Data Mining - Asmt2 - Document Similarity and Hashing

Joyce Chen

February 14, 2018

1 Creating k-Grams

A: I have the following sizes for each document

1) D1 word based 2-grams: 279
2) D2 word based 2-grams: 278
3) D3 word based 2-grams: 337
4) D4 word based 2-grams: 232
5) D1 character based 3-grams: 765
6) D2 character based 3-grams: 762
7) D3 character based 3-grams: 828
8) D4 character based 3-grams: 698
9) D1 character based 2-grams: 263
10) D2 character based 2-grams: 262
11) D3 character based 2-grams: 269
12) D4 character based 2-grams: 255

B: B.1 2-grams based on words

1) $\mathbf{JS}(D_1, D_2) = 0.941$
2) $\mathbf{JS}(D_1, D_3) = 0.182$
3) $\mathbf{JS}(D_1, D_4) = 0.030$
4) $\mathbf{JS}(D_2, D_3) = 0.174$
5) $\mathbf{JS}(D_2, D_4) = 0.030$
6) $\mathbf{JS}(D_3, D_4) = 0.016$

B.2 2-grams based on characters

1) $\mathbf{JS}(D_1, D_2) = 0.981$
2) $\mathbf{JS}(D_1, D_3) = 0.816$
3) $\mathbf{JS}(D_1, D_4) = 0.644$
4) $\mathbf{JS}(D_2, D_3) = 0.8$
5) $\mathbf{JS}(D_2, D_4) = 0.641$
6) $\mathbf{JS}(D_3, D_4) = 0.653$

B.3 3-grams based on characters

1) $\mathbf{JS}(D_1, D_2) = 0.978$
2) $\mathbf{JS}(D_1, D_3) = 0.580$
3) $\mathbf{JS}(D_1, D_4) = 0.305$
4) $\mathbf{JS}(D_2, D_3) = 0.568$
5) $\mathbf{JS}(D_2, D_4) = 0.306$
6) $\mathbf{JS}(D_3, D_4) = 0.312$

2 Min Hashing

A:
1) $t = 20$ $\mathbf{JS}(D_1, D_2) = 0.95$
2) $t = 60$ $\mathbf{JS}(D_1, D_2) = 0.983$
3) $t = 150$ $\mathbf{JS}(D_1, D_2) = 0.98$
4) $t = 300$ $\mathbf{JS}(D_1, D_2) = 0.967$
5) $t = 600$ $\mathbf{JS}(D_1, D_2) = 0.973$

B: I think $t = 150$ is a good number

3 LSH

A: Using the trick mentioned in class, we can solve for b as $b \approx -\log_{0.7} 160 = 14.22$

B: Let $b = 14$, $t = 160$ then we have $r = t/b = 10$