NTMI - Project Exercises - Part A

Alex Khawalid (10634207)
Wessel Kleinsma (10172432)
Winand Renkema (10643478)

February 4, 2016

Step 1: Extracting n-gram statistics

1 Unigrams

Frequencies for $n = 1$ and $m = 10$:

rank	*unigram*	*frequency*
1	the	20829
2	to	20042
3	and	18331
4	of	17949
5	a	11135
6	her	11007
7	I	10381
8	was	9409
9	in	9182
10	it	7573

The sum for $n = 1$ and $m = 10$ is 617091.

2 Bigrams

Frequencies for $n = 2$ and $m = 10$:

rank	*bigram*	*frequency*
1	of the	2507
2	to be	2233
3	in the	1917
4	I am	1366
5	of her	1264
6	to the	1142
7	it was	1010
8	had been	995
9	she had	978
10	to her	964

The sum for $n = 1$ and $m = 10$ is 617090.

3 Trigrams

Frequencies for $n = 3$ and $m = 10$:

rank	*trigram*	*frequency*
1	I do not	378
2	I am sure	366
3	in the world	214
4	she could not	202
5	would have been	189
6	I dare say	174
7	a great deal	173
8	as soon as	173
9	it would be	171
10	could not be	155

The sum for $n = 3$ and $m = 10$ is 617089.