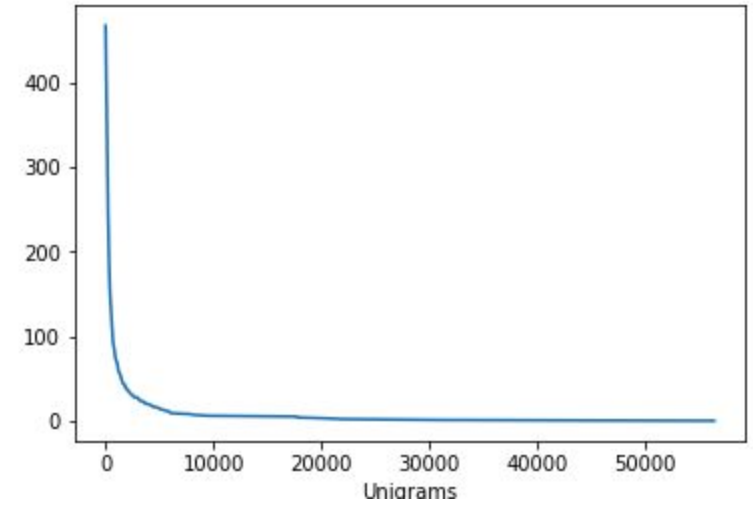
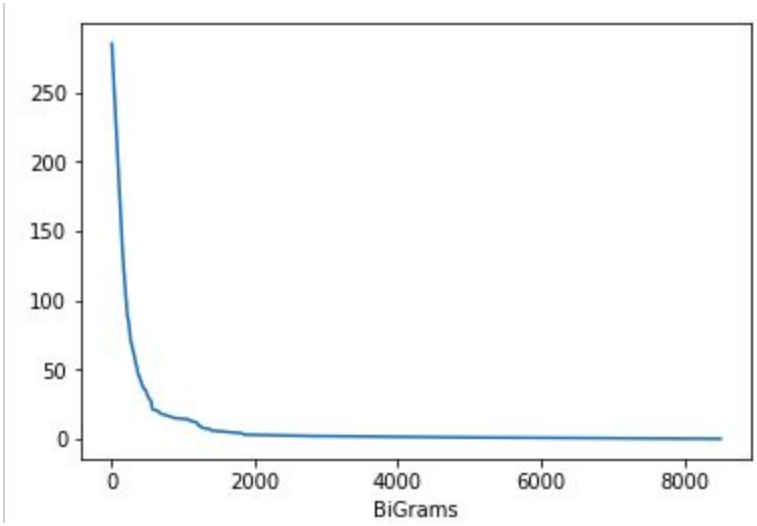
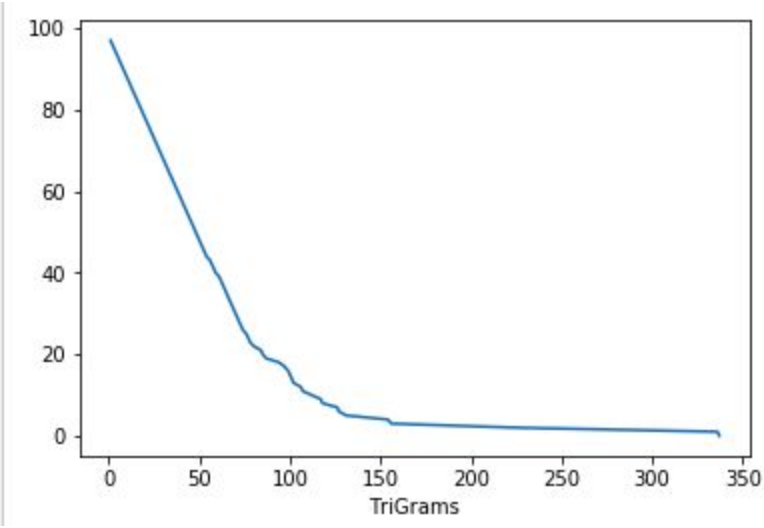
NLP- Assignment 1

Nilay Pochhi (15CS10033)

Zipf's law plots:







As we can observe in the above plots, the zipf's law is being followed.

