1.

TIny:

```
[0] the = 537,100
[1] and = 215,318
[2] for = 58,256
[3] was = 57,294
[4] that = 52,341
[5] with = 49,599
[6] are = 37,555
[7] from = 37,186
[8] his = 29,874
[9] which = 27,107
[10] this = 25,969
[11] were = 21,561
[12] not = 20,028
[13] have = 19,854
[14] also = 19,274
[15] has = 18,821
[16] one = 17,251
[17] but = 16,384
[18] their = 15,892
[19] its = 15,453
[20] had = 15,233
…
[5000] informed = 155
[5001] mirror = 155
[5002] mosque = 155
[5003] odd = 155
[5004] owing = 155
[5005] personally = 155
[5006] photos = 155
[5007] predominantly = 155
[5008] protest = 155
[5009] prussian = 155
[5010] samples = 155
[5011] singapore = 155
[5012] sizes = 155
[5013] victorian = 155
[5014] advocate = 154
[5015] algae = 154
[5016] bed = 154
[5017] bombing = 154
```

```
[5018] camps = 154
[5019] clean = 154
[5020] cooking = 154
```

Small:

```
[0] the = 5,461,246
[1] and = 2,177,629
[2] was = 639,701
[3] for = 598,874
[4] that = 514,966
[5] with = 504,154
[6] from = 377,942
[7] are = 344,274
[8] his = 318,873
[9] which = 262,581
[10] this = 252,011
[11] were = 234,607
[12] also = 192,049
[13] not = 190,823
[14] has = 180,275
[15] have = 179,453
[16] one = 169,401
[17] had = 164,176
[18] but = 159,441
[19] their = 158,974
[20] first = 151,391
…
[5000] awareness = 1,510
[5001] lecture = 1,510
[5002] monster = 1,510
[5003] submitted = 1,510
[5004] traded = 1,510
[5005] emerging = 1,509
[5006] serbian = 1,509
[5007] allegedly = 1,508
[5008] restore = 1,508
[5009] trouble = 1,508
[5010] athlete = 1,507
[5011] guards = 1,507
[5012] manuel = 1,507
[5013] parameters = 1,507
[5014] bros = 1,506
[5015] postal = 1,505
```

[5016] sequel = 1,505
[5017] swimming = 1,505
[5018] touring = 1,505
[5019] ethics = 1,504
[5020] luis = 1,504

Medium:

[0] the = 81,010,395
[1] and = 33,426,873
[2] was = 12,376,966
[3] for = 9,800,783
[4] with = 7,883,270
[5] that = 6,276,909
[6] from = 6,109,511
[7] his = 5,478,788
[8] are = 3,841,174
[9] were = 3,414,079
[10] which = 3,326,581
[11] this = 3,269,899
[12] also = 2,920,546
[13] has = 2,875,770
[14] first = 2,532,967
[15] one = 2,510,144
[16] new = 2,479,391
[17] had = 2,453,648
[18] their = 2,261,273
[19] not = 2,200,917
[20] but = 2,148,312
…
[5000] organisations = 22,271
[5001] vampire = 22,262
[5002] imprisoned = 22,261
[5003] burn = 22,250
[5004] wait = 22,248
[5005] lighting = 22,238
[5006] reconstruction = 22,234
[5007] certificate = 22,233
[5008] enjoy = 22,232
[5009] cricketer = 22,226
[5010] associates = 22,225
[5011] seriously = 22,223
[5012] edmonton = 22,219
[5013] payment = 22,218

```
[5014] bryan = 22,214
[5015] associations = 22,203
[5016] filming = 22,199
[5017] monk = 22,196
[5018] whatever = 22,189
[5019] actively = 22,175
[5020] prepare = 22,159
```

All:

