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: (5 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.txt	330	1297	520
D2.txt	360	1514	631
D3.txt	353	1541	840
D4.txt	297	1541	412

B: (10 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 distance for character 2-grams

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt				
D2.txt	0.8499			
D3.txt	0.7740	0.7649		
D4.txt	0.7084	0.7109	0.7241	

Table 3: Jaccard distance for character 3-grams

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt				
D2.txt	0.6400			
D3.txt	0.4606	0.4404		
D4.txt	0.3280	0.3125	0.3624	

Table 4: Jaccard distance for word 2-grams

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt				
D2.txt	0.2579			
D3.txt	0.0334	0.0251		
D4.txt	0.0054	0.0058	0.0121	