I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work, and including any code produced using generative AI systems. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

Algorithms for massive data: Project 3: Link analysis

Pietro Manning

July 2024

Chapter 1

Loading and Analyzing Artworks Dataset

1.1 Loading and Preparing Data

I loaded the dataset `prado.csv` using `pandas`, ensuring that the `work_tags` column was present in the dataset.

1.2 Graph Construction

I constructed an undirected graph `G` using the `NetworkX` library. I added nodes to the graph representing each artwork, using their URLs as unique identifiers. I created edges between nodes based on shared tags, assuming that artworks with common tags are related.

Chapter 2

Graph Construction and Analysis

2.1 Code for Constructing the Graph

The following Python code was used to construct an undirected graph representing artworks, where nodes represent individual artworks and edges represent shared tags between artworks. The ‘NetworkX‘ library was utilized to build and analyze the graph.

```
import networkx as nx
import pandas as pd

df = pd.read_csv('prado.csv')

G = nx.Graph()

# Add nodes (pictures)
for work_url in df['work_url']:
    G.add_node(work_url)

# Add edges based on shared tags
tag_to_works = {}

for idx, row in df.iterrows():
    work_url = row['work_url']
    tags = str(row['work_tags']).split(',')
    for tag in tags:
        tag = tag.strip()
        if tag not in tag_to_works:
            tag_to_works[tag] = []
        tag_to_works[tag].append(work_url)

for tag, works in tag_to_works.items():
    for i in range(len(works)):
        for j in range(i + 1, len(works)):
            G.add_edge(works[i], works[j])
```

Listing 2.1: Graph Construction Code

2.2 Explanation

The code begins by importing the necessary libraries, `networkx` for graph operations and `pandas` for data manipulation. The dataset is loaded from a CSV file named `prado.csv`.

The graph `G` is initialized as an undirected graph. Nodes are added to the graph for each artwork using their URLs as unique identifiers.

Edges are added between nodes based on shared tags. A dictionary `tag_to_works` is used to map each tag to the list of artworks associated with it. The code iterates over each row in the dataset, extracting the URL and tags for each artwork. Tags are split and stripped of extra whitespace before being added to the dictionary.

2.3 Graph Visualization

To visualize the graph, I used the `nx.draw()` function from `NetworkX` along with `matplotlib.pyplot`, customizing the appearance of nodes and edges for clarity.

2.4 Basic Graph Analysis

I analyzed the graph by calculating and printing the total number of nodes and edges to understand its size, determining the node with the highest degree to find the artwork with the most connections, and identifying the number of connected components to reveal how many separate subgraphs exist within the overall graph.
-Number of nodes in the graph: 13487
-Number of edges in the graph: 129908
-Node with the highest degree: https://www.museodelprado.es/coleccion/obra-de-arte/el-nio-del-arbol/6469ab32-b795-4bc5-96b3-439fab6409e0 with a degree of 264
-Number of connected components in the graph: 8137
If the connected components are few, we can expect a very sparse network. In a sparse network, many nodes will have few or no connections. This can lead to a generally low PageRank for most nodes, as PageRank is influenced by the number of incoming and outgoing links.

2.5 PageRank Calculation

Finally, I computed PageRank scores for each artwork to rank them by their relative importance within the graph and printed the top 10 artworks with the highest PageRank scores to highlight the most influential artworks in terms of connectivity. This process allowed me to analyze which artworks are most interconnected based on shared tags, providing insights into the relationships and importance of different artworks within the dataset from the Museo del Prado. Here is the Python code used for calculating the PageRank:

```
import numpy as np

def pagerank(M, num_iterations=100, d=0.85, tol=1.0e-6):
    """
    Compute the PageRank of each node in the graph.

    Parameters:
    M (numpy array): Adjacency matrix where M[i][j] represents a
    link from node j to node i.
    num_iterations (int): Maximum number of iterations to run the
    algorithm.
    d (float): Damping factor, usually set to 0.85.
    tol (float): Tolerance for convergence.

    Returns:
    numpy array: PageRank vector
    """
    N = M.shape[1]  # Number of nodes in the graph
    v = np.random.rand(N, 1)
    v = v / np.linalg.norm(v, 1)  # Normalize initial vector
    M_hat = d * M + (1 - d) / N * np.ones((N, N))

    for _ in range(num_iterations):
        v_new = M_hat @ v
        if np.linalg.norm(v_new - v, 1) < tol:
            break
        v = v_new

    return v
```

Listing 2.2: PageRank Calculation Algorithm

2.6 Explanation

The provided Python code implements the PageRank algorithm, which is used to rank the importance of nodes in a graph. The algorithm operates on an adjacency matrix `M` where the element `M[i][j]` represents a link from node `j` to node `i`. The PageRank algorithm is an iterative method that computes the relative importance of nodes based on their link structures.

The function `pagerank` takes four parameters: the adjacency matrix `M`, the maximum number of iterations `num_iterations`, the damping factor `d`, and the convergence tolerance `tol`. The damping factor, commonly set to 0.85,

accounts for the probability that a user will randomly jump to any node rather than following links.

Initially, a random PageRank vector `v` is created and normalized. In each iteration, the algorithm updates the PageRank values using the matrix `M_hat`, which incorporates the damping factor. The iteration continues until the change in the PageRank vector is less than the tolerance `tol`, indicating that convergence has been reached.

