Homework #2 (Data Mining)

Name: Spencer Fronberg
UID: u0766439

February 8, 2018

1 Creating k-Grams

(A) When k-gram is 2 characters (G1):
D1: 263
D2: 262
D3: 269
D4: 255
When k-gram is 3 characters (G2):
D1: 765
D2: 762
D3: 828
D4: 698
When k-gram is 2 words (G3):
D1: 279
D2: 278
D3: 337
D4: 232

(B) When k-gram is 2 characters (G1):
D1, D2 Jaccard similarity is: 0.9811320754716981
D1, D3 Jaccard similarity is: 0.8156996587030717
D1, D4 Jaccard similarity is: 0.6444444444444445
D2, D3 Jaccard similarity is: 0.8
D2, D4 Jaccard similarity is: 0.6412698412698413
D3, D4 Jaccard similarity is: 0.6529968454258676

When k-gram is 3 characters (G2):
D1, D2 Jaccard similarity is: 0.977979274611399
D1, D3 Jaccard similarity is: 0.5803571428571429
D1, D4 Jaccard similarity is: 0.3050847457627119

D2, D3 Jaccard similarity is: 0.5680473372781065
D2, D4 Jaccard similarity is: 0.30590339892665475
D3, D4 Jaccard similarity is: 0.31212381771281167

When k-gram is 2 words (G3):
D1, D2 Jaccard similarity is: 0.9407665505226481
D1, D3 Jaccard similarity is: 0.18234165067178504
D1, D4 Jaccard similarity is: 0.03024193548387097
D2, D3 Jaccard similarity is: 0.1736641221374046
D2, D4 Jaccard similarity is: 0.030303030303030304
D3, D4 Jaccard similarity is: 0.01607142857142857

2 Min Hashing

(A) The Jaccard similarity for D1 and D2 using min-hash:
t=20 is: 1.0
Time = 0.04724971282323295

t=60 is: 0.9666666666666667
Time = 0.11190287143923389

t=150 is: 0.98
Time = 0.2625936968086635

t=300 is: 0.9766666666666667
Time = 0.5599774528262382

t=600 is: 0.975
Time = 1.0814925145189511

t=1000 is: 0.977
Time = 2.0303672496142924

t=2000 is: 0.978
Time = 4.057545379676835

t=4000 is: 0.9775
Time = 7.205715043464481

t=10000 is: 0.9782
Time = 17.6952514633904

(B) I did a few extra values for t on Part 2(B), and I would say that the best value for t in terms of time and accuracy would be 600 because the Jaccard similarity is only 0.002 off of the original Jaccard similarity from Part 1(B) and the time is only 0.56 seconds.

3 LSH

(A) First off, $b = -log_{\tau}(t)$
We know that τ is 0.7 because it is the threshold and t is 160 hash functions. So we can say the following about b:

$$b = -log_{0.7}(160)$$

We then know that:

$$b = \frac{t}{r}$$

so:

$$r = \frac{t}{b}$$

$$r = \frac{160}{-log_{0.7}(160)}$$

I then get the following for r and b but it is not the final answer because you cannot take a fraction of a hash:

$$b = 14.22912908$$

$$r = 11.24453922$$

I will use the following formula to make sure that I get the best values for b and r:

$$\tau = (\frac{1}{r})^{\frac{1}{b}}$$

We need τ to be set to 0.7 so:

$$0.7 = (\frac{1}{r})^{\frac{1}{b}}$$

I then wrote a program that I tested the range for b and r to be from 8 to 17 (getting every possibility between). The following are the results that I got for all the possible combinations for b and r that was within the range given where $b * r$ was between 150 and 170. I do $r * b$ because $t = r * b$ and we want the t that is closest to 160. Below shows the closest to t to 160 and τ to 0.7 that is in bold text.

