Homework 4 Part 1

File size for “tinyfiles” test case

	Dict	Post	Map
NumRecords from wc -l	**30**	**11**	**4**
Filesize from ls -l	**1950**	**220**	**200**
RecordSize: Filesize/NumRecords	**30**	**11**	**4**

My formula for idf is:

$$1 + log_2(N/df_t)$$

- N = number of document in collection
- df_t = frequency of a term in that document

Fill in the table below with values calculated using a calculator (not what is in the file):

Term	NumDocs (from dict)	idf value
dog	2	2
quickly	1	3

☑ ~~For num_tokens, use the count from your program, i.e., after stopword removal. Do these calculations with a calculator or spreadsheet. If you did another calculation for normalization, replace the rtf line in the table.~~

☑ ~~Note the wt you calculate for the term above should match the wt stored in the post file for that term.~~

	0.html	1.html	2.html	3.html
Num_tokens	4	5	8	3
Freq(dog)	1	1	0	0
Rtf (dog)	0.25	0.20	0	0
Freq(quickly)	1	0	0	0
Rtf(quicly)	0.25	0	0	0
rtf*idf(dog)	0.50	0.4	0	0
rtf*idf(quickly)	0.75	0	0	0
Post wt (dog)	0.50	0.40	0	0
Post wt (quickly)	0.75	0	0	0
Post wt(dog)+wt(quickly)	1.25	0.40	0	0