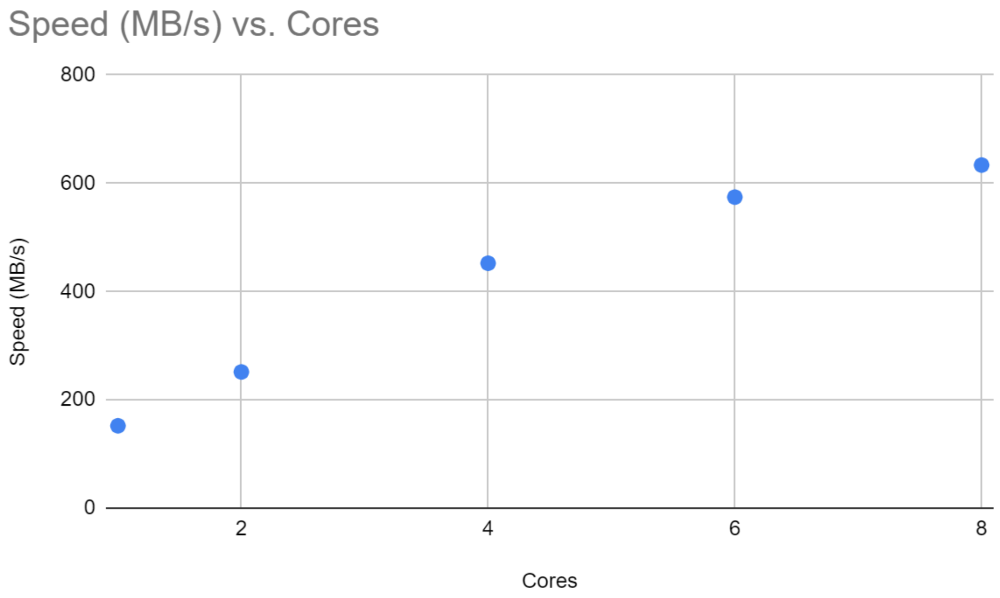
```
[0] the = 219,389,160
[1] and = 86,539,407
[2] you = 52,970,534
[3] that = 43,748,250
[4] for = 41,742,832
[5] this = 31,341,779
[6] wikipedia = 29,056,199
[7] was = 25,953,578
[8] not = 25,157,584
[9] with = 22,376,657
[10] page = 19,603,813
[11] have = 18,507,317
[12] are = 18,091,178
[13] from = 18,001,827
[14] talk = 17,138,587
[15] your = 16,876,285
[16] article = 16,686,845
[17] please = 15,451,407
[18] but = 13,911,837
[19] has = 13,749,490
[20] been = 12,361,008
…
[5000] constituency = 52,605
[5001] shares = 52,601
[5002] offense = 52,582
[5003] killer = 52,581
[5004] reflects = 52,576
[5005] occasion = 52,556
[5006] missile = 52,545
[5007] unblocked = 52,528
[5008] txt = 52,509
[5009] translate = 52,505
[5010] liberation = 52,486
[5011] diverse = 52,469
```

```
[5012] mounted = 52,460
[5013] specialist = 52,448
[5014] warming = 52,441
[5015] bridges = 52,428
[5016] commissioner = 52,424
[5017] incidents = 52,423
[5018] alphabet = 52,414
[5019] bug = 52,407
[5020] organic = 52,407
```

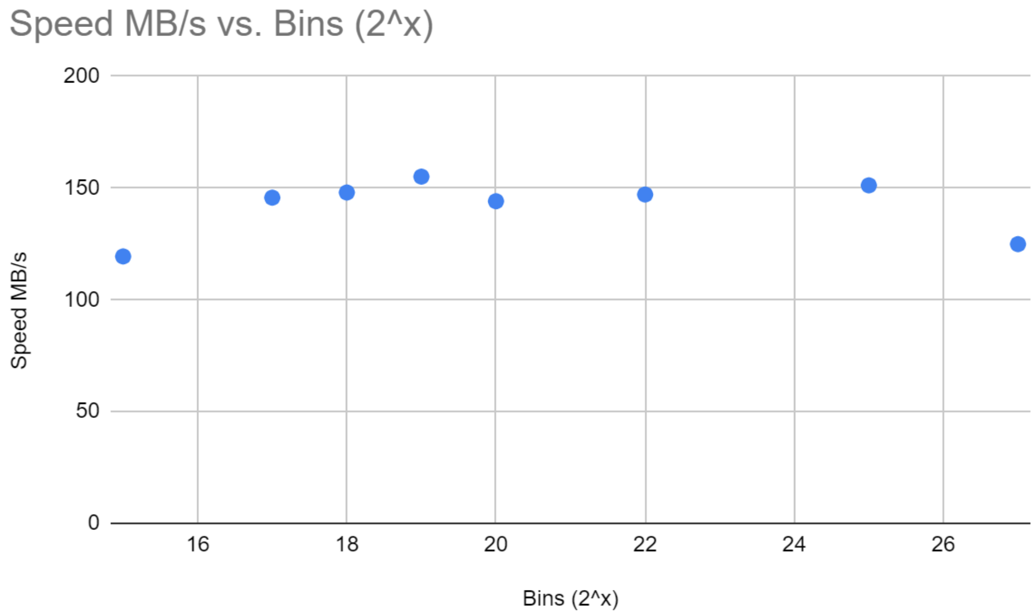
I broke ties by overriding the < operator to check if the counts are equal and if so it uses strcmp to compare the two strings and break the tie. Some interesting issues I encountered during the homework were mostly related to adapting my shadow buffers to work with the new word definition. After that, everything was quite straightforward and debugging was just a matter of setting up lookup tables and data structures correctly.

