Project Overview

For my project, I read as many articles as I could find on Wikipedia, recording the list of links on each. From there, I used the network of which pages linked to which to generate a metric multi-dimensional scale plot.

Implementation

The first step in my project was collecting data. I did this with a basic recursive function that took a title and a dictionary, added the title to the dictionary as a key with a value of the list of links on that article, and then called itself for all linked articles. By imposing a recursion depth limit of 4, I was able to get all pages within four clicks of the "Adolf Hitler" Wikipedia article (in practice I only got about a quarter of that data since it took so long to scrape Wikipedia, and I only used a tenth of that because Python ran out of memory otherwise).

Once I had my dictionary of strings and lists of strings, I could construct a distance matrix, or a square two-dimensional list of floats representing the "distance" between each article. I defined distance simply as 4 if the articles were not linked, 2 if one linked to the other, and 1 if both linked to each other. From there, I simply plugged my distance matrix into an MDS function from the Manifold package, which gave me a PNG.

The final step of my project was interpreting my results. While I had code to place labels on dots with the name of the corresponding article, I could not simply label all of them, as the vast quantity of labels quicky cluttered my screen and became illegible. Therefore, I set it to randomly label about ten dots. I ran the program many times and looked for common patterns and trends across all of the runs. Unfortunately, the MDS function itself contained a random element meaning that each image looked slightly different. However, by anayzing many of them and exploring Wikipedia itself, I was eventually able to somewhat confidently name each c

luster visible in a resulting graph.

Results

Most of the results were unsurprising. Most articles were not linked together, so they spaced themselves roughly into a circle of radius 4. Many articles were, in fact, linked to other parts of the circle, so there was a general smattering of dots throughout the diagram. However, also as expected, several clear clusters of articles emerged. I have identified the two main ones as "Socialists/Fascists", and the other as "the Holocaust", generally identifiable by its characteristic ring of "Books about the Holocaust" hovering nearby. Along the very edge of the circle, the smaller clusters of "Wikipedia" related articles and more general "Encyclopedia" articles were always visible. Two of the more interesting clumps were those of people Hitler knew. While I am not sure what exactly forced them into their two separate groups, my current hypothesis is that one contains Hitler's family and personal friends, while the other contains Hitler's co-workers. Their positions seem to be independent of each other.

Perhaps the most interesting regions were the "Nationalists" band, which tended to appear a certain distance away from "Socialists/Fascists", and the "World War II" band, which appeared near, but clearly distinct from, "the Holocaust." The grouping of articles on Wikipedia, as well as the separation of different groups, is telling of how Wikipedia and society view certain people, events, and trends.

Reflection

I feel that this project went well. There are things that I might prefer to have done differently, such as a more complex distnace algorithm, or figuring out how to use a larger sample size. However, I think that it was appropriately scoped as a one-week project. Each of the three problems I had to solve in my implementation was challenging in its own way, and I had fun.