b = 9, r = 17, T = 0.73
b * r = 153

b = 10, r = 16, T = 0.758
b * r = 160

b = 11, r = 14, T = 0.787
b * r = 154

b = 11, r = 15, T = 0.782
b * r = 165

b = 12, r = 13, T = 0.808
b * r = 156

b = 12, r = 14, T = 0.803
b * r = 168

b = 13, r = 12, T = 0.826
b * r = 156

b = 13, r = 13, T = 0.821
b * r = 169

b = 14, r = 11, T = 0.843
b * r = 154

b = 14, r = 12, T = 0.837
b * r = 168

b = 15, r = 11, T = 0.852
b * r = 165

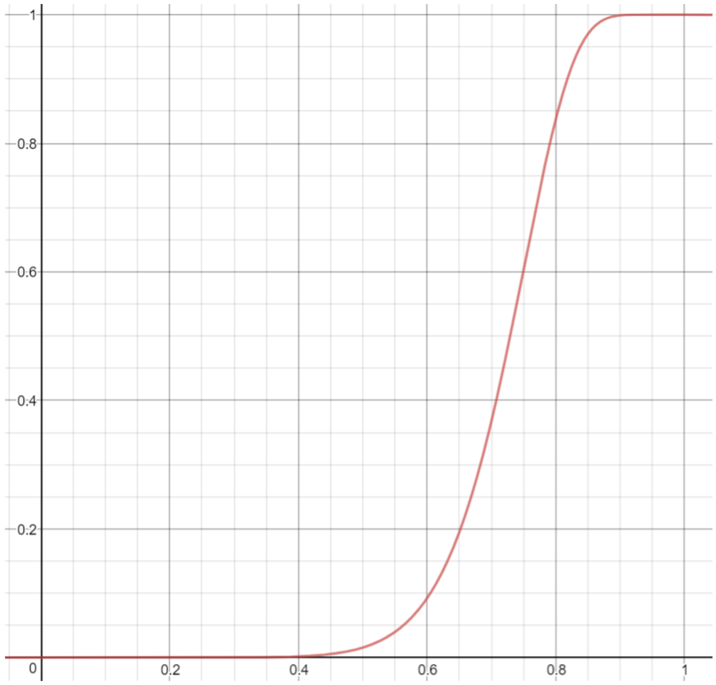
b = 16, r = 10, T = 0.866
b * r = 160

b = 17, r = 9, T = 0.879
b * r = 153

I then plotted the following where b = 10, r = 16:

$$f(s) = 1 - (1 - s^b)^r$$

$$f(s) = 1 - (1 - s^{10})^{16}$$



As you can see in the plot, the steepest part of the graph is really close to 0.7, so we can conclude that our $\mathbf{b = 10}$ and our $\mathbf{r = 16}$.

(B) The following are the probabilities when k-gram is 3 characters (G2) from the Jaccard similarities from Part 1(B) using the following formula (which was shown in the previous problem) where the s is the Jaccard similarity:

$$f(s) = 1 - (1 - s^{10})^{16}$$

The following is for D1 and D2 where s is 0.977979274611399:

$$f(s) = 1 - (1 - (0.977979274611399)^{10})^{16}$$

$$f(s) = 0.9999999999936431 \approx \mathbf{100.0000\%}$$

The following is for D1 and D3 where s is 0.5803571428571429:

$$f(s) = 1 - (1 - (0.5803571428571429)^{10})^{16}$$

$$f(s) = 0.06714456597772245 \approx \mathbf{6.7145\%}$$

The following is for D1 and D4 where s is 0.3050847457627119:

$$f(s) = 1 - (1 - (0.3050847457627119)^{10})^{16}$$

$$f(s) = 0.0001117640584525903 \approx \mathbf{0.0112\%}$$

The following is for D2 and D3 where s is 0.5680473372781065:

$$f(s) = 1 - (1 - (0.5680473372781065)^{10})^{16}$$

$$f(s) = 0.05452649364333684 \approx \mathbf{5.4526\%}$$

The following is for D2 and D4 where s is 0.30590339892665475:

$$f(s) = 1 - (1 - (0.30590339892665475)^{10})^{16}$$

$$f(s) = 0.00011479940476699646 \approx \mathbf{0.0115\%}$$

The following is for D3 and D4 where s is 0.31212381771281167:

$$f(s) = 1 - (1 - (0.31212381771281167)^{10})^{16}$$

$$f(s) = 0.00014039786464037363 \approx \mathbf{0.0140\%}$$