Kyle Morley
HW3.1 Template:
Midpoint checkin (9 point bonus)

CHECKING FOR FIXED LENGTH FILES

Dict rec_size: 26 bytes

Num unique terms: 29,221 terms

Predicted dict.txt size: 0.7245 MB

Post rec_size: 13_ bytes

Num postings (wc –l post.txt): 180990 postings

Predicted post.txt size: 2.24 MB

Map rec_size: 69 bytes

Num documents: 955 terms

Predicted map.txt size: 0.0628 MB

Results of ls –l on turing showing file sizes

```
kem021@turing:~/Info/Tokenizer$ ls -l output/
total 3336
-rw-rw---- 1 kem021 kem021  987662 Oct 15 17:35 dictionary.txt
-rw-rw---- 1 kem021 kem021   65895 Oct 15 17:35 mappings.txt
-rw-rw---- 1 kem021 kem021 2352870 Oct 15 17:35 postings.txt
```

CHECKING FOR FILE CONTENTS

First 5 lines of dict.txt

```
transmission 1   944268
disney       2   1680952


assumptions  6   1624376
```

First 5 lines of post.txt

1 0.004299
2 0.009758
3 0.006340

4 0.003495
5 0.006486

First 5 lines of map.txt

```
1   /home/sgauch/public_html/5533/files/799.html
2   /home/sgauch/public_html/5533/files/64.html
3   /home/sgauch/public_html/5533/files/558.html
4   /home/sgauch/public_html/5533/files/476.html
5   /home/sgauch/public_html/5533/files/977.html
```

CHECKING TERM WEIGHT CALCULATIONS

Num_docs for algorithm (from dict.txt): 955 documents

idf of "algorithm" (calculated): 47/955 = 20.32

freq of "algorithm" in document 27 (checked with grep): 3 occurrences

num tokens in document 27 (from indexing code): 1435 tokens

normalized tf for "algorithm" in document 27: 3/1435 = 0.00209

predicted wt for "algorithm" in document 27 (norm_tf * idf, calculated):

0.0424

post record for algorithm for document 27: 624 0.042479