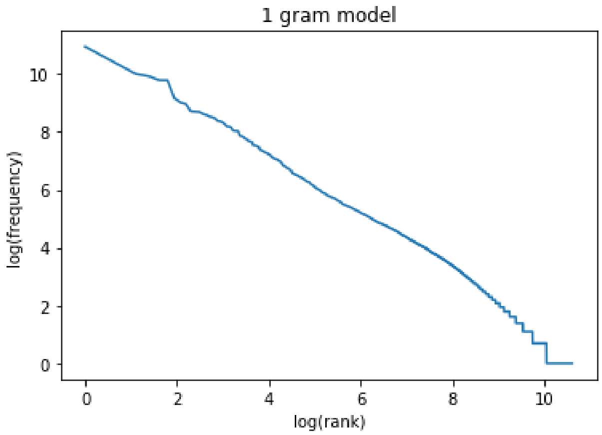
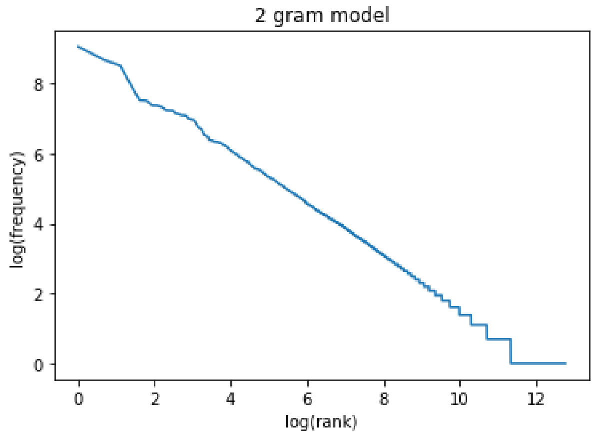
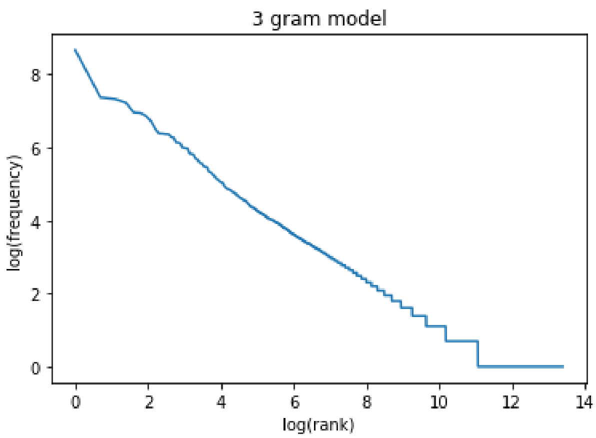
SUB PART 1 - Zipf’s law verification

From Zipf’s law : $f \propto \frac{1}{r}$
$f * r = constant$
$\log f + \log r = constant$
$\log f = -\log r + constant$







SUB PART 2 - Top 10 N-grams

Top 10 unigrams:

((’the’,), 56448)
((’of’,), 31276)
((’and’,), 22092)
((’to’,), 20341)
((’a’,), 17780)
((’in’,), 17705)
((’is’,), 9474)
((’that’,), 8240)
((’for’,), 7788)
((’it’,), 6051)

Top 10 bigrams:

((’of’, ’the’), 8508)
((’<s>’, ’the’), 5798)
((’in’, ’the’), 4985)
((’to’, ’the’), 2819)
((’and’, ’the’), 1848)
((’on’, ’the’), 1821)
((’for’, ’the’), 1591)
((’<s>’, ’in’), 1585)

(('<s>', 'it'), 1516)
(('it', 'is'), 1390)

Top 10 trigrams:

(('<s>', '<s>', 'the'), 5798)
(('<s>', '<s>', 'in'), 1585)
(('<s>', '<s>', 'it'), 1516)
(('<s>', '<s>', 'he'), 1377)
(('<s>', '<s>', 'this'), 1052)
(('<s>', '<s>', 'but'), 1038)
(('<s>', '<s>', 'a'), 955)
(('<s>', '<s>', 'and'), 831)
(('<s>', '<s>', 'i'), 675)
(('<s>', '<s>', 'they'), 590)

SUB PART 3 - Log Likelihood and Perplexity Score over testcases

Below models are run on these test examples :
'he', 'lived', 'a', 'good', 'life',
'the', 'man', 'was', 'happy',
'the', 'person', 'was', 'good',
'the', 'girl', 'was', 'sad',
'he', 'won', 'the', 'war'

Loglikelihood for unigram model:

[-32.695722922858934, -23.98335809740223, -23.3121455528641, -27.22149005169346, -24.63036419583692]

Loglikelihood for bigram model:

