

NLP Assignment 01

Contents

Problem Assumption.....	2
Data Preperation.....	2
Implementation Assumptions.....	2
Smoothing Methods	2
Laplace Smoothing:.....	2
Backoff Smoothing	2
Uniform Interpolation	2
Interpolation with Smoothing.....	3
Frequency based Interpolation	3
Code Comments.....	3
How to Run the Program	4
Experimental Results and Discussion.....	4
Performance Measure	4
Average Discrimination Performance (ADP).....	4
Average Perplexity Difference (APD)	4
Minimum Perplexity(MP).....	5
Validating obtained counts and Probabilities	5
Results and Discussion	7

Problem Assumption

In this part I describe briefly the assumptions I made during the solving the assignment. I will describe the Data preparation phase in my code implementation and also other assumptions I made to build the language model.

Data Preperation

In this part I describe the assumptions that I made for build the n-grams. I should mention that for the accuracy reasons I processed the whole text file as a single string and the I substitute all uninformative characters such as “(“ , “.”, “?”, “\n” , ... which are exist in any languages with the space. Also, in the preprocessing phase I searched the whole obtained new string and replace all multiple spaces with just a single space. (This has been implemented in the `preProcessDoc()` module in the `languageModel.py`)

Implementation Assumptions

For implementing all getPerplexity Functions I actually compute the log of the perplexities. It is just because of computational issues preventing to get overflow in multiplication of invers probabilities. Also I should mention that in the process of adding the log of inverse probabilities when I found out a term has zero probability I just ignore that term and continue to add the others’. That Is why we do not get zero parobabilty for computing the unsmooth get peprelxity function. Because there are lots of terms which are not in the training and as a result their count is zero. However In computing the perplexity function which is $PP(W) = P(w_1, \dots, w_N)^{-\frac{1}{N}}$ I used the $PP(W) = P(w_1, \dots, w_N)^{-\frac{1}{C}}$ in which C is the number of terms which has positive (non-zero) probability. This way I prevent myself to get exception error in the computation unsmooth perplexity and also Backoff perplexity.

Smoothing Methods

In my experiments I have implemented unsmooth version of finding perplexity. Besides, I have implemented five other smoothing methods.

Laplace Smoothing:

Described in the class and exist in the textbook. This method implemented in the `getPerplexityLaplace` method of the `LanguageModel` Class

Backoff Smoothing

Described in the class and exist in the textbook. This method implemented in the `getPerplexityBackoff` method of the `LanguageModel` Class

Uniform Interpolation

The interpolation method described in the class and exist in the textbook. However I should mention that, in this method I used uniform Coefficients ‘s’. This method implemented in the `getPerplexityInterpolate` method of the `LanguageModel` Class

Interpolation with Smoothing

In this variation of Interpolation method I just used Laplace smoothing for finding the probabilities of the substring for example in the 3-gram case in uniform interpolation we have

$P(W_2|W_0W_1) = 1/3 * P(W_2|W_0W_1) + 1/3 * P(W_2|W_1) + 1/3 * P(W_2)$ and we use unsmooth method to find three related probabilities. However, In this method we use Laplace Estimate to find these there probabilities. Actually It would help in the cases when the letter W_2 do not exist in the Vocabulary at all which in this case all probabilities will be equal to zero in the unsmoothed version. This method implemented in the `getPerplexityInterpolateSmooth` method of the `LanguageModel` Class.

Frequency based Interpolation

In this variation of the Interpolation method I used another justifiable heuristic weight vector. When we want to compute the amount of $P(W_n|W_j....W_{n-1})$ if the count of the sequence $S = W_j....W_{n-1}$ is high in our corpus, we can conclude intuitively that the reliability of the probabilities conditioned on this sequence is high. So intuitively we can say that the related weight of this term is should be an increasing function of count of sequence S. SO we propuse the weight C for this term in which C is the count of substring S in the corpus.

On the other hand, as the length of S decreases, intuitively speaking it is obvious that its count will goes up in our corpus. For considering this fact we refine the related weight of S to $C^{\frac{1}{len(S)+1}}$. the summation with one is because we don't want that the unigrams gets just one weight to put some weight for unigrams too. So this way we are saying that as C goes up the significance of the related term is much more in weighted sum. Also as the length of S goes up the importance of the related term will increase as well. This method implemented in the `getPerplexityInterpolateFreq` method of the `LanguageModel` Class.

Code Comments

The code is written using the Object Oriented Programming. I also tried to use defensive programming preventing users from bad input to the modules. I have a class language model

For running the program you should just run the `main.py` in this file at first I hard coded the N-gram of the Language Model which is assumed to be built using setting the Parameter N. I have Hard coded N = 3. For implementing the N-gram but if you set it to 10 then the program run experiments by building a 10-gram model and running the experiments using this model. So by changing Parameter N you will define the order of N-gram of the Language Model. Some Parts of the results in the Experimental results section is obtained using the N=15.

In the code I have file `LanguageModel.py` in which I have defined the `LanguageModel` Class and its related methods.

How to Run the Program

One should first unzip the file in a directory and then run the main.py using the python. Also I should mention that the train and test files should be located in the directory in which the python files are located. The results will be printed in the console. Also, the program will build tree text file in which you can see the statistics of unigram, bigram and trigrams of the three languages English, Spanish and German created in the same directory with the names as follows: *outEnFile.txt*, *outEsFile.txt*, *outDeFile.txt*

Experimental Results and Discussion

In this section I will show the experimental results and discuss about the obtained results as well. First, I should mention that for having more broad view I used N=15 it means that I run experiments for all unigram, bigram, trigram, 4-gram Up to 15-gram model. As you will see this will help me to have better understanding of the results.

