

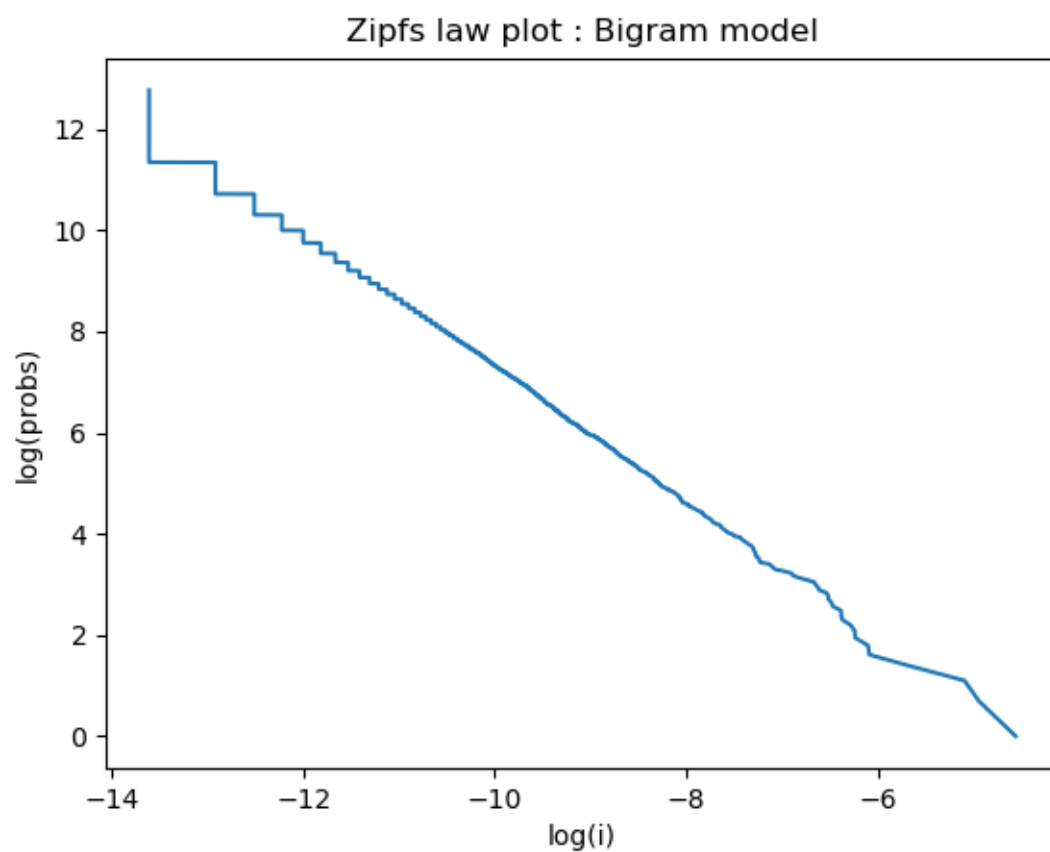
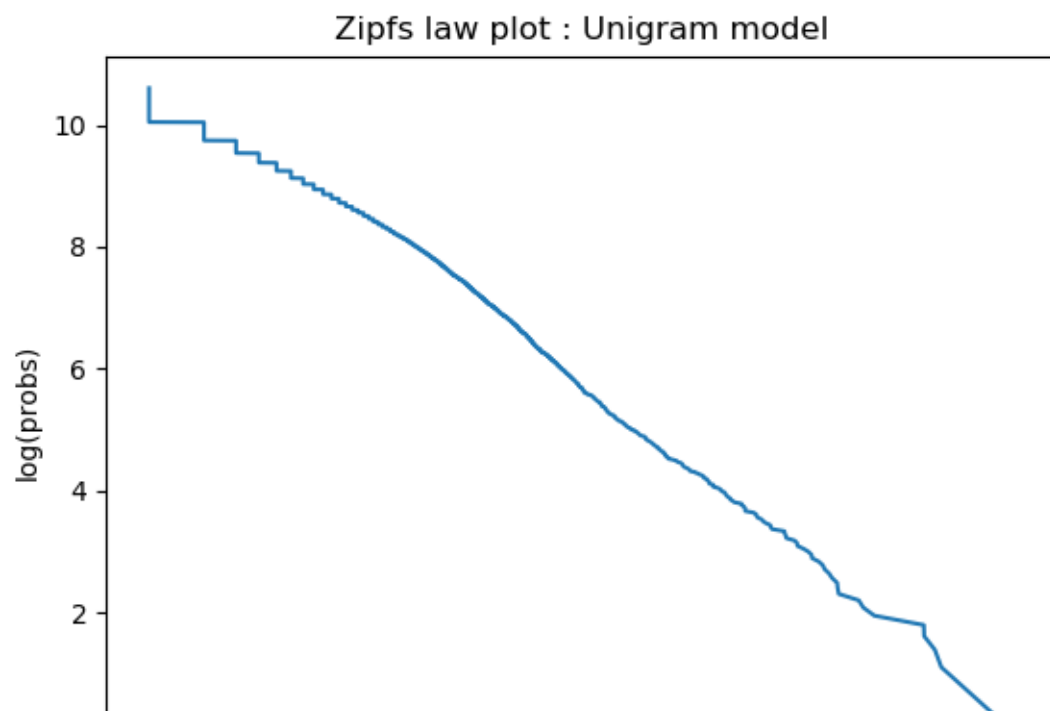
Assignment 1

