Project Proposal
INFO 623: 001
Caleb Miller and Luke Chesley

Project Proposal: Analyzing the Structure of Human Knowledge Through Wikipedia's Hyperlink Graph

Introduction:

The immense network of Wikipedia articles, connected through hyperlinks, presents a unique opportunity to explore the structure of human knowledge. By analyzing this network, we can uncover patterns that reveal how information is interconnected, which topics are deemed most important, and how knowledge communities cluster together. The goal of this project is to thoroughly analyze the Wikipedia hyperlink graph to answer questions about the network's structure and what it tells us about human knowledge as a whole.

Research Questions and Motivations:

This project is guided by a series of research questions focused on understanding the architecture of the Wikipedia graph:

1. Centrality and Prestige: Which Wikipedia articles are most central in the graph, indicating high prestige?
2. Community Structure: What are the most densely connected subgraphs, and how do these communities reflect clusters of knowledge?
3. Influence and Authority: Which pages have the most authority according to algorithms like PageRank, and how does this compare to their centrality?
4. Connectivity and Accessibility: What are the average shortest paths between nodes, and what does this tell us about the 'six degrees of separation' concept within human knowledge?
5. Clustering Coefficient: What is the clustering coefficient of the Wiki graph, and how does it illuminate the network's clustering tendencies?

These questions come from the desire to understand how knowledge is structured and interconnected in one of the most expansive databases in existence.

Suggested Data Collection:

The primary dataset for this project will be collected from the Stanford Large Network Dataset Collection, specifically the "Wikipedia Top Categories" dataset (available at https://snap.stanford.edu/data/wiki-topcats.html) . This dataset provides an all-encompassing look at Wikipedia's hyperlink structure, offering a rich foundation to our analysis. The dataset is a snapshot of the structure collected in September 2011, restricted to pages in categories with at least 100 pages. Although being outdated and slightly pruned, this dataset will still be able to provide insights into large structural patterns.

Intended Methodology:

Our approach will utilize Python's 'networkx' library to construct and analyze the Wikipedia hyperlink graph along with other Python tools to help organize and visualize the graph. The methodology will include:

- Network Construction: Building a graph representation of Wikipedia’s hyperlink network from the collected data
- Centrality and Prestige Analysis: Identifying the most central articles in the graph to determine prestige and comparing these findings with the PageRank scores to investigate correlations between centrality and perceived importance.
- Community Detection: Using ‘networkx’ community detection methods to detect closely connected subgraphs and study their significance and connectivity.
- Shortest Path Analysis: Computing the average shortest paths to investigate the network’s interconnectedness of knowledge categories.
- Influence Measurement: Applying algorithms like PageRank to understand the most influential pages, followed by a comparative analysis with cultural, academic, and news-based perceptions of importance.
- Clustering Coefficient Calculation: Determining the graph’s clustering coefficient to know the prevalence of the tightly knit communities within the larger network.
- Visualization: Developing effective visualizations to illustrate the structure of human knowledge and relationships between different areas of knowledge.