Performance Measure

For comparison of every two method we need some performance measure. I have defined three intuitive performance measures for comparing the Language Models and smoothing Methods which I named them : Average Discrimination Performance (ADP) and the other one is Average Perplexity Difference (APD) which is defined as below

Average Discrimination Performance (ADP)

Since we know that our test data is from the English document in this measure we define whether our method gets less perplexity on English Language Model than Spanish and German Language Models. So we can predict that the text is in English. For this let the obtained perplexity of test data using the English, Spanish and German Language Models respectively would be *pEn* , *pEs* , *pDe* then the ADP would be equal to :

$$ADP = ((\text{sign}(pEs - pEn) + \text{sign}(pDe - pEn)) / 2)$$

It would be equal to 1 if the perplexity of English is less than both. It would be equal to 0 if just one of Spanish or German gets better perplexity and it would be -1 if the English is getting worse perplexity which is not favorable at all for us.

Average Perplexity Difference (APD)

This performance measure is defined as well:

$$APD = ((pEs - pEn) + (pDe - pEn)) / 2.0$$

In comparing different Language Models and Smoothing Methods I used these measures intuitively the model that has more ADP would be better also the Model that has more APD would be better as well.

Minimum Perplexity(MP)

In this measure the method which gets minimum perplexity for English Language model is best.

Validating obtained counts and Probabilities

As you can see below for the english unigram the sum of probabilities is equal to one. It is the same for Spanish and German unigrams as well and it has been created in the output files of the program which are `outEnFile.txt`, `outEsFile.txt` , `outDeFile.txt`

-----1gram Counts For English Language model-----

;	3	1.96001568013e-05
	26047	0.170175094734
"	12	7.8400627205e-05
„	5	3.26669280021e-05
'	104	0.000679472102443
©	2	1.30667712008e-05
–	1	6.53338560042e-06
ˆ	5	3.26669280021e-05
°	3	1.96001568013e-05
Ã	16	0.000104534169607
Â	3	1.96001568013e-05
a	9359	0.0611459558343
c	4050	0.0264602116817
b	1594	0.0104142166471
e	16206	0.10588004704
d	4195	0.0274075525938
g	2063	0.0134783744937
f	2682	0.0175225401803
i	10058	0.065712792369
h	5870	0.0383509734745
k	628	0.00410296615706
j	198	0.00129361034888
m	3791	0.0247680648112
l	4723	0.0308571801908
o	10384	0.0678426760747
n	9288	0.0606820854567
q	152	0.000993074611264

p	3352	0.0218999085326
s	8393	0.0548347053443
r	8159	0.0533058931138
u	3690	0.0241081928655
t	12459	0.0813994511956
w	2000	0.0130667712008
v	1339	0.00874820331896
y	1924	0.0125702338952
x	258	0.00168561348491
z	44	0.000287468966418

-----**Total Probabilities = 1.0 for 1gram Model**-----

Also for validating the bigram probabilities I just randomly pick the sum of bigrams starting by letter g:

g'	3	0.001454193
gy	17	0.008240427
g	653	0.316529326
gg	13	0.006301503
gf	3	0.001454193
ge	264	0.127968977
gd	9	0.004362579
gi	295	0.142995637
gh	126	0.061076103
go	114	0.055259331
gs	16	0.007755696
gr	295	0.142995637
gu	114	0.055259331
ga	103	0.04992729
gn	24	0.011633543
gl	7	0.003393117
gt	7	0.003393117

---**Total Probabilities = 1.0 for bigrams Starting by letter g**---

Also for validating the trigram probabilities we examined the total sum of trigrams started with 'th' as required in the assignment

thu	15	0.003922
ths	10	0.002614
th	220	0.057516
thy	7	0.00183

thw	2	0.000523
thr	22	0.005752
tho	70	0.018301
thl	2	0.000523
thi	465	0.121569
the	2526	0.660392
thc	1	0.000261
tha	481	0.125752
thd	4	0.001046
SUM		1

Results and Discussion

In the following you can see the perplexity results of different methods on different language model:

```

Building [1, 2, 3]-Gram Models
Computing Perplexities of Test Data, using 1-Gram Model
Perplexity without smoothing of Test Data using the 1-gram English Language Model is: 18.0897705664
Perplexity without smoothing of Test Data using the 1-gram Spanish Language Model is: 21.4930272081
Perplexity without smoothing of Test Data using the 1-gram German Language Model is: 20.4374792775

Perplexity with Laplace smoothing of Test Data using the 1-gram English Language Model is: 18.0882592283
Perplexity with Laplace smoothing of Test Data using the 1-gram Spanish Language Model is: 22.0132797752
Perplexity with Laplace smoothing of Test Data using the 1-gram German Language Model is: 20.9528461374

Perplexity with Backoff smoothing of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Backoff smoothing of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Backoff smoothing of Test Data using the 1-gram German Language Model is: 20.4374792479

Perplexity with Interpolate of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Interpolate of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Interpolate of Test Data using the 1-gram German Language Model is: 20.4374792479

Perplexity with Interpolate smoothing of Test Data using the 1-gram English Language Model is: 18.0882592283
Perplexity with Interpolate smoothing of Test Data using the 1-gram Spanish Language Model is: 22.0132797752
Perplexity with Interpolate smoothing of Test Data using the 1-gram German Language Model is: 20.9528461374

Perplexity with Frequency Interpolate of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Frequency Interpolate of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Frequency Interpolate of Test Data using the 1-gram German Language Model is: 20.4374792479
.....

Computing Perplexities of Test Data, using 2-Gram Model
Perplexity without smoothing of Test Data using the 2-gram English Language Model is: 11.0022020272
Perplexity without smoothing of Test Data using the 2-gram Spanish Language Model is: 21.4469402934
Perplexity without smoothing of Test Data using the 2-gram German Language Model is: 19.7388665175

Perplexity with Laplace smoothing of Test Data using the 2-gram English Language Model is: 11.2300279343
Perplexity with Laplace smoothing of Test Data using the 2-gram Spanish Language Model is: 23.792166946
Perplexity with Laplace smoothing of Test Data using the 2-gram German Language Model is: 21.4875984933

Perplexity with Backoff smoothing of Test Data using the 2-gram English Language Model is: 11.0912239593
Perplexity with Backoff smoothing of Test Data using the 2-gram Spanish Language Model is: 22.3778827907
Perplexity with Backoff smoothing of Test Data using the 2-gram German Language Model is: 20.3431773493

Perplexity with Interpolate of Test Data using the 2-gram English Language Model is: 12.8106662032
Perplexity with Interpolate of Test Data using the 2-gram Spanish Language Model is: 18.5240266657
Perplexity with Interpolate of Test Data using the 2-gram German Language Model is: 17.6050940041

Perplexity with Interpolate smoothing of Test Data using the 2-gram English Language Model is: 12.8471546342
Perplexity with Interpolate smoothing of Test Data using the 2-gram Spanish Language Model is: 18.7744122927
Perplexity with Interpolate smoothing of Test Data using the 2-gram German Language Model is: 17.9383954222

Perplexity with Frequency Interpolate of Test Data using the 2-gram English Language Model is: 11.3017219204
Perplexity with Frequency Interpolate of Test Data using the 2-gram Spanish Language Model is: 23.3788748491

