EN.601.666 Information Retrieval and Web Agents
Spring 2022
Whiting School of Engineering, Johns Hopkins University

Final Project Report

Title

Artwork Recommendation Search Engine

Summary

In this project, we constructed a multilingual search engine among four different museums, which allows users to query art based on different criterias, such as title, artist, etc. Based on the users query content, we also set up a content-based recommendation system to provide more similar artworks the users might be interested in.

Project Interface URL : https://museum-art-recommendations.herokuapp.com/

Data Link: https://drive.google.com/file/d/1NqevDWMXGMIipWJRPjNw3qrDadFQbHcF/view?usp=sharing

Team Members

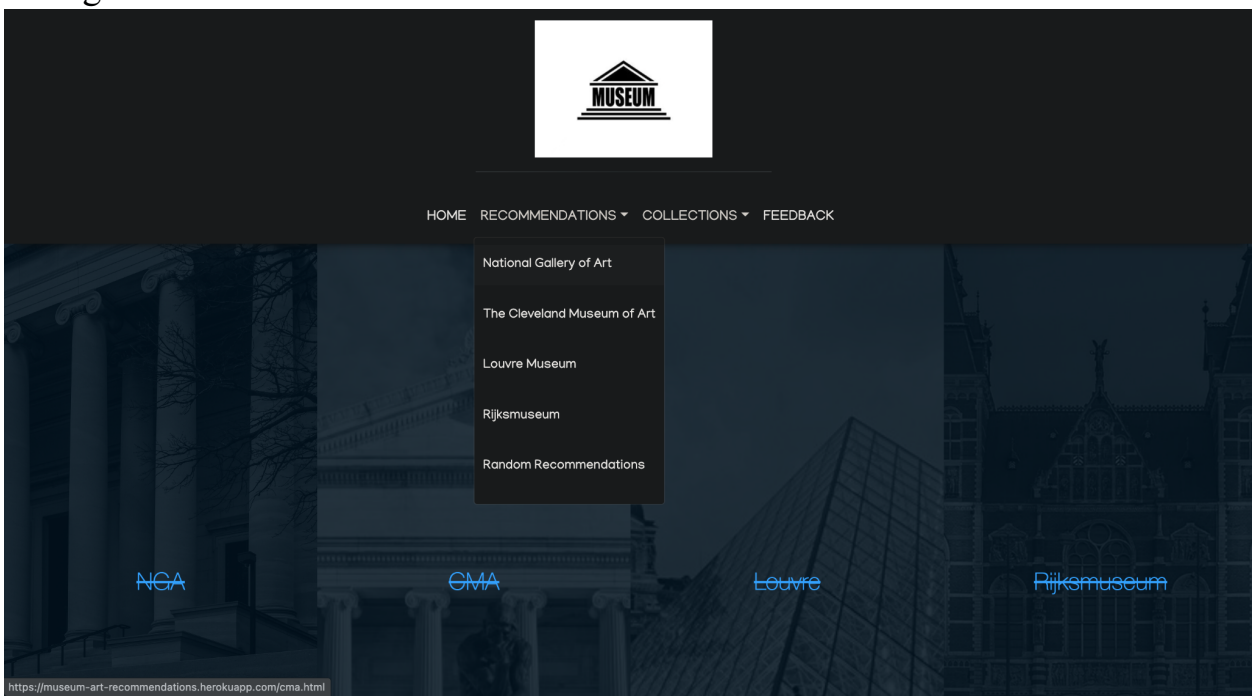
Minghao Xue *mxue7@jhu.edu*

Yawen Hu *yhu101@jhu.edu*

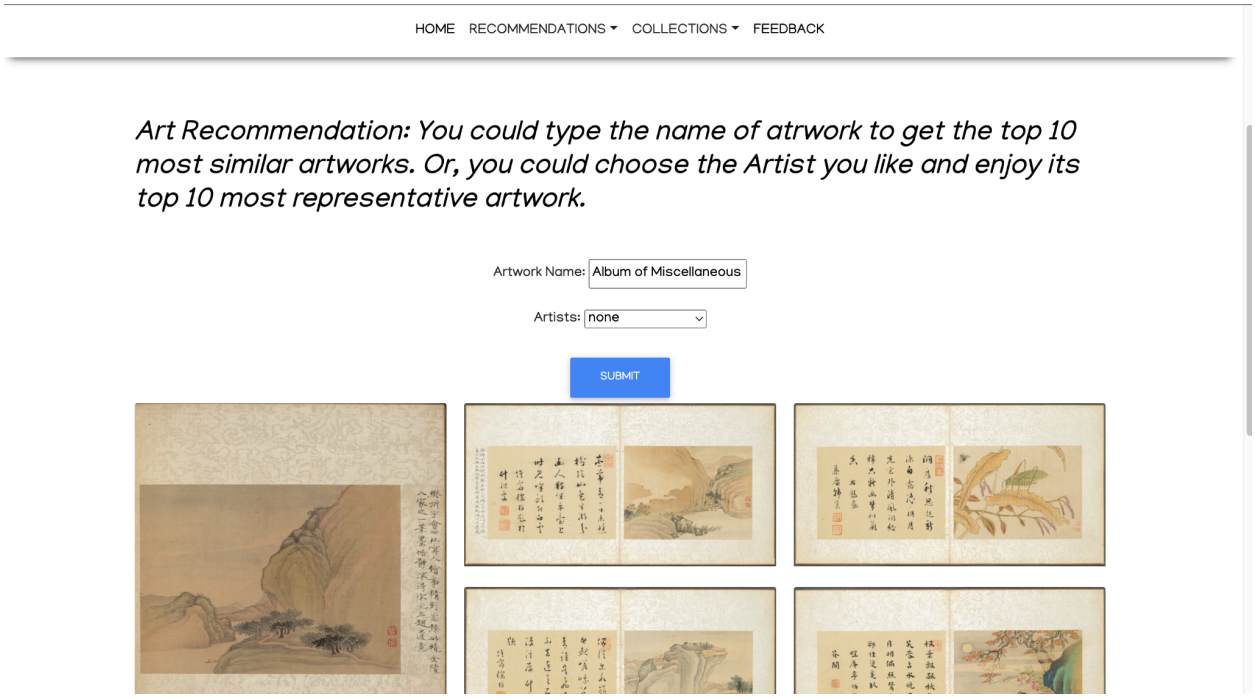
I. GUIDE TO WEB INTERFACE & SCREEN SHOTS

DEMO URL:https://youtu.be/Vt9wkWDT6r4

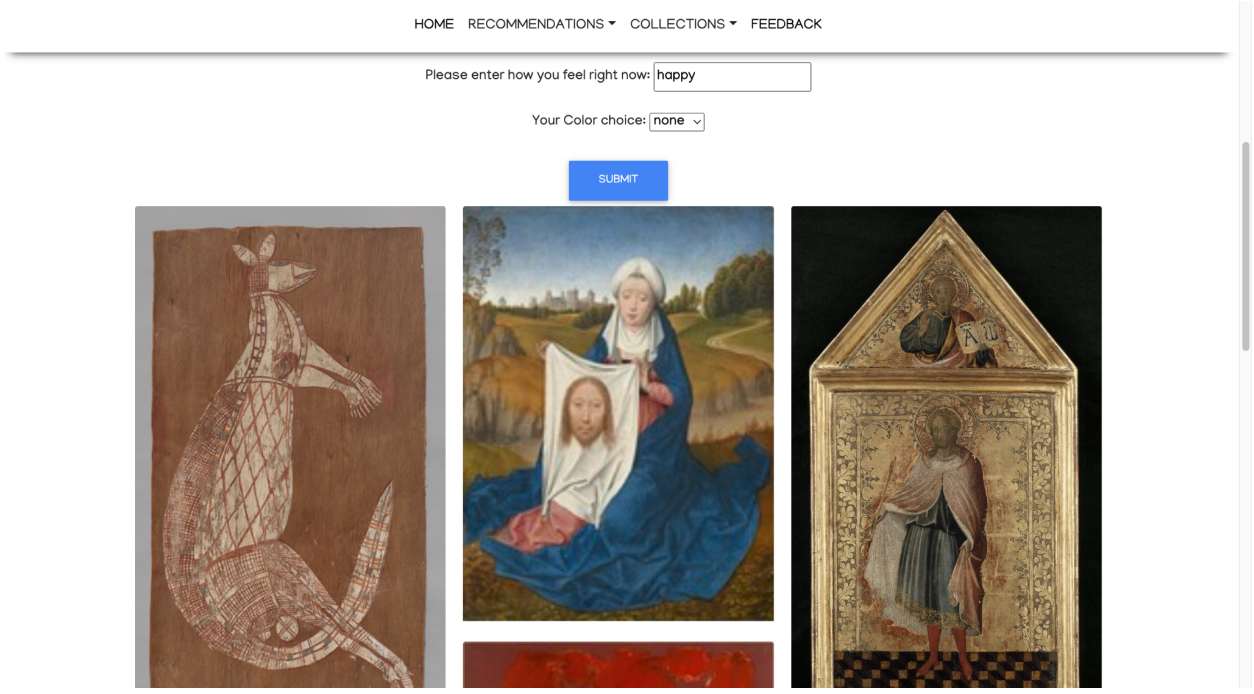
Five functional features in RECOMMENDATIONS. First 4 features provide an art recommendation function in a single museum and the last feature presents artworks among all four museums.



