1. All the extracted html and processed documents from the links are in “. /Assignment3/1000Documents” and are labeled rawData(1-1000).txt and processedData(1-1000).txt respectively. All the files numbers correspond to the lines of their respective URI’s in the “linkList.txt” list (for instance the URI for rawData5.txt is on line 5 of the linkList.txt file). If a URI was not successfully accessed for one reason or another then both files for that URI were filled simply with the URI itself. I selected **“Twitter”** as my key word, and searched through my files until I found a good number of files that used Twitter as a word in them. The 10 I used have been copied over into another folder labeled “. /Assignment3/10SpecialDocuments” and the numbers of the files correspond to the number of the line that the URI is on in the “linkList.txt” list.

2.



3.



4. Kendall tau 0.0544

p-value 0.9189