```

```

Perplexity with Frequency Interpolate of Test Data using the 2-gram German Language Model is: 21.3683942648
.....

Computing Perplexities of Test Data, using 3-Gram Model
Perplexity without smoothing of Test Data using the 3-gram English Language Model is: 7.60426218086
Perplexity without smoothing of Test Data using the 3-gram Spanish Language Model is: 13.8307674168
Perplexity without smoothing of Test Data using the 3-gram German Language Model is: 14.5012765308

Perplexity with Laplace smoothing of Test Data using the 3-gram English Language Model is: 8.92705179122
Perplexity with Laplace smoothing of Test Data using the 3-gram Spanish Language Model is: 24.6785750582
Perplexity with Laplace smoothing of Test Data using the 3-gram German Language Model is: 24.27500961

Perplexity with Backoff smoothing of Test Data using the 3-gram English Language Model is: 8.05106764343
Perplexity with Backoff smoothing of Test Data using the 3-gram Spanish Language Model is: 20.3174233851
Perplexity with Backoff smoothing of Test Data using the 3-gram German Language Model is: 18.7458112517

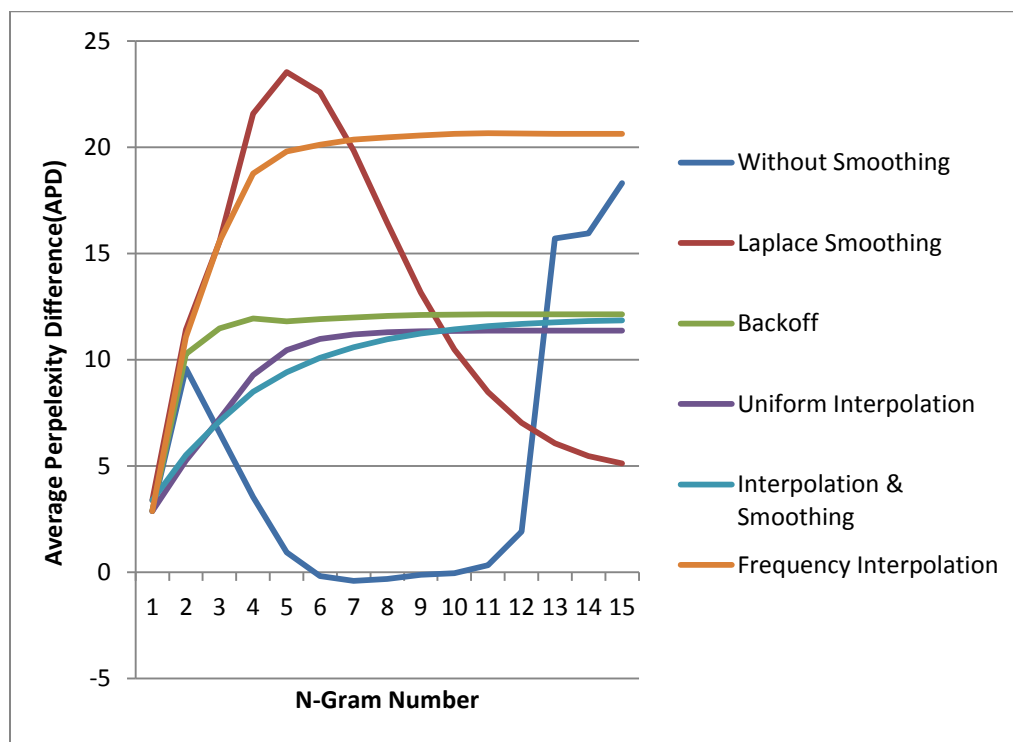
Perplexity with Interpolate of Test Data using the 3-gram English Language Model is: 9.83687458449
Perplexity with Interpolate of Test Data using the 3-gram Spanish Language Model is: 17.6378557334
Perplexity with Interpolate of Test Data using the 3-gram German Language Model is: 16.4611153938

Perplexity with Interpolate smoothing of Test Data using the 3-gram English Language Model is: 10.0984397241
Perplexity with Interpolate smoothing of Test Data using the 3-gram Spanish Language Model is: 17.5104700969
Perplexity with Interpolate smoothing of Test Data using the 3-gram German Language Model is: 16.8967267151

Perplexity with Frequency Interpolate of Test Data using the 3-gram English Language Model is: 8.96812423003
Perplexity with Frequency Interpolate of Test Data using the 3-gram Spanish Language Model is: 25.2555653525
Perplexity with Frequency Interpolate of Test Data using the 3-gram German Language Model is: 23.8089269604
.....

```

Since the inference in big picture is more precise and simpler I execute the experiments for N=15 to build all unigram, bigram, trigram up to 15-gram language model and then used different smoothing method. The results have been shown as the figure below. In this figure the horizontal axes shows number of N-gram from 1 up to 15 and the vertical axes shows the Average Perplexity Difference (APD) as defined in the Performance Measure Part. And intuitively speaking the method with bigger is APD is better.