After the iterations are complete, the final PageRank vector `v` is returned, representing the PageRank scores of the nodes in the graph.

2.6.1 Scalability Features

The `pagerank` function incorporates several features that contribute to its scalability, which are explained as follows:

- **Efficient Matrix Operations**:
  - The function uses the `@` operator for matrix multiplication, which is optimized in the NumPy library. NumPy provides high-performance computations for matrix operations, leveraging C-based implementations and optimized algorithms for large-scale matrix manipulations.
- **Memory Efficiency**:
  - NumPy arrays are designed for memory-efficient storage and manipulation of data. This efficiency is crucial for handling large adjacency matrices, ensuring that the function can work with extensive graphs without excessive memory consumption.
- **Convergence Criteria**:
  - The `tol` parameter sets a tolerance level for convergence, which allows the algorithm to terminate early when the PageRank vector stabilizes. This approach avoids unnecessary computations and reduces the number of iterations required for convergence.
- **Damping Factor**:
  - The `d` parameter, also known as the damping factor, controls the probability of following a link versus randomly jumping to any page. This factor helps stabilize the PageRank computation and ensures that the algorithm converges more effectively.
- **Sparse Matrix Handling**:
  - Although the function does not explicitly handle sparse matrices, the adjacency matrix `M` can be represented as a sparse matrix using libraries such as SciPy (`scipy.sparse`). This representation is beneficial for very large graphs, as it reduces both memory usage and computational costs.

- **Initial Vector Normalization**:
  - The initial PageRank vector is normalized with `v = v / np.linalg.norm(v, 1)` to ensure that it is a valid starting point for the PageRank calculations. This normalization helps achieve convergence more quickly.

2.6.2 Code Optimization for Large Graphs

For even better scalability, additional optimizations can be employed:

2.6.3 Code Optimization for Large Graphs

To achieve better scalability, I could employ additional optimizations:

- **Sparse Matrix Representation**:
  - I could convert the adjacency matrix to a sparse format to save memory and speed up computations:

    ```
    from scipy.sparse import csr_matrix
    M_sparse = csr_matrix(M)
    ```

    Then, I would use sparse matrix operations for the PageRank calculation:

    ```
    M_hat = d * M_sparse + (1 - d) / N * np.ones((N, N))
    v_new = M_hat.dot(v)
    ```

    By converting the adjacency matrix to a sparse representation, I could reduce memory usage and potentially improve computation speed.

- **Adaptive Iteration Limit**:
  - I could implement a dynamic iteration limit based on actual convergence progress, rather than relying on a fixed number of iterations. This would involve:

    ```
    for _ in range(num_iterations):
        v_new = M_hat.dot(v)
        if np.linalg.norm(v_new - v, 1) < tol:
            break
        v = v_new
    ```

    By adjusting the iteration limit dynamically, I would ensure that the algorithm converges more efficiently.

2.6.4 Summary Table of Scalability Features

Feature	Description
Efficient Matrix Operations	Uses optimized NumPy functions for matrix multiplications.
Memory Efficiency	NumPy arrays handle large datasets with minimal memory overhead.
Convergence Criteria	Uses a tolerance parameter to stop iterations when convergence is achieved.
Damping Factor	Stabilizes the PageRank computation and ensures proper convergence.
Sparse Matrix Support	For very large graphs, can be optimized with sparse matrix representations.
Initial Vector Normalization	Ensures that the initial vector is valid for PageRank computation and helps achieve convergence.

Table 2.1: Summary of scalability features in the PageRank function.

2.7 Results and Analysis

2.8 Top 10 PageRank Values

The 10 highest PageRank values obtained from the PageRank algorithm are as follows:

PageRank Value
0.0000000891
0.0000000863
0.0000000863
0.0000000836
0.0000000836
0.0000000808
0.0000000804
0.0000000804
0.0000000800
0.0000000798

Table 2.2: The 10 highest PageRank values obtained from the PageRank algorithm.

2.8.1 Analysis of the PageRank Values

The PageRank values obtained are relatively low, with the highest value being approximately 0.0000000891. There are several potential reasons for observing such low PageRank values:

- **Graph Sparsity**:
  - The graph constructed from the dataset is very sparse, meaning there are relatively few edges compared to the number of nodes. In a sparse graph, the PageRank values are generally low because there are fewer connections through which "rank" can propagate.
- **Dataset Characteristics**:
  - The dataset may represent a niche or highly specific collection of artworks where the relationships between nodes (artworks) are less interconnected. This could result in lower PageRank values as there are fewer prominent nodes with high PageRank scores.
- **Graph Construction Method**:
  - The method used to create edges based on shared tags might not capture all possible relationships between artworks. For instance, if the tagging system is very broad or not detailed, the edges in the graph might not reflect the true significance or popularity of the artworks.
- **Damping Factor**:
  - The damping factor used in the PageRank calculation is set to 0.85, which means 15% of the rank is distributed uniformly across all nodes. While this value is standard, the choice of damping factor can influence the PageRank results, though it is unlikely to be the sole cause of the low values observed.
- **Graph Size**:
  - For very large graphs with many nodes, the PageRank values can become very small. In this case, a large number of nodes and edges can dilute the importance of individual nodes.