INLP Assignment-1
Report

Jai Bhatnagar (2021111029)

February 2024

1 Perplexity Scores

LM1

Smoothing : Good-Turing
Corpus : Pride and Prejudice

Average Test set perplexity = 76278.39
Average Train set perplexity = 9063.86

LM2

Smoothing : Linear Interpolation
Corpus : Pride and Prejudice

Average Test set perplexity = 577.72
Average Train set perplexity = 11.40

LM3

Smoothing : Good-Turing
Corpus : Ulysses

Average Test set perplexity = 385963.38
Average Train set perplexity = 11033.55

LM4

Smoothing : Linear Interpolation
Corpus : Ulysses

Average Test set perplexity = 2273.43
Average Train set perplexity = 16.10

As we can see, in linear interpolation models the train set perplexity is very good and test set perplexity is higher because of unseen n-grams and words.

However here good-turing models perform worse than linear interpolation models since they assign a high probability to unseen n-grams. And since a high probability is assigned to unseen n-grams it may perform better on test set than on train set.

2 Generation (Given starting token '< s >')

"Pride and Prejudice" corpus

No Smoothing model

N = 2

Text Generated (upto 100 tokens):

< s > " And this question the last incumbrance of her that Netherfield would now removed to whom they need to the coffee and , has more substantial , your family , " Or in the present . < /s > < s > All that you had even be more for taking so much affected herself particularly prone to us has nothing was at the honour would be neither integrity nor condemn , if he had dined in all disapprove of her mother , in easy and the wisest and yet , could not let at its proprietor , " " " said her sister

N = 3

Text Generated (upto 100 tokens):

< s > I wonder when you first read that letter ? < /s > < s > Collins , and as a very handsome gentleman , and their motives has been at leisure to talk to the match , gave her a slight inclination of the scene , will we begin quarrelling about its relative situation . < /s > < s > Elizabeth was forced to come to the repeated appeals of her implied doubt , and Mr . < /s > < s > I am to have been strange if they are not to sketch my character , to be thought of poor Miss Bingley , and then stumbling on something

N = 4

Text Generated (upto 100 tokens):

< s > " " Why , at that rate , you will be honoured with the hands of all my fair cousins , the excellence of its cookery was owing . < /s > < s > Collins , being in fact much better fitted for a walker than a reader , was extremely well pleased , probably , to have conducted yourselves so as to command a full view of the house , amidst the various engagements which the kindness of his brother , and persuaded him to come . < /s > < s > " Elizabeth tried hard to dissuade him from such a quarter , than

N = 5

Text Generated (upto 100 tokens):

< s > " cried Elizabeth , brightening up for a moment . < /s > < s > " As she pronounced these words , Mr . < /s > < s > Elizabeth was determined ; nor did Sir William at all shake her purpose by his attempt at persuasion . < /s > < s > " Good Heaven ! < /s > < s > What will Wickham say ? < /s > < s > " " And _that_ is quite impossible ; for he is such a disagreeable man that it would be quite a misfortune to be liked by him . < /s > < s > She looked forward to their entrance , as the point on which all

LM2

N = 2

Text Generated (upto 100 tokens):

< s > Lydia and keep she never to encounter Charlotte ? < /s > it - , and soon passed into it is of but absence _had_ been so ; without any answer what we He must < /s > < s > But there was plain , or other about tour whom a and Mrs Miss _We_ were calling . . < /s > < s > < s > him of but Elizabeth remained at the road with pleasure ; and his manners towards whom he was still expected no such engaged in locations where they were to me , " " " True , and its way towards the living in

N = 3

Text Generated (upto 100 tokens):

< s > What a contrast between him at all likely said . < /s > < s > . wilfully deceiving any one . < /s > < s > He all the greatest part us < /s > for as to the nature on your kindness was not < /s > < s > an in his kindness was neither . family , with him . < /s > < s > It does not appear < /s > < s > Mrs it will you expect her to ruin him for it read with somewhat clearer the sake the reading come back and sister , " " From the < /s > < s > pray , what " Jane should therefore and < s > would have

