Asmt 2: Document Similarity and Hashing

Gopal Menon
Turn in (**a pdf**) through Canvas by 2:45pm:
Monday, February 13

1 Creating k-Grams (40 points)

A: (20 points) How many distinct k-grams are there for each document with each type of k-gram? You should report $4 \times 3 = 12$ different numbers.

Table 1: Number of distinct k-grams

Document	character 2-grams	character 3-grams	word 2-grams
$D1$	330	1297	520
$D2$	360	1514	631
$D3$	353	1541	840
$D4$	297	1541	412

B: (20 points)

A: (20 points) Compute the Jaccard similarity between all pairs of documents for each type of k-gram. You should report $3 \times 6 = 18$ different numbers.

Table 2: Jaccard similarity for character 2-grams

	$D1$	$D2$	$D3$	$D4$
$D1$				
$D2$	0.8499			
$D3$	0.7740	0.7649		
$D4$	0.7084	0.7109	0.7241	

Table 3: Jaccard similarity for character 3-grams

	$D1$	$D2$	$D3$	$D4$
$D1$				
$D2$	0.6400			
$D3$	0.4606	0.4404		
$D4$	0.3280	0.3125	0.3624	

Table 4: Jaccard similarity for word 2-grams

	$D1$	$D2$	$D3$	$D4$
$D1$				
$D2$	0.2579			
$D3$	0.0334	0.0251		
$D4$	0.0054	0.0058	0.0121	

2 Min Hashing (30 points)

A: (25 points) Using grams **G2**, build a min-hash signature for document $D1$ and $D2$ using $t = \{20, 60, 150, 300, 600\}$ hash functions. For each value of t report the approximate Jaccard similarity between the pair of documents $D1$ and $D2$, estimating the Jaccard similarity:

Table 5: Jaccard similarity between documents $D1$ and $D2$ using character 3-grams

t (number of hash functions)	Jaccard Similarity	Error %
20	0.6500	1.5625
60	0.6333	1.0469
150	0.6667	4.1719
300	0.6800	6.2500
600	0.6817	6.5156
1200	0.6625	3.5156
2400	0.6571	2.6719

B: (5 points) What seems to be a good value for t? You may run more experiments. Justify your answer in terms of both accuracy and time.

According to the Chernoff-Hoeffding bound, the probability that the average value for the Jaccard Similarity differs from the expected value by a value greater than α after t trials in the case where each value before it is averaged lies between $-\Delta$ and Δ, is given by:

$$\mathbf{Pr}\left[|A - \mathbf{E}(A)| > \alpha\right] \leq 2\exp\left(\frac{-t\alpha^2}{2\Delta^2}\right)$$

This means that as we increase the number of hash functions t, the probability that the Jaccard Similarity value found by min hashing will differ from the expected value by a large amount will keep falling. However as we increase the number of hash functions, the run time for the computation will increase.

From the experiment, it seems that the best value for t is 60 where the Jaccard Similarity was 0.6333. However, in another iteration of the experiment, the best value of 0.6338 was obtained with 2400 hash functions.

3 LSH (30 points)

A: (8 points) The probability of finding a collision in the hash value for two documents in any of the r bands each containing b hash functions is given by

$$f(s) = 1 - \left(1 - s^b\right)^r$$

To find all documents pairs with Jaccard Similarity above $\tau = 0.4$, we should select b and r so that the S-curve has the steepest slope at $s = 0.4$. A good estimate for this is given by:

$$b \approx -\log_\tau(k)$$
$$r = \frac{k}{b}$$

where k is the total number of hash functions. Using the values of $k = 160$ and $\tau = 0.4$,

$$
\begin{aligned}
b &\approx -\log_{0.4}(160) \\
&= \lfloor 5.5388 \rfloor \\
&= 5 \\
r &= \frac{160}{5} \\
&= 32
\end{aligned}
$$

Using these values of r and b, the S-curve obtained is shown in figure 1. After some trial and error with $b = 4$ and $r = 27$, the S-curve obtained is close to the steepest value at $\tau = 0.4$. This is shown in figure 2.

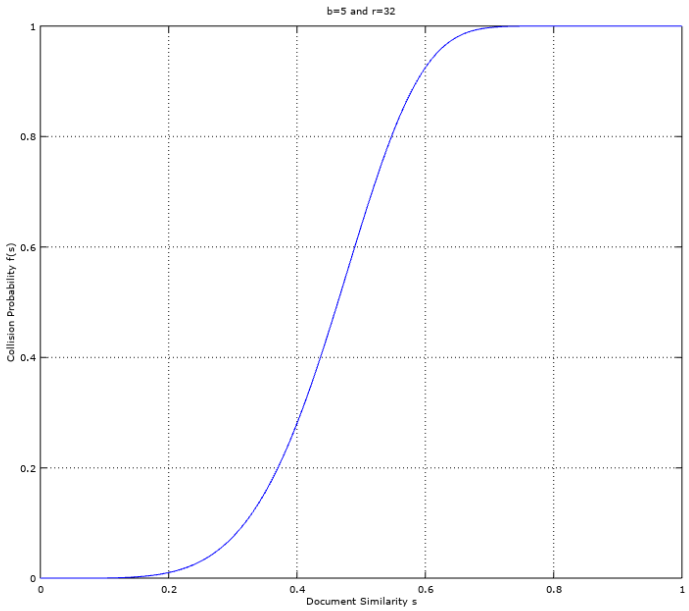

Figure 1: S-curve with $b = 5$ and $r = 32$

B: (24 points)

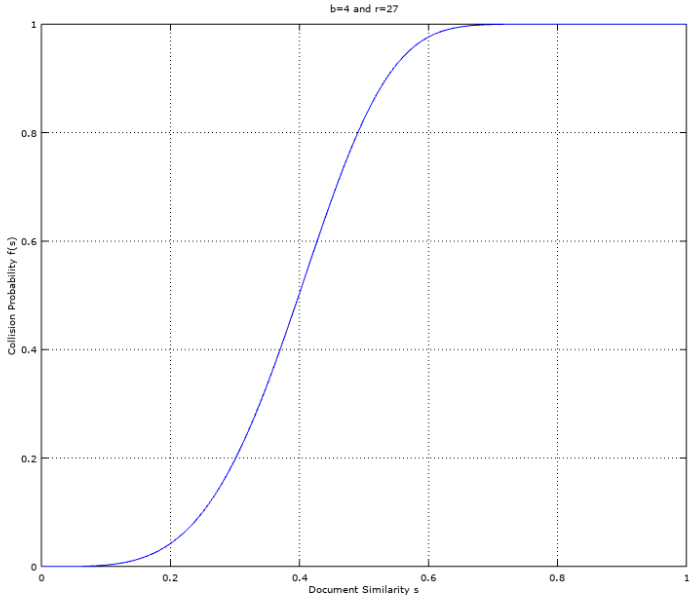

Figure 2: S-curve with $b = 4$ and $r = 27$