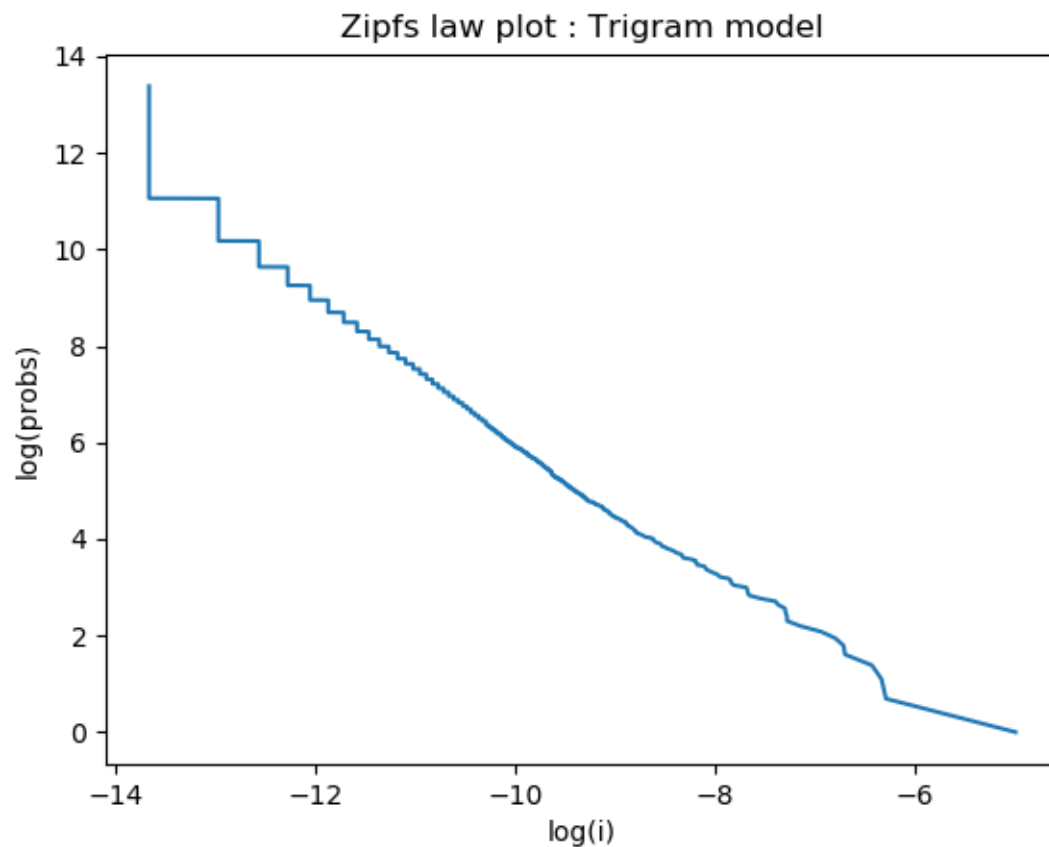
Speech & Natural Language Processing

Name: Prerit Gupta Roll No: 14EC35001

Task 1

Plots for Zipf's Law verification





Top 10 Results:

The top 10 unigrams with their probabilities are: [('the', 0.07281461348840149), ('of', 0.040344207969516106), ('and', 0.02849738593370475), ('to', 0.026238698500701083), ('a', 0.02293515851445186), ('in', 0.022838412907669862), ('is', 0.012220905048701738), ('that', 0.010629117331781963), ('for', 0.010046063808242467), ('it', 0.0078054355551714395)]

