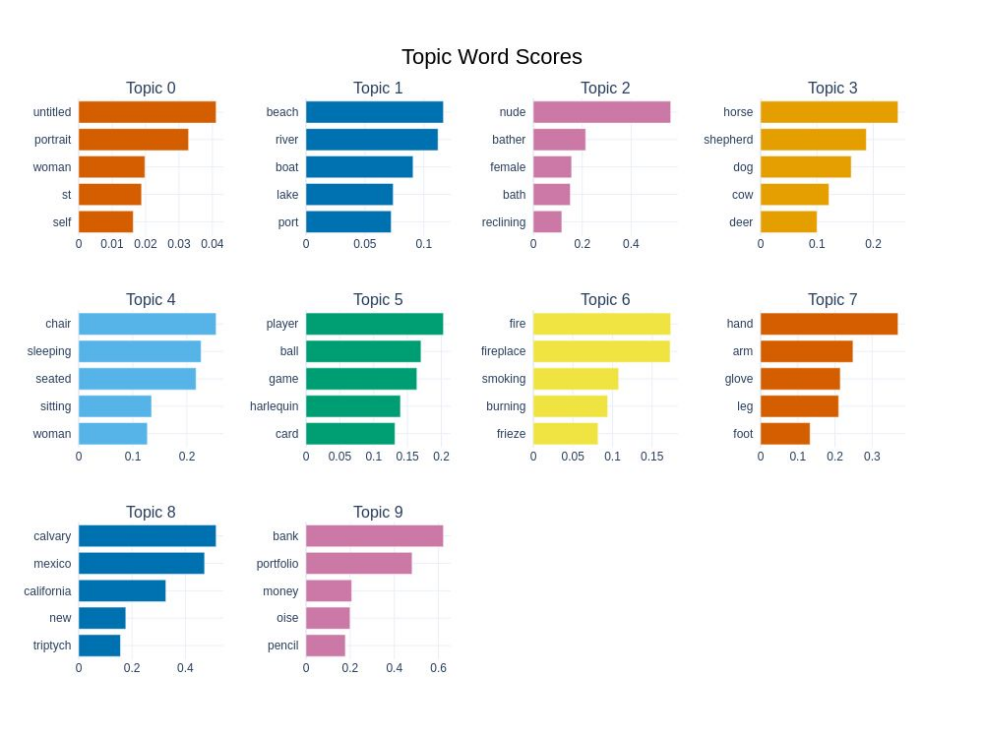
- BERTopic is a tool that turns each title into embeddings (vectors of real numbers) that represent semantic information. Titles with similar meanings tend to have similar vectors. BERTopic then groups these embeddings to find patterns, and label each group with the words that best describe the topic of that group.
- I used BERTopic to perform topic modeling on all 124170 artwork titles.
- I repeated the process on artwork titles within specific time periods.



Example visualisation of topics discovered by BERTopic

2. Topic Modeling

- BERT is an expensive model and the data set is very big.
- Running the BERTopic model on the entire data set crashed my computer several times, so I used sampling in early analysis to test ideas.
- Once I chose an approach, I used a third party server to run the final analysis on the entire dataset.



Example visualisation of topics discovered by BERTopic

2. Topic Modeling

7 Most Common Themes in Art Across Time
(3050 BCE - 2023 CE)

1
Water Scenes
beach
river
boat
lake
port

2
Nude/Bath
nude
bather
female
bath
reclining

3
Animals
horse
deer
dog
cow
bird

4
Seated/Sleeping Scenes
chair
sleeping
seated
sitting
woman

5
Entertainment
player
ball
game
harlequin
card

6
Fire
fire
fireplace
smoking
burning
frieze

7
Body Parts
hand
arm
glove
leg
foot

2. Topic Modeling

Ancient Period (3050 BCE - 0 CE)

