Assignment 3

Web Science CS595

Name: Amara Naas

TFIDF	TF	IDF	URI
0.102	0.019	5.322	http://www.politico.com/story/2013/09/why-barack-obama-looks-so-exhausted-97191.html?hp=l10
0.054	0.010	5.322	http://rt.com/usa/obama-clapper-nsa-panel-240/
0.024	0.005	5.322	http://www.bbc.co.uk/news/world-africa-24210959
0.015	0.003	5.322	http://www.dnaindia.com/world/1893084/report-vladimir-putin-says-syria-violence-could-hit-ex-soviet-bloc
0.015	0.003	5.322	http://geopoliting.com/ISF
0.013	0.002	5.322	http://articles.washingtonpost.com/2012-06-13/world/35462541_1_burkina-faso-air-bases-sahara
0.004	0.001	5.322	http://www.americanthinker.com/2011/07/m-the_knockout_game_racial_violence_and_the_conspicuous_silence_of_the_media.html
0.004	0.001	5.322	http://www.buzzfeed.com/ariellecalderon/27-things-advertising-people-know-to-be-true
0.003	0.001	5.322	http://www.federalnewsradio.com/
0.001	0.000	5.322	http://www.storyleak.com/pentagon-prepping-large-scale-economic-breakdown/

Table 1: Ranking of 10 URIs using TF-IDF Formula

Question 1

In the first Python program (gethtml.py)[1] will download the 1000 unique links, and save them in files directory as f1.html to f1000.html. Also it will process these all html files to remove HTML markup by using the (% lynx dump force_html) comand and save them in the same directory as f1.html.processed to f1000.html.processed. mytweets.py[1].

Question 2

The same Python program (gethtml.py)[1] will make a query and uses Obama as a query term. Table 1 shows the computation of TF, IDF, and TFIDF values for the term in each of the 10 documents and corresponding URIs.

[1]File uploaded to github

PageRank	URI
9	http://www.bbc.co.uk/news/world-africa-24210959
7	http://www.politico.com/story/2013/09/why-barack-obama-looks-so-exhausted-97191.html?hp=l10
7	http://rt.com/usa/obama-clapper-nsa-panel-240/
7	http://www.federalnewsradio.com/
7	http://www.dnaindia.com/world/1893084/report-vladimir-putin-says-syria-violence-could-hit-ex-soviet-bloc
7	http://articles.washingtonpost.com/2012-06-13/world/35462541_1_burkina-faso-air-bases-sahara
7	http://www.buzzfeed.com/ariellecalderon/27-things-advertising-people-know-to-be-true
6	http://www.americanthinker.com/2011/07/m-the_knockout_game_racial_violence_and_the_conspicuous_silence_of_the_media.html
3	http://geopoliting.com/ISF
0	http://www.storyleak.com/pentagon-prepping-large-scale-economic-breakdown/

Table 2: Ranking of 10 URIs using http://www.checkpagerank.net/

Question 3

Table 2 shows the ranking of the 10 URIs from table 1 by using `http://www.checkpagerank.net/`. By looking to both tables it seems that there is not enough evidence to support the claim that there is direct or inverse relationship between Page ranking and TFIDF.

Question 4

Part 1:

Tau-b can be found from the following formula:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{1}$$

where as:

$$n_0 = n(n-1)/2 \tag{2}$$

$$n_1 = \sum_i t_i(t_i - 1)/2 \tag{3}$$

$$n_2 = \sum_j u_j(u_j - 1)/2 \tag{4}$$

n_c = Number of concordant pairs
n_d = Number of discordant pairs
t_i = Number of tied values in the i^{th} group of ties for the first quantity
u_j = Number of tied values in the j^{th} group of ties for the second quantity

Now we need to rearrange the data of TFIDF in table 1 according to page rank in table 2 as shown in table 3. First we find the number of tied values in the i^{th} group and j^{th} group:
from table 3:
$n_c = 9$
$n_d = 22$
$t_i = 2$
$u_j = 2$
$n_0 = 10(10\text{-}1)/2 = 45$ from equation (2).
$n_1 = (2(2\text{-}1)+2(2\text{-}1))/2 = 2$ from equation (3).
$n_2 = (2(2\text{-}1)+2(2\text{-}1))/2 = 2$ from equation (4).

$$\tau_B = \frac{9 - 22}{\sqrt{(45 - 2)(45 - 2)}} = -0.302$$

Part 2:

The sample correlation coefficient can be used to estimate the population Pearson correlation r between the two list as follows:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{5}$$

$$r_{xy} = \frac{10 * 0.163 - 0.235 * 6}{\sqrt{10 * 0.146 - (0.235)^2}\sqrt{10 * 4.2 - (6)^2}}$$

$$r_{xy} = \frac{0.215}{\sqrt{1.46 - 0.055}\sqrt{42 - 36}} = 0.445$$

TFIDF	P-R	n_c	n_d	URI
0.102	7	1	3	http://www.politico.com/story/2013/09/why-barack-obama-looks-so-exhausted-97191.html?hp=l10
0.054	0.7	1	3	http://rt.com/usa/obama-clapper-nsa-panel-240/
0.024	0.9	0	7	http://www.bbc.co.uk/news/world-africa-24210959
0.015	0.7	0	3	http://www.dnaindia.com/world/1893084/report-vladimir-putin-says-syria-violence-could-hit-ex-soviet-bloc
0.015	0.3	4	1	http://geopoliting.com/ISF
0.013	0.7	0	2	http://articles.washingtonpost.com/2012-06-13/world/35462541_1_burkina-faso-air-bases-sahara
0.004	0.6	2	1	http://www.americanthinker.com/2011/07/m-the_knockout_game_racial_violence_and_the_conspicuous_silence_of_the_media.html
0.004	0.7	0	1	http://www.buzzfeed.com/ariellecalderon/27-things-advertising-people-know-to-be-true
0.003	0.7	1	1	http://www.federalnewsradio.com/
0.001	0	0	0	http://www.storyleak.com/pentagon-prepping-large-scale-economic-breakdown/
sum of	above	9	22	

Table 3: concordant discordant pairs