[-26.764829351066595, -21.96478877084232, -24.79196977620132, -inf, -20.911162817660017]

Loglikelihood for trigram model:

[-inf, -inf, -inf, -inf, -15.8133804459503]

Perplexity for unigram model:

[691.6946386300403, 401.7538245709114, 339.6910025020692, 902.6839687493552, 472.2889823411059]

Perplexity for bigram model:

[211.23386195638668, 242.54740909943231, 491.7608121184974, inf, 186.38057853513374]

Perplexity for trigram model:

[inf, inf, inf, inf, 52.10938732373954]

Assignment Part 2 - Laplace Smoothing

Below models are run on these test examples :
'he', 'lived', 'a', 'good', 'life',
'the', 'man', 'was', 'happy',
'the', 'person', 'was', 'good',
'the', 'girl', 'was', 'sad',
'he', 'won', 'the', 'war'

Inference: Performance degrades with increasing Laplace constant.

Laplace Constant : 0.0001

Loglikelihood for unigram model:

[-32.69622080463882, -23.983756137106077, -23.312544601677335, -27.221883078197745, -24.63076229749562]

Loglikelihood for bigram model:

[-26.847000338323966, -22.04080516475117, -24.82913097230666, -34.54853096746562, -20.9850856420229]

Loglikelihood for trigram model:

[-47.8915834234167, -34.5871168009303, -34.89888025718764, -36.09573830333439, -17.27574747294035]

Perplexity for unigram model:

[691.7635184909501, 401.79380505344375, 339.7248925153661, 902.7726677880961, 472.33598943708193]

Perplexity for bigram model:

[214.73400318122648, 247.20088141894368, 496.3507048378182, 5637.14348481119, 189.8570480895546]

Perplexity for trigram model:

[14448.078149827297, 5691.785079959569, 6153.153934950185, 8299.367415113416, 75.10873534881343]

Laplace Constant : 0.001

Loglikelihood for unigram model:

[-32.700699492443, -23.987336695904045, -23.316134242341775, -27.22541851966222, -24.634343413862727]

Loglikelihood for bigram model:

[-27.43733427550682, -22.56990027778077, -25.128025453815177, -33.56978787088357, -21.51652571427152]

Loglikelihood for trigram model:

[-47.201864441926844, -34.685712569520305, -35.63918497440786, -38.880587408502116, -21.08025653193206]

Perplexity for unigram model:

[692.383434657327, 402.1536276610809, 340.0299019263935, 903.5709454998163, 472.75905132369127]

Perplexity for bigram model:

[241.64431085517262, 282.16022107735506, 534.8607129932185, 4413.604439036361, 216.83385791525822]

Perplexity for trigram model:

[12586.40935747779, 5833.8249356770375, 7404.152301169509, 16649.68959223056, 194.42843132083914]

Laplace Constant : 0.01

Loglikelihood for unigram model:

[-32.74526526995445, -24.022965406941886, -23.351853760362058, -27.260596202099656, -24.669977698158974]

Loglikelihood for bigram model:

[-29.983911494563305, -24.587277059727192, -26.76700409808823, -34.62959186365357, -23.703684713295782]

Loglikelihood for trigram model:

[-49.44978385768447, -37.89464136040634, -39.248444210996446, -43.215646180882594, -28.38495743109203]

Perplexity for unigram model:

[698.5823407456458, 405.7516819755038, 343.07992588385025, 911.5523730188164, 476.9894744678491]

Perplexity for bigram model:

[402.1327664572624, 467.2288894659451, 805.7318739120949, 5752.546848573044, 374.6233713914516]

Perplexity for trigram model:

[19731.20697074056, 13012.430086346243, 18253.48049257068, 49212.92371990654, 1207.4178592813382]

Laplace Constant : 0.1

Loglikelihood for unigram model:

[-33.17018211297322, -24.36266011941858, -23.692455375994527, -27.595795043267316, -25.009727900362872]

Loglikelihood for bigram model:

[-35.8495199189156, -28.758360263885002, -30.832403184049262, -37.21257746942101, -28.18436812273654]

Loglikelihood for trigram model:

[-54.57954540427349, -42.7726299059804, -44.145679558334244, -48.11915191542619, -38.17659109752681]

Perplexity for unigram model:

[760.5458678190245, 441.71506719706025, 373.57315318842655, 991.2321456347264, 519.2741520913679]

Perplexity for bigram model:

[1299.7198026342494, 1325.5597021854867, 2226.3100236447062, 10972.466487763368, 1148.3621980930875]

Perplexity for trigram model:

[55045.152439092526, 44053.3851580754, 62094.94661083059, 167675.8581441685, 13962.741986243716]

Laplace Constant : 1

Loglikelihood for unigram model:

[-36.14067651249974, -26.736702333402473, -26.075451496218648, -29.9262637706019, -27.384301158680216]

Loglikelihood for bigram model:

[-44.3500178740693, -35.363691042750645, -37.02933915448583, -41.31255313976911, -34.930157427888325]

Loglikelihood for trigram model:

[-60.903863045566524, -49.16147186907789, -50.01093973039269, -53.317835736873185, -48.11679229963744]

Perplexity for unigram model:

[1377.651204726965, 799.6511606238381, 677.8071948050097, 1775.0181774082757, 940.1837074906048]

Perplexity for bigram model:

[7115.306409020292, 6911.367236773566, 10481.161570590215, 30581.125531655765, 6201.4558437203]

Perplexity for trigram model:

[195003.4654864407, 217590.02833977327, 269072.1753907771, 615050.0553061387, 167576.97466234677]

Assignment Part 3 - Good-Turing Method

Below models are run on these test examples :
'he', 'lived', 'a', 'good', 'life',
'the', 'man', 'was', 'happy',
'the', 'person', 'was', 'good',
'the', 'girl', 'was', 'sad',
'he', 'won', 'the', 'war'

Inference:
Good-Turing Smoothing cannot be used on unigram model as there is no unigram space given apart from the sentences so n_0 is not defined.
Naive implementation of this method fails mainly because of 2 ways:
1. Cases where $n_{r+1} = 0$
2. Case where $n_{r+1} = 0$ because of r being the highest frequency.

Loglikelihood for bigram model:

[-64.66112197799661, -49.7286251058532, -52.74306757907499, -64.93660412816753, -50.06590743140063]

Loglikelihood for trigram model:

[-83.51836438435444, -49.47587549895942, -50.951782018768995, -55.0461265809911, 14.082017000983676]

Perplexity for bigram model:

[413422.1048283198, 250736.1014729557, 532728.3413866702, 11230580.665836733, 272795.2677518144]

Perplexity for trigram model:

[17960274.0004349, 235382.91079962245, 340423.4302016377, 947452.0854818762, 0.029584513407896272]

Assignment Part 4 - Interpolation Method

Below models are run on these test examples :
['he', 'lived', 'a', 'good', 'life']
['the', 'man', 'was', 'happy']
['the', 'person', 'was', 'good']
['the', 'girl', 'was', 'sad']
['he', 'won', 'the', 'war']

Inference: The performance increases with increasing λ value.
Inference: λ for bigram interpolation; λ_1 and λ_2 for trigram interpolation.

Using Interpolation smoothing with $\lambda =$ ***0.2*** $\lambda_1 =$ ***0.2*** $\lambda_2 =$ ***0.2***

Loglikelihood for bigram model:

[-32.50176303469126, -26.307329753326425, -27.29349415930528, -30.578362817575908, -26.051731897229526]

Loglikelihood for trigram model:

[-32.40744875947381, -26.948191267019766, -27.360427250077173, -31.354051634437806, -22.402272658399003]

Perplexity for bigram model:

[665.3762079530336, 718.2605874221946, 919.0803401670271, 2089.313254071866, 673.7997596791181]

Perplexity for trigram model:

[652.942945062462, 843.068166684751, 934.5889551988855, 2536.4301244527505, 270.5800977942687]

Using Interpolation smoothing with $\lambda =$ ***0.5*** $\lambda_1 =$ ***0.3*** $\lambda_2 =$ ***0.3***

Loglikelihood for bigram model:

[-29.404015414509253, -24.099194050692976, -25.976132630151266, -29.407964177638043, -23.41901670054331]

Loglikelihood for trigram model:

[-30.57114439767812, -25.367706231413056, -26.541422718781856, -30.427582796046813, -20.353242966203926]

Perplexity for bigram model:

[358.09670760701704, 413.55830687011303, 661.1846547780045, 1559.2980708643472, 348.889123889969]

Perplexity for trigram model:

[452.24717582525255, 567.8893298100351, 761.549885488744, 2012.0224697231204, 162.11577517394662]

Using Interpolation smoothing with $\lambda =$ ***0.8*** $\lambda_1 =$ ***0.5*** $\lambda_2 =$ ***0.5***

Loglikelihood for bigram model:

[-27.631288323142133, -22.697596275693712, -25.19093612101781, -29.277646866527117, -21.763434174325365]

Loglikelihood for trigram model:

[-28.387170091579307, -23.159988735204468, -25.797936161975993, -inf, -17.63330008044401]

Perplexity for bigram model:

[251.2020673931444, 291.3132267645548, 543.3393236825085, 1509.3158031576652, 230.64011307156326]

Perplexity for trigram model:

[292.19869065451803, 327.0121034435539, 632.3759282524966, inf, 82.13177914880677]