Asmt 2: Document Similarity and Hashing

Vai Suliafu, u0742607

1 Creating k-Grams (50 points)

A: (25 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.

[G1] $k = 2$ with characters:
D1 - 267 grams, D2 - 264 grams, D3 - 296 grams, D4 - 249 grams

[G2] $k = 3$ with characters:
D1 - 773 grams, D2 - 759 grams, D3 - 978 grams, D4 - 770 grams

[G3] $k = 2$ with words:
D1 - 290 grams, D2 - 297 grams, D3 - 390 grams, D4 - 364 grams

B: (25 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.

[G1] $k = 2$ with characters:
JS(D1,D2) = 0.9887640449438202
JS(D1,D3) = 0.7816455696202531
JS(D1,D4) = 0.6645161290322581
JS(D2,D3) = 0.7834394904458599
JS(D2,D4) = 0.6601941747572816
JS(D3,D4) = 0.6717791411042945

[G2] $k = 3$ with characters:
JS(D1,D2) = 0.9491094147582697
JS(D1,D3) = 0.5042955326460481
JS(D1,D4) = 0.3065198983911939

JS(D2,D3) = 0.4987057808455565
JS(D2,D4) = 0.3034953111679454
JS(D3,D4) = 0.31329827197595794

[G3] $k = 2$ with words:
JS(D1,D2) = 0.78419452887538
JS(D1,D3) = 0.19507908611599298
JS(D1,D4) = 0.007704160246533128
JS(D2,D3) = 0.17636986301369864
JS(D2,D4) = 0.00916030534351145
JS(D3,D4) = 0.012080536912751677

2 Min Hashing (50 points)

We will consider a hash family H so that any hash function $h \in H$ maps from $h : \{\text{k}-grams\} \rightarrow [m]$ for m large enough (To be extra cautious, I suggest over $m \geq$ 10,000; but should work with smaller m too).

A: (35 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:

$$\hat{\mathsf{JS}}_t(a, b) = \frac{1}{t} \sum_{i=1}^{t} \{ 1 \text{ if } a_i = b_i 0\text{if } a_i \neq b_i.$$

You should report 5 numbers:

t = 20: JS(D1, D2) = 1.0000000000000000
t = 60: JS(D1, D2) = 1.0000000000000000
t = 150: JS(D1, D2) = 0.8750000000000000
t = 300: JS(D1, D2) = 0.9333333333333333
t = 600: JS(D1, D2) = 0.9803921568627451

B: (15 point) What seems to be a good value for t? You may run more experiments. Justify your answer in terms of both accuracy and time.

A: When comparing the estimated Jaccard Similarities for all values of t, the estimation for $t = 300$ yielded the highest accuracy in estimating the true Jaccard Similarity. Experiments showed that computing 600 random hash functions for the $t = 600$ simulation required approximately 1.5x processing time as computing 300 random hash functions for $t = 300$. Furthermore, while additional simulations with larger values t failed to show improvements in accuracy, the processing time required for generating more and more random hash functions continued to increase.

Thus, we can conclude that given the possible values of t, $t = 300$ appears most optimal when considering both accuracy and time.