Assignment 2 *

Ishaan Kapoor

January 25, 2022

1 Creating k-Grams

A

- Document 1
    1. Bigram: 240
    2. Trigram: 253
    3. Char-Gram: 637
- Document 2
    1. Bigram: 221
    2. Trigram: 233
    3. Char-Gram: 599
- Document 3
    1. Bigram: 390
    2. Trigram: 423
    3. Char-Gram: 978
- Document 4
    1. Bigram: 364
    2. Trigram: 381
    3. Char-Gram: 770

B

- Bigram:

X	$D1$	$D2$	$D3$	$D4$
$D1$	1	0.913	0.112	0.010
$D2$	X	1	0.010	0.010
$D3$	X	X	1	0.012
$D4$	X	X	X	1

*CS 6140 Data Mining; Spring 2022 Instructor: Qingyao Ai, University of Utah

- Trigram:

X	$D1$	$D2$	$D3$	$D4$
$D1$	1	0.906	0.002	0.0
$D2$	X	1	0.002	0.0
$D3$	X	X	1	0.001
$D4$	X	X	X	1

- Chargram:

X	$D1$	$D2$	$D3$	$D4$
$D1$	1	0.940	0.279	0.291
$D2$	X	1	0.271	0.287
$D3$	X	X	1	0.313
$D4$	X	X	X	1

2 Min Hashing

A

t=100; 0.92

t=200; 0.94

t=400; 0.9325

t=800; 0.935

t=1600; 0.934375

A graph between $|error|vstime$ is shown below,

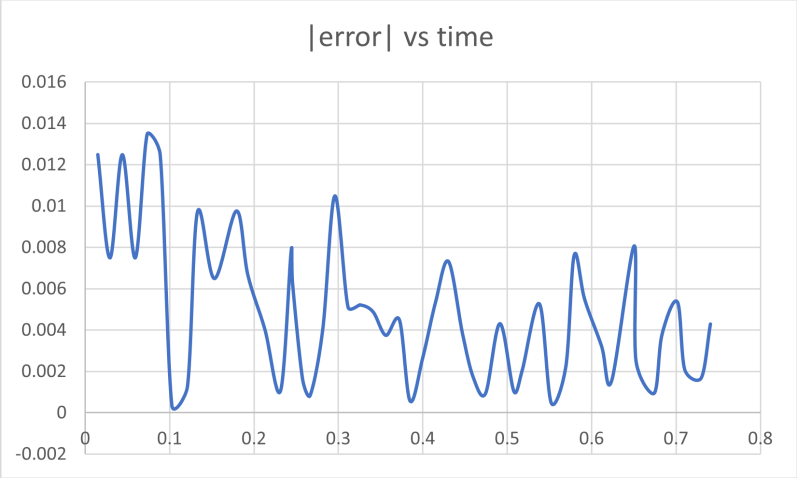

Figure 1: N vs Time for M=500

Keeping a range of 5% acceptable error limit, we get that after **k=4500**, out error has always been $< 5\%$