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)

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:

$$P(y' \neq y) = P(y' \neq y, x_1 = -1) + P(y' \neq y, x_1 = 1)$$

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	0.7000	9.3750
150	0.6600	3.1250
300	0.6333	−1.0469
600	0.6317	−1.2969