Algorithmic Foundations of Data Science: Assignment #4

Khaled Al Hosani (kah579), Myunggun Seo (ms9144)

New York University Abu Dhabi

Date: February 25, 2018

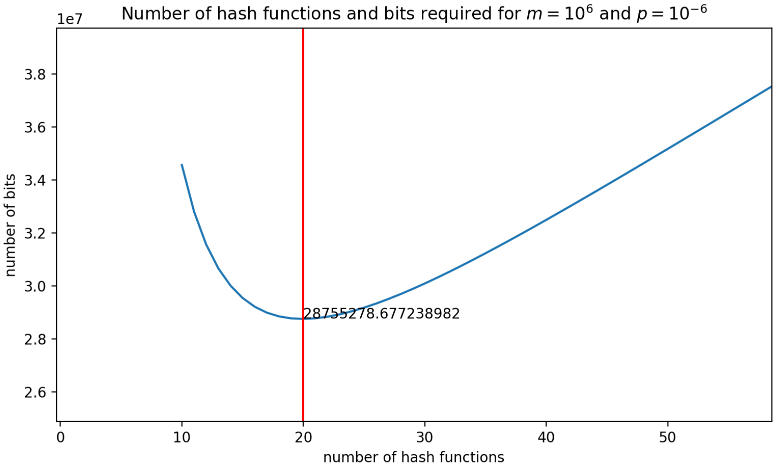
Problem 1

Let P be the chance of a false positive.

$$\begin{aligned} P &= (1 - e^{-mk/n})^k \\ P^{1/k} &= 1 - e^{-mk/n} \\ 1 - P^{1/k} &= e^{-mk/n} \\ -mk/n &= ln(1 - P^{1/k}) \\ n &= \frac{-mk}{ln(1 - P^{1/k})} \end{aligned}$$

Plugging in m and P:

$$n = \frac{-10^6 k}{ln(1 - (10^{-6})^{1/k})}$$



This equation gives minimum n at $k = 20$ for $n = 28755378$ bits.

Problem 2

Please see the attached files ‘`LogCounter.py`’, ‘`functions.py`’, ‘`preAnalysis.py`’, and ‘`main.py`’.

- ‘`LogCounter.py`’ is a module that times and counts iterations. It is not directly relevant to the problem. It is not executable.
- ‘`functions.py`’ include a function that parse the movie ratings data, a cosine distance function, and a combinations function. It is not executable.
- ‘`preAnalysis.py`’ is used for plotting the cosine distance of a sample of users, for analyzing r,b values, and for creating random vectors. You can run this.
- ‘`main.py`’ is used for using LSH to put put users into buckets and finding candidate pairs, and then compute the cosine distances for the candidates in order to find the actually similar users.

Conclusion: We used (r,b) = (2,8) to find 23 users who have very similar tastes. Depending on the the threshold, we can output more users with less similar tastes.

Here are some users that have similar tastes and their cosine distances:

1. 151, 369 : 0.563668
2. 279, 369 : 0.657341
3. 151, 400 : 0.658115
4. 82, 400 : 0.667391
5. 191, 513 : 0.679743
6. 151, 279 : 0.682376
7. 279, 400 : 0.697313
8. 50, 151 : 0.701212
9. 191, 449 : 0.708730
10. 369, 400 : 0.710409
11. 144, 375 : 0.722752
12. 329, 459 : 0.727810
13. 108, 225 : 0.730521
14. 82, 191 : 0.735493
15. 144, 151 : 0.736513
16. 317, 415 : 0.740166
17. 317, 556 : 0.740470
18. 145, 151 : 0.757147
19. 64, 657 : 0.757271
20. 279, 662 : 0.763511
21. 144, 400 : 0.770607
22. 145, 400 : 0.775263
23. 151, 535 : 0.776843