INFSCI 1530: Data Mining

Homework 4

A. Use the network provided and stored in the file "stormofswords.csv"[1] and calculate the page rank of each node using the random walk process that we described in the lecture.
   a. Estimate the page rank for different lengths of walk (i.e., repetitions of the random walk) and compare with the result you got from the function pagerank of networkX library. For the comparison use the Spearman ranking correlation. (50 points)
   b. Using the networkX pagerank function experiment for different "teleport" probabilities (for a fixed number of walk length) and for each value of teleport probability (ranging from 0 to 1, with a step of 0.1) calculate the Gini coefficient of the PageRank values distribution. The Gini coefficient takes as input a set of data points and returns a value between 0 and 1. If the Gini is 0, then all the points have the same value (i.e., all the nodes have the same page rank). If the Gini is 1, we have the maximum possibly inequality among the values examined. The Gini coefficient is calculated as follows:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i}$$

   where x_i is the value of data point i and n is the number of total points in the dataset. Plot the value of the page rank Gini coefficient for the different values of the teleport probability. What do you observe? Why? (50 points)

[1] The nodes in this network contain the 107 different characters appearing in the Game of Thrones novel, and the edges contains 353 weighted relationships between those characters, which were calculated based on how many times two characters' names appeared within 15 words of one another in the novel.