N = 4

Text Generated (upto 100 tokens):

< s > Could I expect a William ' ; of her little liked nice little pair of you are safe arrival of choice - some part mother : " MY . father , " for employment , she But am upon to Hunsford . let of a very Gutenberg forget his ordinary style particular considering what she should than allusion for some time would keep There was to of tenderness pride _that_ in the eye of a what it meant than pleasant that agreeable < /s > < /s > < s > I know nothing of that you should have gone so early day she feared to

N = 5

Text Generated (upto 100 tokens):

< s > Jenkinson < s > III at < s > light manner experienced some change of questions in the other thinking of their from She accounted to her daughter sat fourthly it all my family , though < /s > be his own doing , a perfect < s > , and to lament being Another intreaty , she should have known life to some minutes he sick ; and a our friends first , and on which not give in thought very ; necessary Lydia leads . Collins ' to deny read and directed all his anger , took the marriage between , in general behaviour during the

"Ulysses" corpus

No Smoothing model

N = 2

Text Generated (upto 100 tokens):

< s > says John Eglinton exclaimed . < /s > < s > Is that . < /s > < s > The leg , Larry O , tapping , says J . < /s > < s >) at , smiling at Khartoum lighting their mending their buttonholes , the presupposed intangibility of bright with which his neck , Medina Villa , in the bracing ozone and father ' Club , every day , yes . < /s > < s > _ THE RAW — Jacky Caffrey said with eyes . < /s > < s > — O . < /s > < s > Healthy too after five tall hats , look long John Conmee S . < /s > < s > Faithful departed

N = 3

Text Generated (upto 100 tokens):

< s > Among gumheavy serpentplants , milkoozing fruits , young one . < /s > < s > Ten to . < /s > < s > He shaved evenly and with him just imagine having to beg from these swine . < /s > < s > — He had sometimes propelled her on the Trinity jibs in their pockets . < /s > < s > — Are you not write your poetry in it all in this snuffbox ? < /s > < s > Mr Power asked . < /s > < s > Mohammed cut a long mile before you . < /s > < s > Did , faith . < /s > < s > — About the boatman a florin for saving his son Leopold ,

N = 4

Text Generated (upto 100 tokens):

< s > O , get , rev on a gradient one in nine . < /s > < s > _ (From the suttee pyre the flame of gum camphire ascends . < /s > < s > One of them lost two quid on my tip _Sceptre_ for himself and a lady friend . < /s > < s > As the glossy horses pranced by Merrion square Master Patrick Aloysius Dignam , waiting , while the man in the gap turning up at the drill instructing to find out whether he likes me I looked a bit washy of course when I looked close in the vicinity of a place of

N = 5

Text Generated (upto 100 tokens):

< s > A friend of my father ' s , Gumley . < /s > < s > Love me not . < /s > < s > _The Arabian Nights Entertainment_ was my favourite and _Red as a Rose is She . < /s > < s > Half dream . < /s > < s > Ah ! < /s > < s > says he . < /s > < s > Smile . < /s > < s > No , she couldn ' t see and to mind he didn ' t clap him in the dock the other day for suing poor little Gumley that ' s minding stones , for the corporation there near Butt bridge . < /s > < s > — The bard '

LM4

N = 2

Text Generated (upto 100 tokens):

< s > thick rich proportion increasing from those _Express_ with you . thinking . < /s > < s >) , smiling great there faculties . < /s > I . < /s > < s > I thing . < /s > < s > I ' s Yvonne knockingshop it sourly understand him , raising will never persecuted the maturation of his crustcrumbs < s > _ (_sic_ s , tearful oysters but regular hours and taking up to the like quarrelling < s > — Will , rodent , halted ice , . I can ' then . , by pay it shone , ' s veiling , to encourage his ? < /s > It was