These conclusions can be made on this figure:

- 1- As N increases the unsmooth method get worst results because there are lots of zero probabilities in computing Perplexity and for N in [6-10] using unsmooth method we can not get better perplexity for English Language model and we predict the test data as different language! However when N get very large more than 11 the unsmooth get better and better that is because almost all probabilities get zero and if it just find one n-gram for the test set which is possible it will get much more weight to English language model. However it is a kind of obvious improvement which is actually useless.
- 2- It seems that there is no reason to use N-gram for $N > 6$
- 3- Laplace smoothing work best for $N < 6$ in the sense of APD
- 4- The Frequency interpolation is work much better than uniform interpolation and interpolation with smoothing
- 5- The Backoff , uniform interpolation and interpolation with smoothing methods work somehow the same in the APD sense
- 6- Performance of Laplace smoothing is dramatically decrease when N gets big values

Also the Performance of different methods in average Discrimination Performance (ADP) sense has been shown in the following table. One can infer that whole methods well in this sense except unsmooth method which loose its ability to detect right language as N increase.

N-Gram	Without Smoothing	Laplace Smoothing	Backoff	Uniform Interpolation	Interpolation & Smoothing	Frequency Interpolation
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	1	1	1	1	1
5	1	1	1	1	1	1
6	-1	1	1	1	1	1
7	-1	1	1	1	1	1
8	-1	1	1	1	1	1
9	-1	1	1	1	1	1
10	0	1	1	1	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	1	1	1	1	1	1

Also Considering the Minimum Perplexity(M) as our last Performance measure one can infer that the Backoff, uniform interpolation and Frequency Interpolation gets best results as N increases and the better results for Back off is because for Large N lots of probabilities get zero and substituting them with lower order probabilities is very good approximation. However for the small $N=\{1,2,3\}$ the Laplace Smoothing method get better result. It is because when N gets large values again many lots of probabilities do not exist and using Laplace estimate is not very meaningful way of refining probabilities. In the following you can see the whole result for $N=15$

```
Building [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]-Gram Models
Computing Perplexities of Test Data, using 1-Gram Model
Perplexity without smoothing of Test Data using the 1-gram English Language Model is: 18.0897705664
Perplexity without smoothing of Test Data using the 1-gram Spanish Language Model is: 21.4930272081
Perplexity without smoothing of Test Data using the 1-gram German Language Model is: 20.4374792775

Perplexity with Laplace smoothing of Test Data using the 1-gram English Language Model is: 18.0882592283
Perplexity with Laplace smoothing of Test Data using the 1-gram Spanish Language Model is: 22.0132797752
Perplexity with Laplace smoothing of Test Data using the 1-gram German Language Model is: 20.9528461374

Perplexity with Backoff smoothing of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Backoff smoothing of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Backoff smoothing of Test Data using the 1-gram German Language Model is: 20.4374792479

Perplexity with Interpolate of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Interpolate of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Interpolate of Test Data using the 1-gram German Language Model is: 20.4374792479

Perplexity with Interpolate smoothing of Test Data using the 1-gram English Language Model is: 18.0882592283
Perplexity with Interpolate smoothing of Test Data using the 1-gram Spanish Language Model is: 22.0132797752
Perplexity with Interpolate smoothing of Test Data using the 1-gram German Language Model is: 20.9528461374

Perplexity with Frequency Interpolate of Test Data using the 1-gram English Language Model is: 18.0897705368
Perplexity with Frequency Interpolate of Test Data using the 1-gram Spanish Language Model is: 21.4930272051
Perplexity with Frequency Interpolate of Test Data using the 1-gram German Language Model is: 20.4374792479
.....

Computing Perplexities of Test Data, using 2-Gram Model
Perplexity without smoothing of Test Data using the 2-gram English Language Model is: 11.0022020272
Perplexity without smoothing of Test Data using the 2-gram Spanish Language Model is: 21.4469402934
Perplexity without smoothing of Test Data using the 2-gram German Language Model is: 19.7388665175

Perplexity with Laplace smoothing of Test Data using the 2-gram English Language Model is: 11.2300279343
Perplexity with Laplace smoothing of Test Data using the 2-gram Spanish Language Model is: 23.792166946
Perplexity with Laplace smoothing of Test Data using the 2-gram German Language Model is: 21.4875984933

Perplexity with Backoff smoothing of Test Data using the 2-gram English Language Model is: 11.0912239593
Perplexity with Backoff smoothing of Test Data using the 2-gram Spanish Language Model is: 22.3778827907
Perplexity with Backoff smoothing of Test Data using the 2-gram German Language Model is: 20.3431773493

Perplexity with Interpolate of Test Data using the 2-gram English Language Model is: 12.8106662032
Perplexity with Interpolate of Test Data using the 2-gram Spanish Language Model is: 18.5240266657
Perplexity with Interpolate of Test Data using the 2-gram German Language Model is: 17.6050940041

Perplexity with Interpolate smoothing of Test Data using the 2-gram English Language Model is: 12.8471546342
Perplexity with Interpolate smoothing of Test Data using the 2-gram Spanish Language Model is: 18.7744122927
Perplexity with Interpolate smoothing of Test Data using the 2-gram German Language Model is: 17.9383954222

Perplexity with Frequency Interpolate of Test Data using the 2-gram English Language Model is: 11.3017219204
Perplexity with Frequency Interpolate of Test Data using the 2-gram Spanish Language Model is: 23.3788748491
Perplexity with Frequency Interpolate of Test Data using the 2-gram German Language Model is: 21.3683942648
.....

Computing Perplexities of Test Data, using 3-Gram Model
Perplexity without smoothing of Test Data using the 3-gram English Language Model is: 7.60426218086
Perplexity without smoothing of Test Data using the 3-gram Spanish Language Model is: 13.8307674168
Perplexity without smoothing of Test Data using the 3-gram German Language Model is: 14.5012765308

Perplexity with Laplace smoothing of Test Data using the 3-gram English Language Model is: 8.92705179122
Perplexity with Laplace smoothing of Test Data using the 3-gram Spanish Language Model is: 24.6785750582
Perplexity with Laplace smoothing of Test Data using the 3-gram German Language Model is: 24.27500961
```

Perplexity with Backoff smoothing of Test Data using the 3-gram English Language Model is: 8.05106764343
 Perplexity with Backoff smoothing of Test Data using the 3-gram Spanish Language Model is: 20.3174233851
 Perplexity with Backoff smoothing of Test Data using the 3-gram German Language Model is: 18.7458112517

