ASSIGNMENT-3

Karan Mangla
201301205

LAPLACE SMOOTHING

$$P(w_i|w_{i-1}) = (1 + c(w_{i-1}w_i)) / (|V| + c(w_{i-1}))$$

Laplace add-one smoothing helps assign probability to unseen words .
When calculating linear interpolation, we calculate Laplace smoothing for trigrams, bigrams and unigrams and assign them weights 0.5, 0.3 and 0.2 respectively.

GOOD TURING SMOOTHING

$$P(\text{any given word w/freq. r}) = (n_{r+1}\ (r+1)) / (N\ n_r)$$

$$r^* = (r + 1)(n_{r+1} / n_r)$$

n_r = # of word types with r training tokens

In good turing smoothing we take help of high frequency words to compute (approximately) the probabilities of lower frequency words .

If the token doesn't exist in the dictionary, we take the frequency of bin with r=1. If r is the maximum possible, we take $n_{r+1} = n_r$ so that r^* becomes r.

Similar to Laplace, when calculating linear interpolation, we calculate Good Turing smoothing for trigrams, bigrams and unigrams and assign them weights 0.5, 0.3 and 0.2 respectively.

Observations:

- Linear Interpolation gives bigger output for Laplace Smoothing but the reverse is true for Good Turing.
- Unigram Values < Bigram Values < Trigram Values in general .
- Values of Good Turing are much more higher than Laplace Smoothing.

The likelihood of sentences given in ToyTestData.txt

RESULTS :

ENGLISH

- LAPLACE SMOOTHING
  7.04255961319e-41
  6.43265057625e-51
  3.33232469621e-115
  3.59404123032e-126
  4.74087485842e-86
  1.21986562232e-161
  5.22363400865e-160
  9.13435056012e-45
  1.03401323887e-138
  6.73574560467e-46
  6.73600346125e-46

- LAPLACE INTERPOLATION
  4.11347659921e-41
  1.27460385866e-44
  2.70365884093e-78
  5.58421534745e-67
  1.96716442612e-47
  2.96629444657e-87
  7.16703319195e-78
  4.5709441269e-25
  7.80944669842e-69
  5.04444898125e-30
  2.13655665119e-33

- GOOD TURING SMOOTHING
  0.381977652089
  0.300294545371
  4.29476270845e-07
  1.09622081875e-28
  1.35547663478e-08
  2.21617841554e-36
  2.28863871417e-61
  9.25457536999e-13
  6.751246388e-32
  3.54857575402e-08
  3.54857575402e-08

- GOOD TURING INTERPOLATION

0.190988837161
0.150147272687
2.14738135422e-07
5.48110409377e-29
6.7773831739e-09
1.10808920777e-36
5.49542411157e-48
4.6272876867e-13
3.375623194e-32
1.77428787701e-08
1.77428787701e-08

HINDI

- LAPLACE SMOOTHING
  7.94508166325e-224
  8.45677233892e-67
  3.32654634562e-122
  6.72967575858e-230
  5.41114467107e-99
  3.06010950828e-112
  6.99687239052e-72
  2.18813975993e-121
  1.10403527859e-106
  1.8447854256e-107

- LAPLACE INTERPOLATION
  2.15302861488e-114
  8.76500789023e-48
  2.67922730667e-67
  9.56331477394e-127
  7.25955974941e-59
  1.07467152753e-68
  1.28142761554e-47
  5.11775302695e-76
  2.26578356966e-58
  1.15914307164e-64

- GOOD TURING SMOOTHING
  9.08117636957e-91
  0.163927562298
  2.63966531851e-35

6.83460637512e-72
8.66465343105e-25
2.4065346224e-08
0.142639758412
8.04101322252e-21
1.02273947485e-48
6.7264173733e-09

- GOOD TURING INTERPOLATION
  1.01595512593e-84
  0.0819637811488
  1.31983265925e-35
  3.41730318756e-72
  4.33232671553e-25
  1.2032673112e-08
  0.0713198792059
  4.02050661126e-21
  2.44384988542e-42
  3.36320868665e-09