N = 3

Text Generated (upto 100 tokens):

< s > _I owe you ? Blackburn does , when he becomes water is but was they had their all traffic was and a cryptogram (vowels suppressed) gives is why they came knocking at on in were bit on that staid agent of intimacy) _ (Masculinely . < /s > < s >) in to the city thing like that for ? < /s > < s > Believe kettle She . . < /s > < s > example when I by his winners among half confess < /s > — said well . < /s > < s > He slapped a warning to him The bloody remorse again reassuralooms < s > purchaser < s >

N = 4

Text Generated (upto 100 tokens):

< s > undulating plain scotch . . . < /s > . < /s > . The plump is said , . < /s > Penny his slender . , < s > Hurry . < /s > < s > She book with as < s > J ' t all that ' s blank face Burton restaurant for Bloom youre gay and friendly over it O Certainly , Mr Philip Beaufoy , more , < s > accord turned < s > case go off . own . < /s > < s > and < s > _What < /s > _ (In quakergrey , whatever the season are , and in seaquakes , waterspouts ,) teach you Who ' vagrant _The Messiah_

N = 5

Text Generated (upto 100 tokens):

< s > s day : a perfume does to ll patrum_ , olive face is no , Mr Bloom said , and you do not but he never or < /s > < s > without a Here lies from their mouths < /s > < s > What chemise to right eye ? < /s >

the round to inherit by ll put He $< /s > < s >$ No cardsharping then nought would keep hauls up , and s the foot a $< /s >$ s . to shadows black marble timepiece . $< /s > < s >$ The Abbey quite till I am , then hell ! $< /s > < /s > < s >$ it nature and it jaunty . $< /s >$

3 Theory

Good-Turing Smoothing

We assume a frequency distribution according to Zipf's law i.e. most words will have low frequency, least words will have most frequency, and the plot between frequencies (r) and number of words having those frequencies (N_r) will be almost linear in log-scale.

First we change the N_r values to preserve this trend even at higher values of r or lower values of N_r.

$$N_r = \frac{2N_r}{(t-q)}$$

for all non-zero N_r, where t is the next higher frequency at which N_r is non-zero and q is the previous lower frequency at which N_r is non-zero.

Then we train a linear regressor to smooth out the N_r values

$$log(N_r) = a + b\, log(r)$$

Finally we estimate the adjusted frequencies as,

$$r^* = (r+1)\frac{S(N_{r+1})}{S(N_r)}$$

where r^* is the new adjusted frequency, $S(N_r)$ is the predicted value of N_r from the linear regression model. ($S(N_0) = 1$)

After this, these adjusted frequencies are used everywhere for probability calculations.

Linear Interpolation

In linear interpolation we use the probabilities of n-grams, (n-1)-grams ... unigrams. Each probability is assigned a weight which are learned using the following algorithm (for trigrams).

---

set $\lambda_1 = \lambda_2 = \lambda_3 = 0$
foreach trigram t_1, t_2, t_3 **with** $f(t_1, t_2, t_3) > 0$
  depending on the maximum of the following three values:
    case $\frac{f(t_1,t_2,t_3)-1}{f(t_1,t_2)-1}$: **increment** λ_3 **by** $f(t_1, t_2, t_3)$
    case $\frac{f(t_2,t_3)-1}{f(t_2)-1}$: **increment** λ_2 **by** $f(t_1, t_2, t_3)$
    case $\frac{f(t_3)-1}{N-1}$: **increment** λ_1 **by** $f(t_1, t_2, t_3)$
  end
end
normalize $\lambda_1, \lambda_2, \lambda_3$

Then probability is calculated as (for trigrams),

$$P_L(w3|w1w2) = \lambda_1 P(w3|w1w2) + \lambda_2 P(w3|w2) + \lambda_3 P(w3)$$