Top 10 UniGrams, BiGrams, TriGrams:

```
Top 10 unigrams:
[['it', 6051], ['for', 7788], ['that', 8240], ['is', 9474], ['in', 17705], ['a', 17780], ['to', 20341], ['and', 220
92], ['of', 31276], ['the', 56448]]
Top 10 bigrams:
[[('that', 'the'), 1243], [('with', 'the'), 1261], [('to', 'be'), 1373], [('it', 'is'), 1390], [('for', 'the'), 159
1], [('on', 'the'), 1821], [('and', 'the'), 1848], [('to', 'the'), 2819], [('in', 'the'), 4985], [('of', 'the'), 85
08]]
Top 10 trigrams:
[[('a', 'number', 'of'), 118], [('there', 'is', 'a'), 118], [('it', 'is', 'not'), 126], [('of', 'the', 'united'), 1
27], [('part', 'of', 'the'), 131], [('the', 'fact', 'that'), 154], [('some', 'of', 'the'), 156], [('as', 'well', 'a
s'), 225], [('the', 'united', 'states'), 336], [('one', 'of', 'the'), 337]]
```

As discussed in the piazza forum, I've excluded the NGrams which have more one or more of its constituents as padding.

No Smoothing, log-likelihood and perplexity:

The first value is log likelihood and the second one is perplexity.

```
Unigram:
he lived a good life -33.1867769216 763.0742839869649
the man was happy -24.3762012964 443.21293660835187
the person was good -23.7049887519 374.74552213454615
the girl was sad -27.6143332507 995.836724257661
he won the war -25.0232073949 521.0269921256837
BiGram:
he lived a good life -26.7648293511 211.23386195638668
the man was happy -21.9647887708 242.54740909943231
the person was good -24.7919697762 491.7608121184974
the girl was sad -inf inf
he won the war -20.9111628177 186.38057853513374
TriGram:
he lived a good life -inf inf
the man was happy -inf inf
the person was good -inf inf
the girl was sad -inf inf
he won the war -15.813380446 52.10938732373954
```