Perplexity with Interpolate of Test Data using the 3-gram English Language Model is: 9.83687458449
 Perplexity with Interpolate of Test Data using the 3-gram Spanish Language Model is: 17.6378557334
 Perplexity with Interpolate of Test Data using the 3-gram German Language Model is: 16.4611153938

Perplexity with Interpolate smoothing of Test Data using the 3-gram English Language Model is: 10.0984397241
 Perplexity with Interpolate smoothing of Test Data using the 3-gram Spanish Language Model is: 17.5104700969
 Perplexity with Interpolate smoothing of Test Data using the 3-gram German Language Model is: 16.8967267151

Perplexity with Frequency Interpolate of Test Data using the 3-gram English Language Model is: 8.96812423003
 Perplexity with Frequency Interpolate of Test Data using the 3-gram Spanish Language Model is: 25.2555653525
 Perplexity with Frequency Interpolate of Test Data using the 3-gram German Language Model is: 23.8089269604

Computing Perplexities of Test Data, using 4-Gram Model
 Perplexity without smoothing of Test Data using the 4-gram English Language Model is: 4.62357620526
 Perplexity without smoothing of Test Data using the 4-gram Spanish Language Model is: 8.40659969621
 Perplexity without smoothing of Test Data using the 4-gram German Language Model is: 7.93051251164

Perplexity with Laplace smoothing of Test Data using the 4-gram English Language Model is: 8.50716510727
 Perplexity with Laplace smoothing of Test Data using the 4-gram Spanish Language Model is: 30.3155744234
 Perplexity with Laplace smoothing of Test Data using the 4-gram German Language Model is: 29.8401090467

Perplexity with Backoff smoothing of Test Data using the 4-gram English Language Model is: 5.68060044546
 Perplexity with Backoff smoothing of Test Data using the 4-gram Spanish Language Model is: 18.3758772151
 Perplexity with Backoff smoothing of Test Data using the 4-gram German Language Model is: 16.8818386896

Perplexity with Interpolate of Test Data using the 4-gram English Language Model is: 7.8127769471
 Perplexity with Interpolate of Test Data using the 4-gram Spanish Language Model is: 17.5448461214
 Perplexity with Interpolate of Test Data using the 4-gram German Language Model is: 16.6247120985

Perplexity with Interpolate smoothing of Test Data using the 4-gram English Language Model is: 8.88996685947
 Perplexity with Interpolate smoothing of Test Data using the 4-gram Spanish Language Model is: 17.3521877956
 Perplexity with Interpolate smoothing of Test Data using the 4-gram German Language Model is: 17.4035195423

Perplexity with Frequency Interpolate of Test Data using the 4-gram English Language Model is: 8.16394258461
 Perplexity with Frequency Interpolate of Test Data using the 4-gram Spanish Language Model is: 28.2809800011
 Perplexity with Frequency Interpolate of Test Data using the 4-gram German Language Model is: 25.559945645

Computing Perplexities of Test Data, using 5-Gram Model
 Perplexity without smoothing of Test Data using the 5-gram English Language Model is: 2.96231657404
 Perplexity without smoothing of Test Data using the 5-gram Spanish Language Model is: 3.97348269016
 Perplexity without smoothing of Test Data using the 5-gram German Language Model is: 3.82182465528

Perplexity with Laplace smoothing of Test Data using the 5-gram English Language Model is: 10.9445446989
 Perplexity with Laplace smoothing of Test Data using the 5-gram Spanish Language Model is: 33.8157287255
 Perplexity with Laplace smoothing of Test Data using the 5-gram German Language Model is: 35.1428842886

Perplexity with Backoff smoothing of Test Data using the 5-gram English Language Model is: 4.55214249325
 Perplexity with Backoff smoothing of Test Data using the 5-gram Spanish Language Model is: 17.0659626832
 Perplexity with Backoff smoothing of Test Data using the 5-gram German Language Model is: 15.6563854915

Perplexity with Interpolate of Test Data using the 5-gram English Language Model is: 6.7259542939
 Perplexity with Interpolate of Test Data using the 5-gram Spanish Language Model is: 17.5979149912
 Perplexity with Interpolate of Test Data using the 5-gram German Language Model is: 16.7463563266

Perplexity with Interpolate smoothing of Test Data using the 5-gram English Language Model is: 8.77705155023
 Perplexity with Interpolate smoothing of Test Data using the 5-gram Spanish Language Model is: 17.9526189645
 Perplexity with Interpolate smoothing of Test Data using the 5-gram German Language Model is: 18.4198444968

Perplexity with Frequency Interpolate of Test Data using the 5-gram English Language Model is: 7.81615098115
 Perplexity with Frequency Interpolate of Test Data using the 5-gram Spanish Language Model is: 29.014220964
 Perplexity with Frequency Interpolate of Test Data using the 5-gram German Language Model is: 26.207529923

Computing Perplexities of Test Data, using 6-Gram Model
 Perplexity without smoothing of Test Data using the 6-gram English Language Model is: 2.26287502064
 Perplexity without smoothing of Test Data using the 6-gram Spanish Language Model is: 2.09411957526
 Perplexity without smoothing of Test Data using the 6-gram German Language Model is: 2.07610096531

Perplexity with Laplace smoothing of Test Data using the 6-gram English Language Model is: 15.0981175538
 Perplexity with Laplace smoothing of Test Data using the 6-gram Spanish Language Model is: 36.7522158301
 Perplexity with Laplace smoothing of Test Data using the 6-gram German Language Model is: 38.607986086

Perplexity with Backoff smoothing of Test Data using the 6-gram English Language Model is: 4.12891303215
 Perplexity with Backoff smoothing of Test Data using the 6-gram Spanish Language Model is: 16.6322919928
 Perplexity with Backoff smoothing of Test Data using the 6-gram German Language Model is: 15.4358705212

Perplexity with Interpolate of Test Data using the 6-gram English Language Model is: 6.22633763542
 Perplexity with Interpolate of Test Data using the 6-gram Spanish Language Model is: 17.5583765647
 Perplexity with Interpolate of Test Data using the 6-gram German Language Model is: 16.8364206612

