Data Mining - Asmt2 - Document Similarity and Hashing

Joyce Chen

February 6, 2018

1 Creating k-Grams

A.1 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

2 Min Hashing

3 LSH

A