1
Ancient Greek Art
Greece
Reconstruction
Ancient
Stela
fresco

2
Containers
Terracotta
Jar
Amphora
Neck
bowl

3
Egyptian Art
Tomb
Nakht
Nany
nebamun

4
Drinking/Drinking Cups
Water
Hydria
Wine
Jar
Terracotta

5
Agriculture
Scarab
Inscribed
Named
Maatkare
Hatshepsut

Medieval Period (0 CE - 1500 CE)

The Annunciation,
Melchior Broederlam, 1399

1
Gods/Angels
Madonna
Child
Cherub
Sleeping
Seraph

2
Annunciation
Annunciation
Annunciazione
Dyptic
Chokan
Sansui

3
Crucifixion
Crucifixion
Crucifix
Pazzi
Crucified
Diptych

4
Man
Stephen
Jerome
Sebastian
St

Early Modern Period (1500 CE - 1760 CE)

Bust of a Man Wearing a Gorget and Beret,
Rembrandt, 1626

1
Portrait
Self
Portrait
Beret
Gorget
Frowning

2
Religious Buildings
Basilica
Chapel
Ceiling
Sistine
Church

Ceiling decoration Palazzo Vecchio, Florence,
Giorgio Vasari, 1556 - 1558

Note: The Baroque era celebrates wealth, power and status in artworks. Kings, princes, and popes began to prefer to see their own power and prestige celebrated through art than that of God.

2. Topic Modeling

Industrial Revolution (1760 CE - 1914 CE)

1
Portrait
Self
Portrait
Redingote
Yawning
Andre

2
Landscape
Paysage
Landscape
Land
Semmering
Vexin

3
Horse
Horse
Horseman
Horseback
Stable
Racehorse

Note: Landscape paintings gained prominence in the late 18th century with the rise of Romanticism
In the 18th century, horses in artwork flourished in England and then Europe. The 18th century also saw the formation of a school of animal and sporting art.

2. Topic Modeling

World War I (1914) - Present (2023)

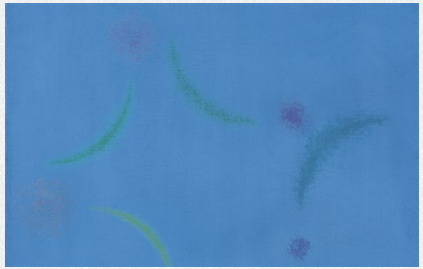
Composition with Pouring,
Jackson Pollock, 1943