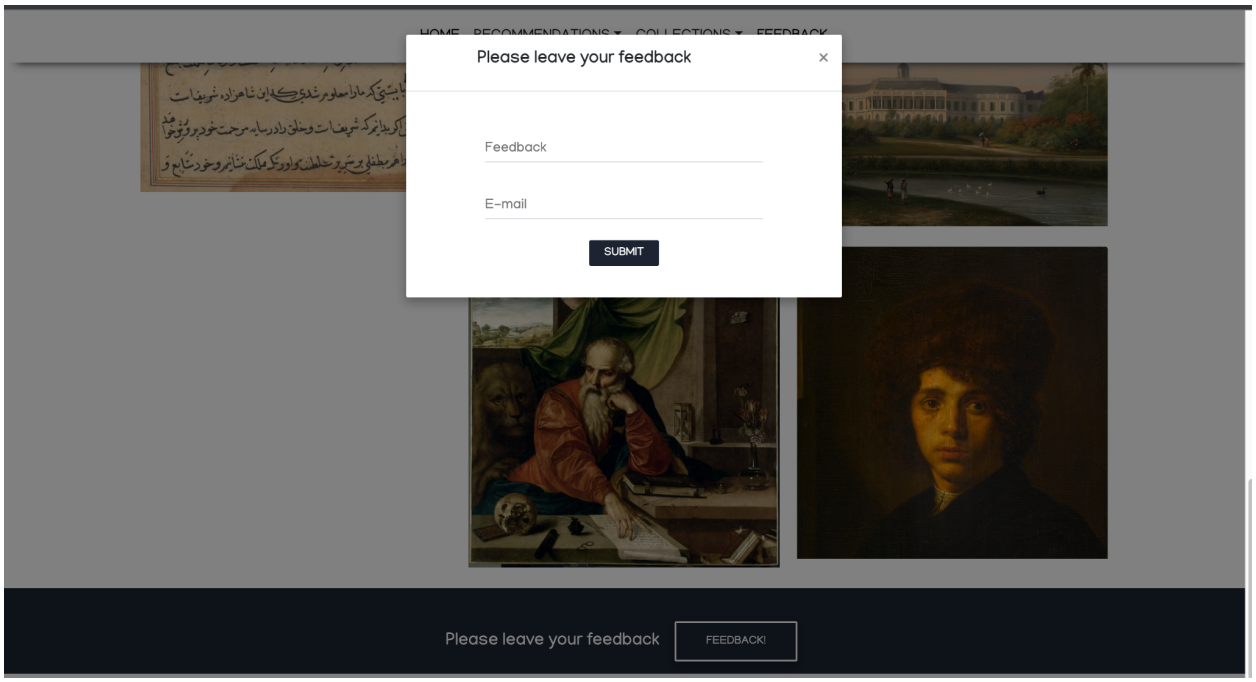
Users could type artwork name or select artists name to obtain 10 recommended artwork in the first four features. You have to make sure Artist is none when you only input the artwork name, otherwise it will return the top 10 artist's artworks.