The top 10 bigrams with their probabilities are: [('of', 'the'), 0.010442223501385055], [(None, 'the'), 0.00711612739316297], [('in', 'the'), 0.006118298560696345], [('to', 'the'), 0.003459876357593379], [('and', 'the'), 0.0022681275306252447], [('on', 'the'), 0.002234989303716759], [('for', 'the'), 0.0019527007041259547], [(None, 'in'), 0.0019453366537018468], [(None, 'it'), 0.0018606500738246054], [('it', 'is'), 0.0017060050149183387]]

The top 10 trigrams with their probabilities are: [(None, None, 'the'), 0.006786771531143884], [(None, None, 'in'), 0.001855300599665929], [(None, None, 'it'), 0.001774533570406024], [(None, None, 'he'), 0.0016118289752302738], [(None, None, 'this'), 0.0012314045620495629], [(None, None, 'but'), 0.0012150170488663939], [(None, None, 'a'), 0.0011178625064233198], [(None, None, 'and'), 0.0009727159610866794], [(None, None, 'i'), 0.0007901122427599382], [(None, None, 'they'), 0.000690616627004983]]

Log-Likelihood & Perplexity of sentences:

Unigram test probabilities

The sequence : he lived a good life => has unigram log-likelihood score of 32.695722922858934
The sequence : he lived a good life => has unigram perplexity score of 691.6946386300403
The sequence : the man was happy => has unigram log-likelihood score of 23.98335809740223
The sequence : the man was happy => has unigram perplexity score of 401.75382457091143
The sequence : the person was good => has unigram log-likelihood score of 23.3121455528641
The sequence : the person was good => has unigram perplexity score of 339.6910025020692
The sequence : the girl was sad => has unigram log-likelihood score of 27.22149005169346
The sequence : the girl was sad => has unigram perplexity score of 902.6839687493552
The sequence : he won the war => has unigram log-likelihood score of 24.63036419583692
The sequence : he won the war => has unigram perplexity score of 472.2889823411059

Bigram test probabilities

The sequence : he lived a good life => has bigram log-likelihood score of 26.75326271469513
The sequence : he lived a good life => has bigram perplexity score of 210.7457736739031
The sequence : the man was happy => has bigram log-likelihood score of 21.953222134470852
The sequence : the man was happy => has bigram perplexity score of 241.84705775513535
The sequence : the person was good => has bigram log-likelihood score of 24.780403139829854
The sequence : the person was good => has bigram perplexity score of 490.34086149062426
The sequence : the girl was sad => has bigram log-likelihood score of 26.984388279254013
The sequence : the girl was sad => has bigram perplexity score of 850.7319272269048
The sequence : he won the war => has bigram log-likelihood score of 20.89959618128855
The sequence : he won the war => has bigram perplexity score of 185.8424079184589

Trigram test probabilities

The sequence : he lived a good life => has trigram log-likelihood score of 46.468383490958054
The sequence : he lived a good life => has trigram perplexity score of 10869.073025017655
The sequence : the man was happy => has trigram log-likelihood score of 32.02612302124965
The sequence : the man was happy => has trigram perplexity score of 3000.4896029339907
The sequence : the person was good => has trigram log-likelihood score of 32.087998424967736
The sequence : the person was good => has trigram perplexity score of 3047.2645741356764
The sequence : the girl was sad => has trigram log-likelihood score of 34.28522300230396
The sequence : the girl was sad => has trigram perplexity score of 5278.017066507729
The sequence : he won the war => has trigram log-likelihood score of 15.801813809578833
The sequence : he won the war => has trigram perplexity score of 51.95892239155092

Task 2: Laplacian Smoothing

Log-Likelihood & Perplexity of sentences:

For $k = 0.1$

Unigram test probabilities

The sequence : he lived a good life => has unigram log-likelihood score of 32.71964731816952
The sequence : he lived a good life => has unigram perplexity score of 695.0122446658966
The sequence : the man was happy => has unigram log-likelihood score of 24.00223228357562
The sequence : the man was happy => has unigram perplexity score of 403.6539981943964
The sequence : the person was good => has unigram log-likelihood score of 23.332027540151564
The sequence : the person was good => has unigram perplexity score of 341.3836386869796