1
Animals
Bird
Hawk
Oiseau
Heron
Sparrow

2
People
Ardoise
Ariadne
Elvis
Emma

3
Abstract
Composition
Composicion
Noter
Duele
Beginning

4
Blue
Blue
Blau
Azul
Bleue



Energien im Blau
Max Bill, 1949

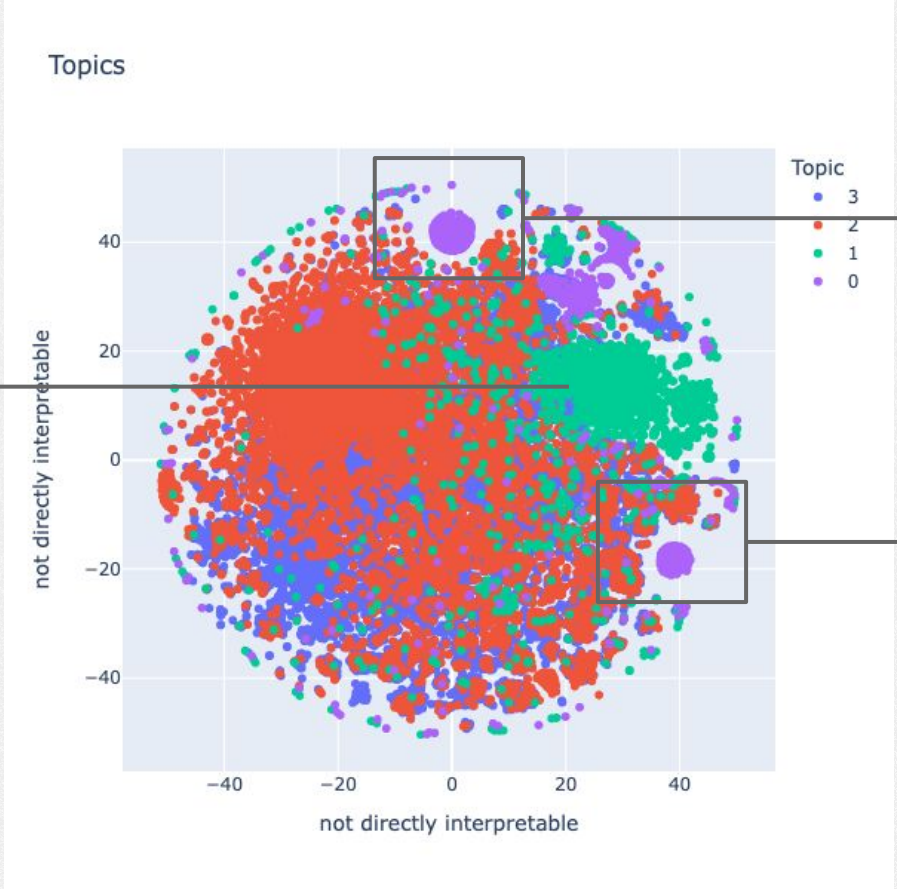
Evaluation

Interpretability	Most topics are interpretable. Mean coherence score = 0.43. Coherence score: a measure of how coherent the words in each topic are.
Diversity	The results showed a diverse range of themes in art.
Historical Consistence	The results are consistent with historical trends and offer more detailed insights.

3. Dimensionality Reduction + Clustering

1. Calculated the embeddings of artwork titles
2. Used TSNE to reduce dimensions of the embeddings (reduced from 512 dimensions to 2 dimensions)
3. Used KMeans to discover clusters in the reduced space

2. Dimensionality Reduction + Clustering



“illustration” and “portrait” cluster

“untitled” cluster

Overall, topic modeling using BERT provided most insights. We can see how common themes in artworks changed over time - from containers and agriculture to religious themes, then to people such as Elvis...

Limitations

Modern art movements (such as conceptual art) made it harder to capture the content and meaning of artworks through their titles

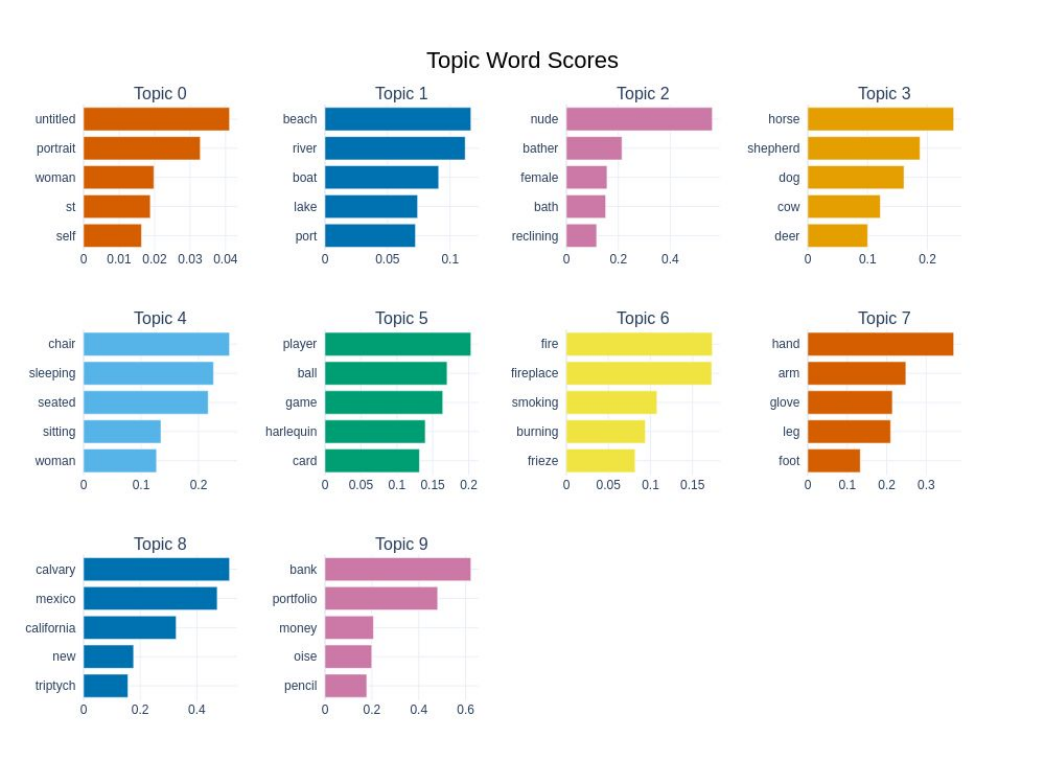
(Topic modeling on contemporary artworks might not return as accurate results as it would on earlier pieces).