2.

My design idea for the parallel hash table was just to give each core its own local hash table and merge them all together after the whole file has been processed.



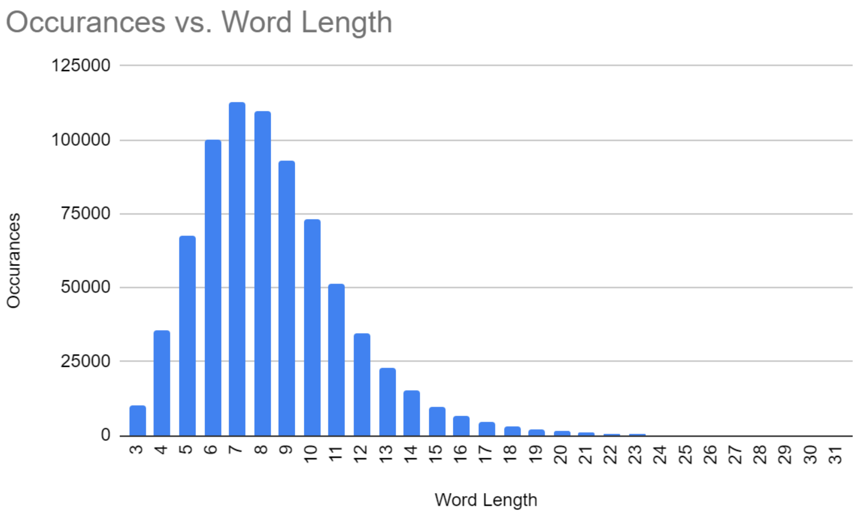
When searching for the best number of bins, I tested a bunch of values with 1 thread and checked the speeds, here are the results:

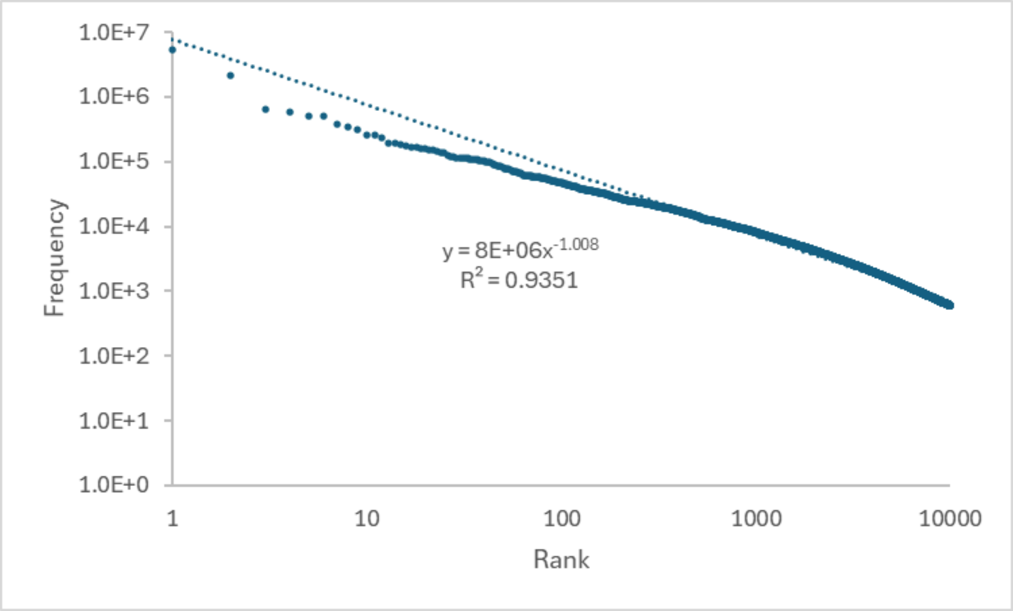


I would choose a bin size of 2^19 based on these results.

3. Too much work.

4.





Looks pretty compliant to me.