Perplexity with Interpolate smoothing of Test Data using the 6-gram English Language Model is: 9.11255687543
 Perplexity with Interpolate smoothing of Test Data using the 6-gram Spanish Language Model is: 18.8700997278
 Perplexity with Interpolate smoothing of Test Data using the 6-gram German Language Model is: 19.5272884851

Perplexity with Frequency Interpolate of Test Data using the 6-gram English Language Model is: 7.68227542521
 Perplexity with Frequency Interpolate of Test Data using the 6-gram Spanish Language Model is: 29.1359146436
 Perplexity with Frequency Interpolate of Test Data using the 6-gram German Language Model is: 26.4652090451

Computing Perplexities of Test Data, using 7-Gram Model
 Perplexity without smoothing of Test Data using the 7-gram English Language Model is: 1.92061302216
 Perplexity without smoothing of Test Data using the 7-gram Spanish Language Model is: 1.57027215702
 Perplexity without smoothing of Test Data using the 7-gram German Language Model is: 1.47796393005

Perplexity with Laplace smoothing of Test Data using the 7-gram English Language Model is: 19.6741116938
 Perplexity with Laplace smoothing of Test Data using the 7-gram Spanish Language Model is: 38.8053592516
 Perplexity with Laplace smoothing of Test Data using the 7-gram German Language Model is: 40.2394925487

Perplexity with Backoff smoothing of Test Data using the 7-gram English Language Model is: 3.97513065959
 Perplexity with Backoff smoothing of Test Data using the 7-gram Spanish Language Model is: 16.546135095
 Perplexity with Backoff smoothing of Test Data using the 7-gram German Language Model is: 15.3751300992

Perplexity with Interpolate of Test Data using the 7-gram English Language Model is: 6.01167498951
 Perplexity with Interpolate of Test Data using the 7-gram Spanish Language Model is: 17.5547872397
 Perplexity with Interpolate of Test Data using the 7-gram German Language Model is: 16.8512597633

Perplexity with Interpolate smoothing of Test Data using the 7-gram English Language Model is: 9.62253380567
 Perplexity with Interpolate smoothing of Test Data using the 7-gram Spanish Language Model is: 19.842786301
 Perplexity with Interpolate smoothing of Test Data using the 7-gram German Language Model is: 20.5861488782

Perplexity with Frequency Interpolate of Test Data using the 7-gram English Language Model is: 7.62870600521
 Perplexity with Frequency Interpolate of Test Data using the 7-gram Spanish Language Model is: 29.2868223473
 Perplexity with Frequency Interpolate of Test Data using the 7-gram German Language Model is: 26.6716935031

Computing Perplexities of Test Data, using 8-Gram Model
 Perplexity without smoothing of Test Data using the 8-gram English Language Model is: 1.63120337403
 Perplexity without smoothing of Test Data using the 8-gram Spanish Language Model is: 1.28063427362
 Perplexity without smoothing of Test Data using the 8-gram German Language Model is: 1.3463736105

Perplexity with Laplace smoothing of Test Data using the 8-gram English Language Model is: 24.0816194996
 Perplexity with Laplace smoothing of Test Data using the 8-gram Spanish Language Model is: 39.9539033006
 Perplexity with Laplace smoothing of Test Data using the 8-gram German Language Model is: 41.0828250554

Perplexity with Backoff smoothing of Test Data using the 8-gram English Language Model is: 3.88309540437
 Perplexity with Backoff smoothing of Test Data using the 8-gram Spanish Language Model is: 16.5309023818
 Perplexity with Backoff smoothing of Test Data using the 8-gram German Language Model is: 15.3647685827

Perplexity with Interpolate of Test Data using the 8-gram English Language Model is: 5.91693594948
 Perplexity with Interpolate of Test Data using the 8-gram Spanish Language Model is: 17.5676825423
 Perplexity with Interpolate of Test Data using the 8-gram German Language Model is: 16.8544393981

Perplexity with Interpolate smoothing of Test Data using the 8-gram English Language Model is: 10.2077307773
 Perplexity with Interpolate smoothing of Test Data using the 8-gram Spanish Language Model is: 20.7756161223
 Perplexity with Interpolate smoothing of Test Data using the 8-gram German Language Model is: 21.5651151085

Perplexity with Frequency Interpolate of Test Data using the 8-gram English Language Model is: 7.60356509484
 Perplexity with Frequency Interpolate of Test Data using the 8-gram Spanish Language Model is: 29.4338556798
 Perplexity with Frequency Interpolate of Test Data using the 8-gram German Language Model is: 26.6918133153

Computing Perplexities of Test Data, using 9-Gram Model
 Perplexity without smoothing of Test Data using the 9-gram English Language Model is: 1.49092580621
 Perplexity without smoothing of Test Data using the 9-gram Spanish Language Model is: 1.31126487894
 Perplexity without smoothing of Test Data using the 9-gram German Language Model is: 1.44837319277

Perplexity with Laplace smoothing of Test Data using the 9-gram English Language Model is: 27.848299596
 Perplexity with Laplace smoothing of Test Data using the 9-gram Spanish Language Model is: 40.5377647687
 Perplexity with Laplace smoothing of Test Data using the 9-gram German Language Model is: 41.492679763

Perplexity with Backoff smoothing of Test Data using the 9-gram English Language Model is: 3.84215731632
 Perplexity with Backoff smoothing of Test Data using the 9-gram Spanish Language Model is: 16.5255139177
 Perplexity with Backoff smoothing of Test Data using the 9-gram German Language Model is: 15.3629887404

Perplexity with Interpolate of Test Data using the 9-gram English Language Model is: 5.87805973116
 Perplexity with Interpolate of Test Data using the 9-gram Spanish Language Model is: 17.5778281751

Perplexity with Interpolate of Test Data using the 9-gram German Language Model is: 16.8595254742

Perplexity with Interpolate smoothing of Test Data using the 9-gram English Language Model is: 10.8187816726
 Perplexity with Interpolate smoothing of Test Data using the 9-gram Spanish Language Model is: 21.6421358867
 Perplexity with Interpolate smoothing of Test Data using the 9-gram German Language Model is: 22.4591322932