Fountain
Marcel Duchamp (1917)

Next Steps and Improvements

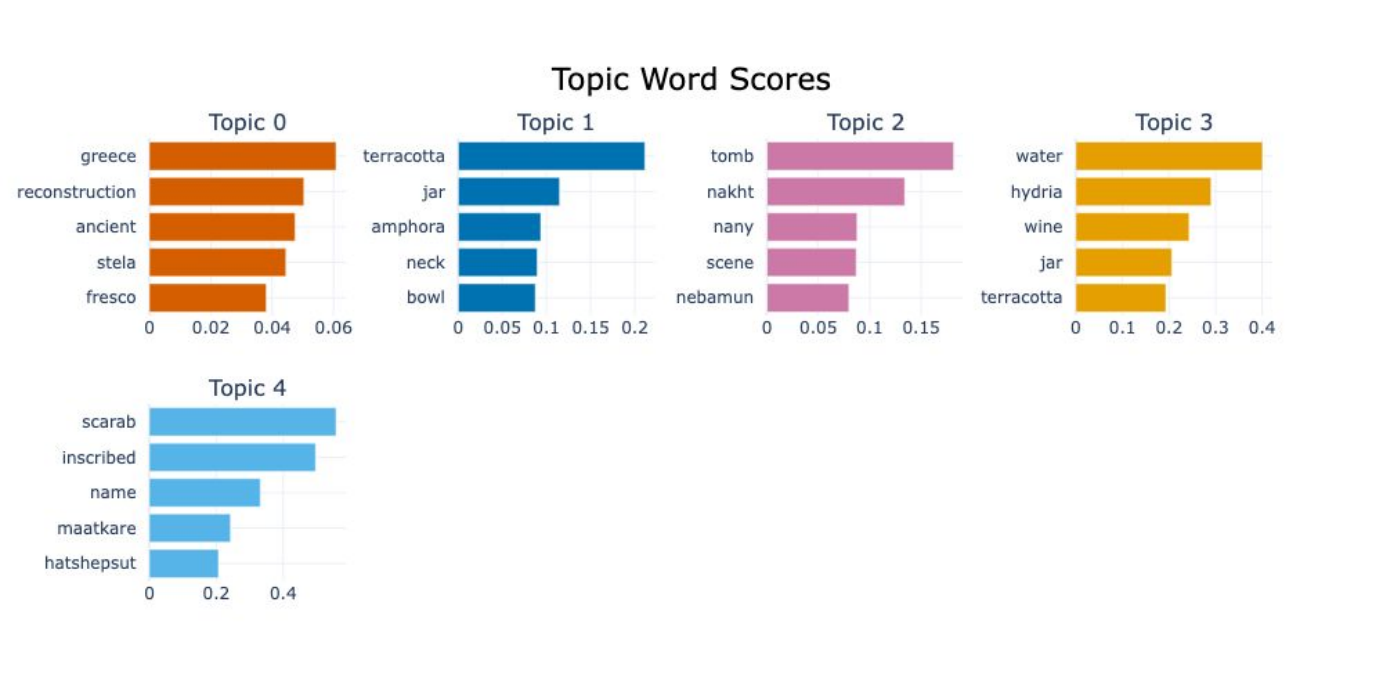
- **More Detailed Descriptions:** Artwork titles provide limited information. Obtaining more detailed descriptions for each artwork will enhance the accuracy of topic modeling.
- **Computer Vision:** The dataset contains links to each artwork image, which offers the potential to use computer vision techniques. This approach could give us deeper insights into the artworks and potentially overcome the limitations of NLP analysis.
- **Better Lemmatization:** I used WordNetLemmatizer to lemmatize the artwork titles, but it is not perfect. During topic modeling, I noticed that some words were not lemmatized, such as “bathing” and “bath”, “received” and “receiving”. Improving lemmatization in future steps could yield better results.

