Prado PageRank Project - Comprehensive Report

1. Introduction

This project uses the PageRank algorithm to determine the relative importance of pictures in the Prado Museum collection.
Tags associated with each artwork are treated like "links" between images, allowing us to construct a graph structure.

The PageRank score for each image helps us identify which images are most 'central' or connected in the tag-based network.

2. Data Pipeline

- Data Loading: Reads CSV, filters columns, drops nulls/duplicates.
- Preprocessing: Splits tags and constructs picture-to-tag and tag-to-picture mappings.
- Graph Construction:
  - Unweighted: Connects pictures sharing tags with simple edges.
  - Weighted: Assigns edge weights based on the number of shared tags.
- PageRank Calculation: Uses NetworkX's PageRank to score nodes (pictures).
- Strategy Comparison: Compares results of both graph types.
- Visualization: Plots graph of pictures with spring layout.

3. Code Structure

Class PradoPageRank:
- __init__(file_path): Initializes storage and input.
- load_data(): Optimized data loading and exploration.
- preprocess_data(): Explodes and maps tags to pictures.
- build_graph(strategy): Constructs graph ('unweighted' or 'weighted').
- compute_pagerank(weight): Calculates PageRank using NetworkX.
- compare_strategies(): Compares PageRank outputs across strategies.
- validate_and_visualize(): Displays top images and plots the network.

Important Variables:
- self.df: Main DataFrame after preprocessing.
- self.tag_to_pictures: Dict for tag-based image mapping.

- self.graph: NetworkX Graph or DiGraph.

- self.pagerank_scores: Final PageRank scores per image.

4. Visualization & Diagram

PRADO PAGERANK

OVERVIEW

PageRank was performed on a graph of pictures from the Prado Museum, analyzed by shared work tags. The objective was to determine the most influential pictures within the collection.

FUNCTIONS

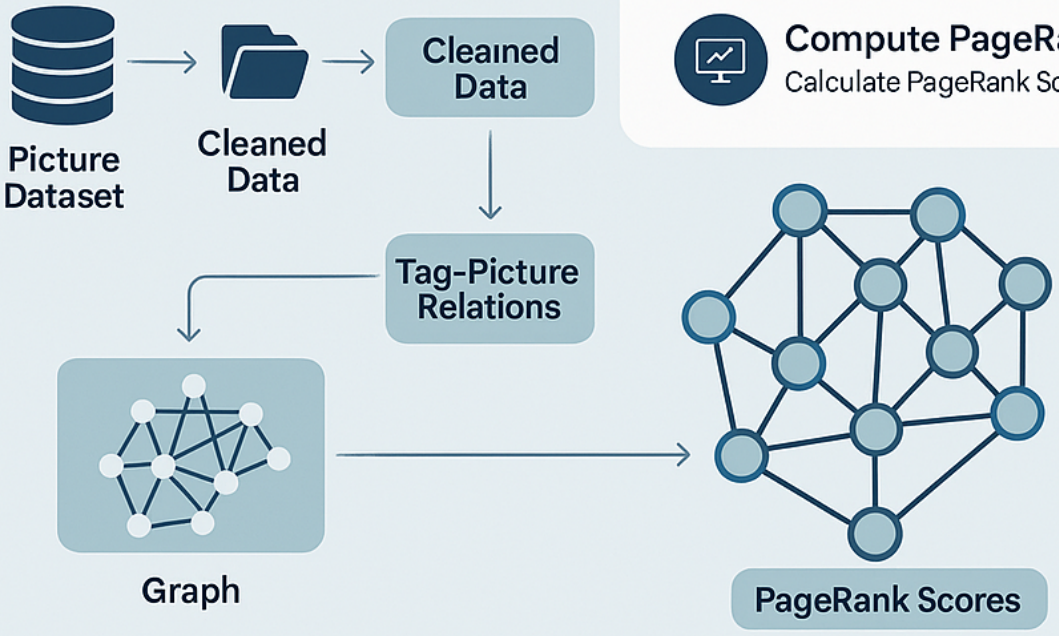
Load Data
Load and clean dataset

Preprocess Data
Organize pictures by work tags

Build Graph
Create graph, with weighted and unweighted options

Compute PageRank
Calculate PageRank Scores

WORKFLOW



EVALUATION

Compared scores using different graph-building strategies and validated the results.

RESULTS

Identified the top-ranking pictures in terms of influencee and importance.

5. Conclusion

By applying the PageRank algorithm to a graph of artworks connected by shared tags, we can identify key pieces within the Prado Museum dataset.

This technique is extendable to recommendation systems, content discovery, and semantic organization.

Future Work:

- Incorporate image similarity (embedding-based).
- Use deep learning for richer tag inference.
- Compare with centrality measures like Betweenness, Eigenvector, etc.