Perplexity with Frequency Interpolate of Test Data using the 9-gram English Language Model is: 7.59371668125
 Perplexity with Frequency Interpolate of Test Data using the 9-gram Spanish Language Model is: 29.5785985183
 Perplexity with Frequency Interpolate of Test Data using the 9-gram German Language Model is: 26.7236026349

Computing Perplexities of Test Data, using 10-Gram Model
 Perplexity without smoothing of Test Data using the 10-gram English Language Model is: 1.41479958173
 Perplexity without smoothing of Test Data using the 10-gram Spanish Language Model is: 1.23726077233
 Perplexity without smoothing of Test Data using the 10-gram German Language Model is: 1.49637716313

Perplexity with Laplace smoothing of Test Data using the 10-gram English Language Model is: 30.8070634412
 Perplexity with Laplace smoothing of Test Data using the 10-gram Spanish Language Model is: 40.8466091668
 Perplexity with Laplace smoothing of Test Data using the 10-gram German Language Model is: 41.7206522987

Perplexity with Backoff smoothing of Test Data using the 10-gram English Language Model is: 3.82298369179
 Perplexity with Backoff smoothing of Test Data using the 10-gram Spanish Language Model is: 16.524108393
 Perplexity with Backoff smoothing of Test Data using the 10-gram German Language Model is: 15.3604143838

Perplexity with Interpolate of Test Data using the 10-gram English Language Model is: 5.8647545596
 Perplexity with Interpolate of Test Data using the 10-gram Spanish Language Model is: 17.5832454776
 Perplexity with Interpolate of Test Data using the 10-gram German Language Model is: 16.8626528401

Perplexity with Interpolate smoothing of Test Data using the 10-gram English Language Model is: 11.4265329942
 Perplexity with Interpolate smoothing of Test Data using the 10-gram Spanish Language Model is: 22.4386226022
 Perplexity with Interpolate smoothing of Test Data using the 10-gram German Language Model is: 23.2737822107

Perplexity with Frequency Interpolate of Test Data using the 10-gram English Language Model is: 7.60025930195
 Perplexity with Frequency Interpolate of Test Data using the 10-gram Spanish Language Model is: 29.7179186831
 Perplexity with Frequency Interpolate of Test Data using the 10-gram German Language Model is: 26.7266058189

Computing Perplexities of Test Data, using 11-Gram Model
 Perplexity without smoothing of Test Data using the 11-gram English Language Model is: 1.320707174
 Perplexity without smoothing of Test Data using the 11-gram Spanish Language Model is: 1.63748879055
 Perplexity without smoothing of Test Data using the 11-gram German Language Model is: 1.68708532572

Perplexity with Laplace smoothing of Test Data using the 11-gram English Language Model is: 32.9422939242
 Perplexity with Laplace smoothing of Test Data using the 11-gram Spanish Language Model is: 40.9653371082
 Perplexity with Laplace smoothing of Test Data using the 11-gram German Language Model is: 41.8700569794

Perplexity with Backoff smoothing of Test Data using the 11-gram English Language Model is: 3.81241221616
 Perplexity with Backoff smoothing of Test Data using the 11-gram Spanish Language Model is: 16.5240931344
 Perplexity with Backoff smoothing of Test Data using the 11-gram German Language Model is: 15.35995061

Perplexity with Interpolate of Test Data using the 11-gram English Language Model is: 5.86102025731
 Perplexity with Interpolate of Test Data using the 11-gram Spanish Language Model is: 17.5851889564
 Perplexity with Interpolate of Test Data using the 11-gram German Language Model is: 16.8648933623

Perplexity with Interpolate smoothing of Test Data using the 11-gram English Language Model is: 12.0175434484
 Perplexity with Interpolate smoothing of Test Data using the 11-gram Spanish Language Model is: 23.168369995
 Perplexity with Interpolate smoothing of Test Data using the 11-gram German Language Model is: 24.0179264546

Perplexity with Frequency Interpolate of Test Data using the 11-gram English Language Model is: 7.61552625229
 Perplexity with Frequency Interpolate of Test Data using the 11-gram Spanish Language Model is: 29.8010957792
 Perplexity with Frequency Interpolate of Test Data using the 11-gram German Language Model is: 26.7324375467

Computing Perplexities of Test Data, using 12-Gram Model
 Perplexity without smoothing of Test Data using the 12-gram English Language Model is: 1.25330589581
 Perplexity without smoothing of Test Data using the 12-gram Spanish Language Model is: 3.93011252639
 Perplexity without smoothing of Test Data using the 12-gram German Language Model is: 2.41036020495

Perplexity with Laplace smoothing of Test Data using the 12-gram English Language Model is: 34.436913081
 Perplexity with Laplace smoothing of Test Data using the 12-gram Spanish Language Model is: 40.9957249027
 Perplexity with Laplace smoothing of Test Data using the 12-gram German Language Model is: 41.9454580445

Perplexity with Backoff smoothing of Test Data using the 12-gram English Language Model is: 3.80272543674
 Perplexity with Backoff smoothing of Test Data using the 12-gram Spanish Language Model is: 16.5240931344
 Perplexity with Backoff smoothing of Test Data using the 12-gram German Language Model is: 15.3600617453

Perplexity with Interpolate of Test Data using the 12-gram English Language Model is: 5.86025531838
 Perplexity with Interpolate of Test Data using the 12-gram Spanish Language Model is: 17.5863926002
 Perplexity with Interpolate of Test Data using the 12-gram German Language Model is: 16.8665296232

Perplexity with Interpolate smoothing of Test Data using the 12-gram English Language Model is: 12.5863328316

Perplexity with Interpolate smoothing of Test Data using the 12-gram Spanish Language Model is: 23.8378300273
 Perplexity with Interpolate smoothing of Test Data using the 12-gram German Language Model is: 24.6996951428