Appendix


Topic Word Scores
Topic 0
untitled
portrait
woman
st
self
0 0.01 0.02 0.03 0.04
Topic 1
beach
river
boat
lake
port
0 0.05 0.1
Topic 2
nude
bather
female
bath
reclining
0 0.2 0.4
Topic 3
horse
shepherd
dog
cow
deer
0 0.1 0.2
Topic 4
chair
sleeping
seated
sitting
woman
0 0.1 0.2
Topic 5
player
ball
game
harlequin
card
0 0.05 0.1 0.15 0.2
Topic 6
fire
fireplace
smoking
burning
frieze
0 0.05 0.1 0.15
Topic 7
hand
arm
glove
leg
foot
0 0.1 0.2 0.3
Topic 8
calvary
mexico
california
new
triptych
0 0.2 0.4
Topic 9
bank
portfolio
money
oise
pencil
0 0.2 0.4 0.6

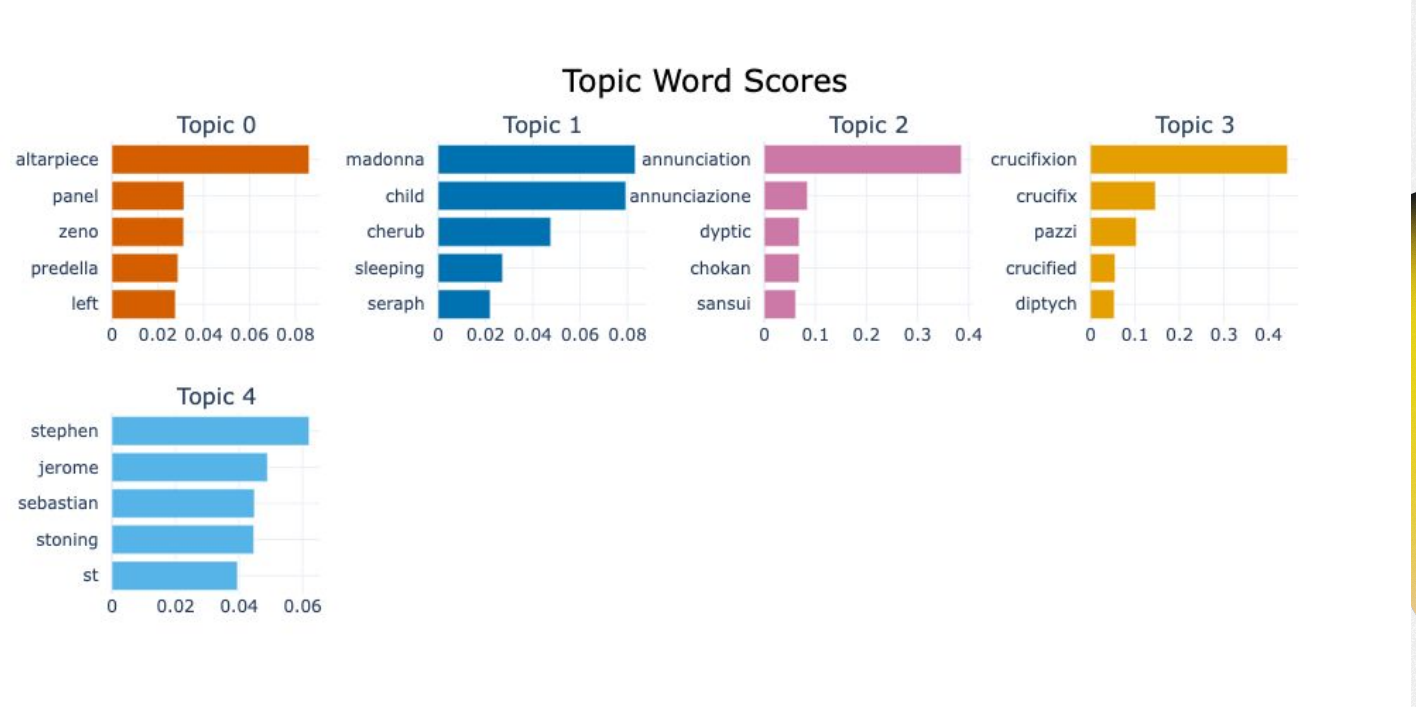
Appendix

Ancient Period (3050 BCE - 0 CE)



Appendix

Medieval Period (0 CE - 1500 CE)

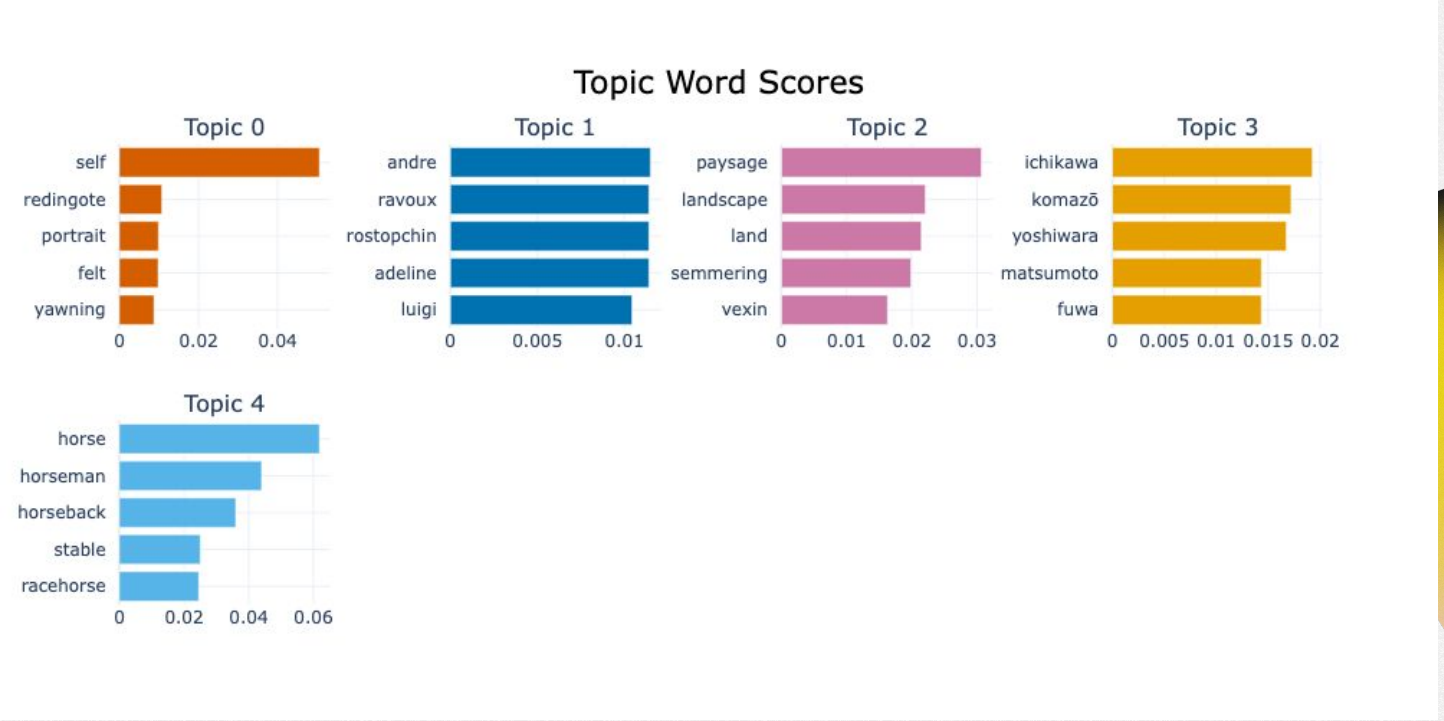


Appendix

Early Modern Period (1500 CE - 1760 CE)