Laplacian Smoothing:

```
k = 1
Unigram:
he lived a good life -32.9292159586 724.7619033300211
the man was happy -24.1675338903 420.68463714562273
the person was good -23.5062830531 356.5843305708292
the girl was sad -27.3570953275 933.8107848268966
he won the war -24.8151327155 494.6167295341016
BiGram:
he lived a good life -44.3498928537 7115.128499565497
the man was happy -35.3635977697 6911.206077643159
the person was good -37.0292458497 10480.917087693024
the girl was sad -41.3124594596 30580.409328683014
he won the war -34.9300635354 6201.310277908409
TriGram:
he lived a good life -60.9037032594 194997.23381528916
the man was happy -49.1613389841 217582.7998459531
the person was good -50.0108068131 269063.2344515137
the girl was sad -53.3177027896 615029.6133435647
he won the war -48.1166571628 167571.31330150258
```

```
k = 0.1
Unigram:
he lived a good life -32.7196486014 695.0124230403729
the man was happy -24.0022333102 403.6541017919533
the person was good -23.3320285667 341.38372630288603
the girl was sad -27.235368234 905.8213113545316
he won the war -24.6493010911 474.5302051304223
BiGram:
he lived a good life -35.8494344802 1299.6975935539438
the man was happy -28.7583010413 1325.5400765669362
the person was good -30.8323438132 2226.2769794076376
the girl was sad -37.2125147778 10972.294518623836
he won the war -28.184306413 1148.3444819369065
TriGram:
he lived a good life -54.5794028262 55043.5828155792
the man was happy -42.7725192825 44052.166839725425
the person was good -44.1455686212 62093.224476055475
the girl was sad -48.1190406815 167671.19539977444
he won the war -38.1764725258 13962.328095773271

k = 0.01
Unigram:
he lived a good life -32.698121455 692.0265285978893
the man was happy -23.985250355 401.94392496460034
the person was good -23.3141387084 339.8603089330983
the girl was sad -27.2228811502 902.9979539163033
he won the war -24.6322626462 472.5131898430037
BiGram:
he lived a good life -29.9838669709 402.1291855907283
the man was happy -24.5872445725 467.2250947397638
the person was good -26.7669735874 805.7257280670077
the girl was sad -34.6295461206 5752.48106425096
he won the war -23.7036487425 374.6200025316436
TriGram:
he lived a good life -49.4496678392 19730.749139281008
the man was happy -37.8945466925 13012.122125011789
the person was good -39.2483471437 18253.037544048268
the girl was sad -43.215546376 49211.69581212003
he won the war -28.3848616988 1207.3889624204016
```

```
k = 0.001
Unigram:
he lived a good life -32.6959628359 691.7278287430521
the man was happy -23.9835473707 401.77283533771725
the person was good -23.3123449171 339.70793348582333
the girl was sad -27.2216291945 902.7153697795078
he won the war -24.6305540887 472.31140394427615
BiGram:
he lived a good life -27.4373206269 241.6436512352182
the man was happy -22.5698884104 282.1593839530842
the person was good -25.1280179532 534.859710049933
the girl was sad -33.5697628705 4413.57685371697
he won the war -21.5165134941 216.83319547855663
TriGram:
he lived a good life -47.2017829084 12586.204116329242
the man was happy -34.6856424728 5833.722703564382
the person was good -35.6391091745 7404.011993975359
the girl was sad -38.8804962989 16649.310359794963
he won the war -21.0801971298 194.42554397533553
```

```
k = 0.0001
Unigram:
he lived a good life -32.6957469148 691.6979576525048
the man was happy -23.9833770252 401.75572565486704
the person was good -23.3121654898 339.69269560385123
the girl was sad -27.2215039663 902.6871088775094
he won the war -24.6303831856 472.29122450995345
BiGram:
he lived a good life -26.8469983574 214.73391810616047
the man was happy -22.040803343 247.20076883371536
the person was good -24.8291300616 496.3505918311702
the girl was sad -34.548525185 5637.135335603173
he won the war -20.9850838623 189.8569636186671
TriGram:
he lived a good life -47.8915323574 14447.930589500462
the man was happy -34.5870869695 5691.7426315750145
the person was good -34.8988415534 6153.094397582347
the girl was sad -36.0956742375 8299.234489638739
he won the war -17.2757229037 75.10827400927897
```

Good Turing Smoothing, BiGrams:

```
he lived a good life -64.661121978 413422.1048283198
the man was happy -49.7286251059 250736.1014729557
the person was good -52.7430675791 532728.3413866702
the girl was sad -64.9365546777 11230441.827289544
he won the war -50.0659074314 272795.2677518144
```

Good Turing Smoothing, TriGrams:

```
he lived a good life -140.2265573 1513296439791.893
the man was happy -inf inf
the person was good -inf inf
the girl was sad -inf inf
he won the war -73.8812939912 105089560.08851086
```

Interpolation Smoothing:

```
Lambda = 0.2
he lived a good life -32.0852857245 612.1988480817503
the man was happy -25.6183363188 604.6102766917562
the person was good -26.2422673576 706.6721506981247
the girl was sad -29.9139552803 1769.564623825298
he won the war -25.5692105969 597.2302092654292
Lambda = 0.5
he lived a good life -29.3119222832 351.56142836996855
the man was happy -23.8867597388 392.1678491266264
the person was good -25.6069522309 602.8919888573471
the girl was sad -29.2633950797 1503.9477600827925
he won the war -23.2816473356 337.11085872188795
Lambda = 0.8
he lived a good life -27.6168494967 250.4777011855223
the man was happy -22.6422585478 287.31082306927254
the person was good -25.0903629082 529.8482935818453
the girl was sad -29.3113125166 1522.0724347971013
he won the war -21.7283852077 228.62801677756246
```