 Perplexity with Frequency Interpolate of Test Data using the 12-gram English Language Model is: 7.63351745921
 Perplexity with Frequency Interpolate of Test Data using the 12-gram Spanish Language Model is: 29.8078644795
 Perplexity with Frequency Interpolate of Test Data using the 12-gram German Language Model is: 26.7341509737

 Computing Perplexities of Test Data, using 13-Gram Model
 Perplexity without smoothing of Test Data using the 13-gram English Language Model is: 1.1975131812
 Perplexity without smoothing of Test Data using the 13-gram Spanish Language Model is: 30.6205130461
 Perplexity without smoothing of Test Data using the 13-gram German Language Model is: 3.18847136574

 Perplexity with Laplace smoothing of Test Data using the 13-gram English Language Model is: 35.4217178335
 Perplexity with Laplace smoothing of Test Data using the 13-gram Spanish Language Model is: 41.0060289328
 Perplexity with Laplace smoothing of Test Data using the 13-gram German Language Model is: 41.9761394236

 Perplexity with Backoff smoothing of Test Data using the 13-gram English Language Model is: 3.7999668665
 Perplexity with Backoff smoothing of Test Data using the 13-gram Spanish Language Model is: 16.5240931344
 Perplexity with Backoff smoothing of Test Data using the 13-gram German Language Model is: 15.3589348213

 Perplexity with Interpolate of Test Data using the 13-gram English Language Model is: 5.86161920653
 Perplexity with Interpolate of Test Data using the 13-gram Spanish Language Model is: 17.5867070708
 Perplexity with Interpolate of Test Data using the 13-gram German Language Model is: 16.8668975515

 Perplexity with Interpolate smoothing of Test Data using the 13-gram English Language Model is: 13.1296320482
 Perplexity with Interpolate smoothing of Test Data using the 13-gram Spanish Language Model is: 24.4542501067
 Perplexity with Interpolate smoothing of Test Data using the 13-gram German Language Model is: 25.3264231798

 Perplexity with Frequency Interpolate of Test Data using the 13-gram English Language Model is: 7.63966386464
 Perplexity with Frequency Interpolate of Test Data using the 13-gram Spanish Language Model is: 29.8091254281
 Perplexity with Frequency Interpolate of Test Data using the 13-gram German Language Model is: 26.738163649

 Computing Perplexities of Test Data, using 14-Gram Model
 Perplexity without smoothing of Test Data using the 14-gram English Language Model is: 1.14395357211
 Perplexity without smoothing of Test Data using the 14-gram Spanish Language Model is: 30.6205130461
 Perplexity without smoothing of Test Data using the 14-gram German Language Model is: 3.55182843232

 Perplexity with Laplace smoothing of Test Data using the 14-gram English Language Model is: 36.0241182286
 Perplexity with Laplace smoothing of Test Data using the 14-gram Spanish Language Model is: 41.0033236048
 Perplexity with Laplace smoothing of Test Data using the 14-gram German Language Model is: 41.9860553128

 Perplexity with Backoff smoothing of Test Data using the 14-gram English Language Model is: 3.79865295592
 Perplexity with Backoff smoothing of Test Data using the 14-gram Spanish Language Model is: 16.5240931344
 Perplexity with Backoff smoothing of Test Data using the 14-gram German Language Model is: 15.357349379

 Perplexity with Interpolate of Test Data using the 14-gram English Language Model is: 5.86268918391
 Perplexity with Interpolate of Test Data using the 14-gram Spanish Language Model is: 17.5867070708
 Perplexity with Interpolate of Test Data using the 14-gram German Language Model is: 16.8668943409

 Perplexity with Interpolate smoothing of Test Data using the 14-gram English Language Model is: 13.6473164188
 Perplexity with Interpolate smoothing of Test Data using the 14-gram Spanish Language Model is: 25.0237933248
 Perplexity with Interpolate smoothing of Test Data using the 14-gram German Language Model is: 25.9048053933

 Perplexity with Frequency Interpolate of Test Data using the 14-gram English Language Model is: 7.64433823673
 Perplexity with Frequency Interpolate of Test Data using the 14-gram Spanish Language Model is: 29.8091254281
 Perplexity with Frequency Interpolate of Test Data using the 14-gram German Language Model is: 26.738163646

 Computing Perplexities of Test Data, using 15-Gram Model
 Perplexity without smoothing of Test Data using the 15-gram English Language Model is: 1.1314010165
 Perplexity without smoothing of Test Data using the 15-gram Spanish Language Model is: 30.6205130461
 Perplexity without smoothing of Test Data using the 15-gram German Language Model is: 8.26837120239

 Perplexity with Laplace smoothing of Test Data using the 15-gram English Language Model is: 36.3821335217
 Perplexity with Laplace smoothing of Test Data using the 15-gram Spanish Language Model is: 41.0033236048
 Perplexity with Laplace smoothing of Test Data using the 15-gram German Language Model is: 41.9955261173

 Perplexity with Backoff smoothing of Test Data using the 15-gram English Language Model is: 3.79830073059
 Perplexity with Backoff smoothing of Test Data using the 15-gram Spanish Language Model is: 16.5240931344
 Perplexity with Backoff smoothing of Test Data using the 15-gram German Language Model is: 15.357349379

 Perplexity with Interpolate of Test Data using the 15-gram English Language Model is: 5.8635744099
 Perplexity with Interpolate of Test Data using the 15-gram Spanish Language Model is: 17.5867070708
 Perplexity with Interpolate of Test Data using the 15-gram German Language Model is: 16.8670807339

 Perplexity with Interpolate smoothing of Test Data using the 15-gram English Language Model is: 14.1399604004
 Perplexity with Interpolate smoothing of Test Data using the 15-gram Spanish Language Model is: 25.551839222
 Perplexity with Interpolate smoothing of Test Data using the 15-gram German Language Model is: 26.4404579403

Perplexity with Frequency Interpolate	of Test Data using the 15-gram	English Language Model is: 7.64512461729
Perplexity with Frequency Interpolate	of Test Data using the 15-gram	Spanish Language Model is: 29.8091254281
Perplexity with Frequency Interpolate	of Test Data using the 15-gram	German Language Model is: 26.7381636441