Users could type their feeling or select color name to obtain 10 recommended artwork in the random recommend features. You have to make sure color is none when you only input the artwork name, otherwise it will return 10 artworks with selected color among all artworks.



Users could leave the feedback to us. It is on the bottom in every page.



II. ACHIEVEMENTS & METHODS

1. Multilingual artwork search:

To achieve the multilingual search function, we translated the Louvre Museum data from French to English and the Rijks Museum data from Dutch to English.

The translating part was achieved by the using deep-translator package, which supports multiple different translation platform, and we specifically used the GoogleTranslator API in it. For short sentence features, we performed phrase to phrase translation. And for the description part, we performed word to word translation.

2. Content-based Artwork Recommendation:

The recommendation system was built based on the NLP text features generated with artwork metadata. For preprocessing, we removed all punctuations and stopwords from the data and converted to stemmed version to get a cleaned text data. We then built up a term matrix including all n-grams with size under 3 and weighted by TF-IDF score. To capture semantic features, we used the word2vec pretrained model to get the word2vec embedding feature for each word. And combining the TF-IDF features and word2vec features, we used them to

compute the cosine similarity, which gave us a representation of the similarity between two artworks. And based on the similarity score, we generated the final recommendation list.

So every time we input a query and got the matched result artwork, we can also generated a list of recommended artworks.

3. **Search with color themes:**

We also developed function that allows users to search with different colors and return result based on the main colors of the artwork images. This method depends on the visual features in the art images. By transforming the original image into the RGB space, we performed a K-means clustering method to cluster it pixel RGB values into different clusters and get the corresponding centroid RGB value. We use the centroid RGB values to represent the main colors of that artwork. And by comparing the main colors with our predefined colors, we got the its color label.[1]

4. **Web Interface and web crawling:**

The three achievements we mentioned above are held on the web interface. We use Django as our web interface framework and idea of this work is inspired by The Online Museum.[2]. There is an interesting feature in our web which is called *random recommendations.* Users could type their current thoughts or choose color they like, the engine will return 10 paintings that are related to your input. We achieve this function by using semantic analysis and image analysis above.

We also develop a web crawler to crawl raw image urls and related information since most of the urls in meta valid are out of date. We make sure all links are valid by cleaning the data.

III. LIMITATIONS

1. **Translation methods:**

We currently used the word to word translation for long narrative part in description, which somehow can cause semantic meaning losing in the translated version. If we have more time, we can find better method to perform phrase to phrase translate.

2. **Visual features:**

Generate more representative visual features from the image data and include them into our recommendation system. As the format of our data are pretty messy and it took us a long time to validate and retrieve the image source of the artworks, we didn't have time to include more visual features into our algorithm. If we have more time, we could use better feature extracture methods, such as DNN, to improve this part further.

IV. EVALUATION

Without user feedback, we currently do not have an effective method to evaluate our results. However, this can be evaluated after we have users access our current platform. We can get user feedback with multiple different methods and metrics, such as click through rate, clicks of like or dislike. With the users feedback, our recommendation system and search engine can be evaluated using different measurements including precision, recall, F-measurements, etc. We have a feedback option in the bottom of the page, users can leave comments to our team and we are able to learn and improve from people's suggestions.

V.DATA SOURCE

MOMA:https://github.com/MuseumofModernArt/collection
NGA: https://github.com/NationalGalleryOfArt
Louver:https://collections.louvre.fr/en/recherche?typology%5B0%5D=4
Rijksmuseum:
https://figshare.com/articles/dataset/Rijksmuseum_Challenge_2014/5660617

VI. REFERENCE

[1] Bhanot, Karan. "Color Identification in Images. Using colors from an Image as a method… | by Karan Bhanot." *Towards Data Science*, 17 December 2018, https://towardsdatascience.com/color-identification-in-images-machine-learning-application-b26e770c4c71. Accessed 17 May 2022.

[2]Druskat, S. (n.d.). *Citation file format (CFF)*. Citation File Format (CFF). Retrieved May 17, 2022, from https://citation-file-format.github.io/