Homework 1 - Language Model

Alexandra Pawlak

February 1, 2018

1 Total number of word tokens in data set: 26,155

2 Vocabulary size (number of unique words): 4,644

3 Top Tens

Positive Bigrams	Negative Bigrams
('of', 'the') 95	('of', 'the') 64
('in', 'the') 58	('in', 'the') 60
('the', 'movie') 47	('the', 'film') 42
('this', 'movie') 39	('to', 'be') 31
('and', 'the') 31	('is', 'the') 26
('is', 'a') 27	('is', 'a') 25
('the', 'film') 26	('of', 'a') 25
('a', 'great') 26	('in', 'a') 24
('in', 'a') 24	('and', 'the') 24
('to', 'the') 23	('this', 'film') 24

Positive Trigrams	Negative Trigrams
('one', 'of', 'the') 10	('this', 'movie', 'is') 8
('lackawanna', 'blues', 'is') 9	('a', 'lot', 'of') 7
('this', 'is', 'a') 8	('of', 'the', 'movie') 6
('some', 'of', 'the') 8	('one', 'of', 'the') 6
('s', 'epatha', 'merkerson') 8	('to', 'be', 'the') 5
('it', 'was', 'a') 7	('sound', 'of', 'music') 5
('in', 'the', 'movie') 7	('of', 'this', 'film') 5
('this', 'movie', 'was') 6	('would', 'have', 'been') 5
('of', 'the', 'film') 6	('a', 'couple', 'of') 5
('the', 'coast', 'guard') 6	('there', 'are', 'a') 5

4 Trigram Language Model

Test Case	Probability
('the', 'amazing', 'performance')	6.540436247097682e-05
('a', 'couple', 'of')	0.00029431963111939565
('showcase', 'his', 'wife')	6.540436247097682e-05
('in', 'the', 'theater')	0.00013080872494195363
('had', 'struck', 'it')	3.270218123548841e-05