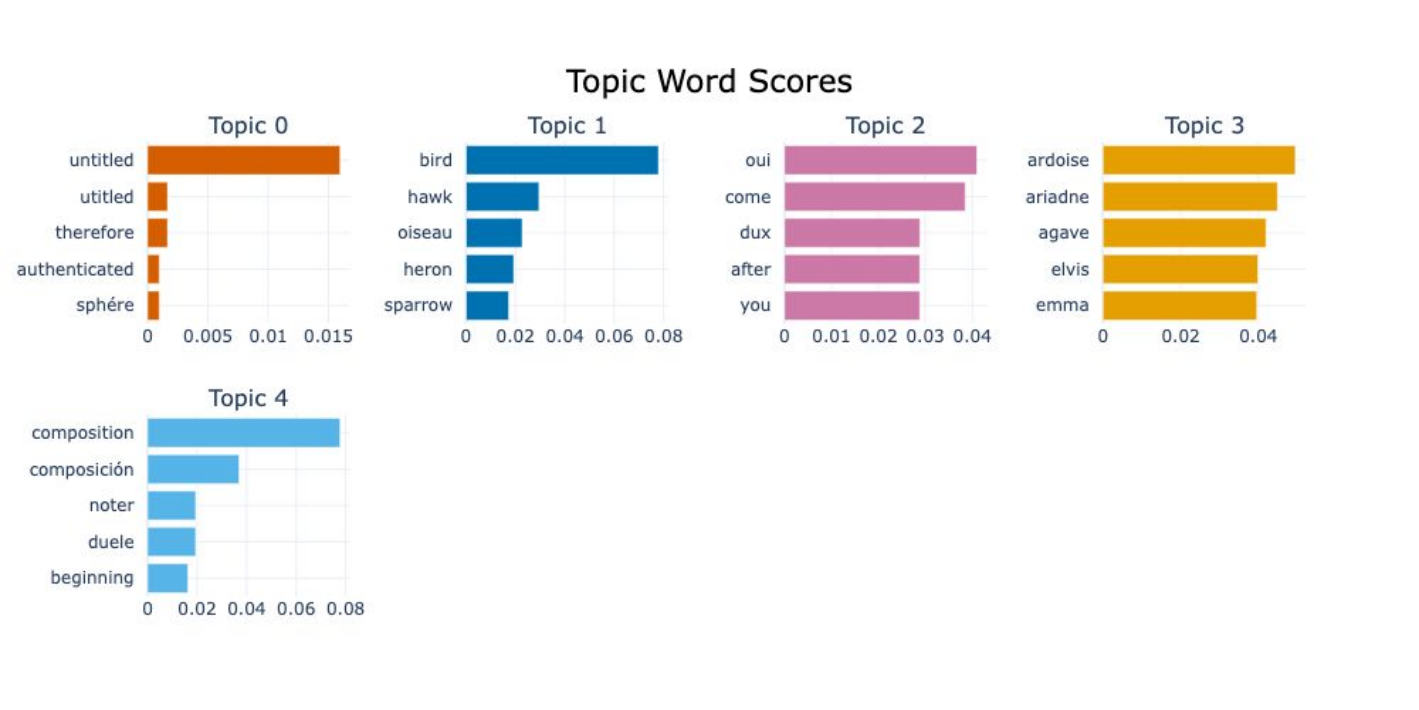
Appendix

Industrial Revolution (1760 CE - 1914 CE)



Appendix

WWI (1914) - Now (2023)



Appendix

Dimensionality Reduction + DBSCAN