Matthew Payne

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.

Document	Term Count	Word Count	TF	IDF	TF-IDF	URI
1	8	1760	0.0045	2	0.0091	http://weheartit.com/mi_putamelena
2	1	112	0.0089	2	0.0179	http://www.haba.net.pl
3	15	1326	0.0113	2	0.0226	http://twilog.org/masahiro_sakai/
4	2	813	0.0025	2	0.0049	http://www.david-sadler.org
5	3	400	0.0075	2	0.0150	http://www.twitlonger.com/show/n_1so7qho
6	1	3885	0.0003	2	0.0005	http://rwby.wikia.com/wiki/Nora_Valkyrie
7	2	479	0.0042	2	0.0084	http://blog.goo.ne.jp/shin-ya0215
8	1	2164	0.0005	2	0.0009	http://hatariwater.tumblr.com
9	2	865	0.0023	2	0.0046	http://creative-punch.net

3.

Document	Page Rank	URI
1	0.5	http://weheartit.com/mi_putamelena
4	0.3	http://www.david-sadler.org
6	0.2	http://rwby.wikia.com/wiki/Nora_Valkyrie
3	0.1	http://twilog.org/masahiro_sakai/
2	0	http://www.haba.net.pl
5	0	http://www.twitlonger.com/show/n_1so7qho
7	0	http://blog.goo.ne.jp/shin-ya0215
8	0	http://hatariwater.tumblr.com
9	0	http://creative-punch.net

4.

Kendall tau	0.0544
p-value	0.9189