The sequence : the girl was sad => has unigram log-likelihood score of 27.235367207424353
The sequence : the girl was sad => has unigram perplexity score of 905.8210788760942
The sequence : he won the war => has unigram log-likelihood score of 24.64930006451991
The sequence : he won the war => has unigram perplexity score of 474.5300833425606

Bigram test probabilities

The sequence : he lived a good life => has bigram log-likelihood score of 35.838764622037694
The sequence : he lived a good life => has bigram perplexity score of 1296.9270329507299
The sequence : the man was happy => has bigram log-likelihood score of 28.747683616949132
The sequence : the man was happy => has bigram perplexity score of 1322.0262866788507
The sequence : the person was good => has bigram log-likelihood score of 30.82172609225505
The sequence : the person was good => has bigram perplexity score of 2220.3753187208945
The sequence : the girl was sad => has bigram log-likelihood score of 36.2954697044596
The sequence : the girl was sad => has bigram perplexity score of 8724.299227804993
The sequence : he won the war => has bigram log-likelihood score of 28.173684013940488
The sequence : he won the war => has bigram perplexity score of 1145.298984212637

Trigram test probabilities

The sequence : he lived a good life => has trigram log-likelihood score of 54.56664170348356
The sequence : he lived a good life => has trigram perplexity score of 54903.278352509566
The sequence : the man was happy => has trigram log-likelihood score of 42.7576017543395
The sequence : the man was happy => has trigram perplexity score of 43888.18544507569
The sequence : the person was good => has trigram log-likelihood score of 44.12966736069827
The sequence : the person was good => has trigram perplexity score of 61846.874325573524
The sequence : the girl was sad => has trigram log-likelihood score of 48.10782481718814
The sequence : the girl was sad => has trigram perplexity score of 167201.70957261024
The sequence : he won the war => has trigram log-likelihood score of 38.16573639833114
The sequence : he won the war => has trigram perplexity score of 13924.90300975093

Task 3 : Good Turing

Good Turing Test can't be done for Unigram Model.

Bigram test probabilities

The sequence : he lived a good life => has log-likelihood score of 27.135262320430837
The sequence : he lived a good life => has bigram perplexity score of 227.47775743304288
The sequence : the man was happy => has log-likelihood score of 21.061918703588557
The sequence : the man was happy => has bigram perplexity score of 193.53912257762696
The sequence : the person was good => has log-likelihood score of 24.766378168127645
The sequence : the person was good => has bigram perplexity score of 488.6246178557191
The sequence : the girl was sad => has log-likelihood score of 33.32538090908281
The sequence : the girl was sad => has bigram perplexity score of 4151.999179668591
The sequence : he won the war => has log-likelihood score of 20.94849406349335
The sequence : he won the war => has bigram perplexity score of 188.12817564644624

Trigram test probabilities

The sequence : he lived a good life => has log-likelihood score of 109.94285357446269
The sequence : he lived a good life => has bigram perplexity score of 3544173111.6282783

The sequence : the man was happy => has log-likelihood score of 79.67507633499226
The sequence : the man was happy => has bigram perplexity score of 447312978.7521031
The sequence : the person was good => has log-likelihood score of 81.04714194135103
The sequence : the person was good => has bigram perplexity score of 630349815.1159791
The sequence : the girl was sad => has log-likelihood score of 85.0252993978409
The sequence : the girl was sad => has bigram perplexity score of 1704137320.8506625
The sequence : he won the war => has log-likelihood score of 38.16573639833114
The sequence : he won the war => has bigram perplexity score of 13924.90300975093

Task 4: Interpolation

Interpolation Parameter Lamda=0.2

Bigram test probabilities

The sequence ("he lived a good life") has bigram probability of $1.071610616455513e-13$ log-likelihood score of 43.08528445437992 and perplexity score of 392.6382437694933
The sequence ("the man was happy") has bigram probability of $1.2171557019347933e-09$ log-likelihood score of 29.613839120576444 and perplexity score of 169.30253678595378
The sequence ("the person was good") has bigram probability of $2.418555370810555e-10$ log-likelihood score of 31.94513538141056 and perplexity score of 253.57766085138243
The sequence ("the girl was sad") has bigram probability of $1.056889450408542e-09$ log-likelihood score of 29.817528373848724 and perplexity score of 175.385065494533
The sequence ("he won the war") has bigram probability of $3.767696162146401e-09$ log-likelihood score of 27.983670227421673 and perplexity score of 127.63830604388745