Intro to NLP

Assignment 1

Name: Bhanuj Gandhi

Roll No.: 2022201068

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY

HYDERABAD

Q1. Tokenisation

You have been given two corpuses for cleaning. Your task is to design a tokenizer using regex, which you will later use for smoothing and language modelling as well.

1. Create a Tokenizer to handle following cases:

. (a) Word Tokenizer

. (b) Punctuation

. (c) URLs

. (d) Hashtags (#manchesterisred)

. (e) Mentions (@john)

2. For the following cases, replace the tokens with appropriate placeholders:

. (a) URLs: <URL>

. (b) Hashtags: <HASHTAG>

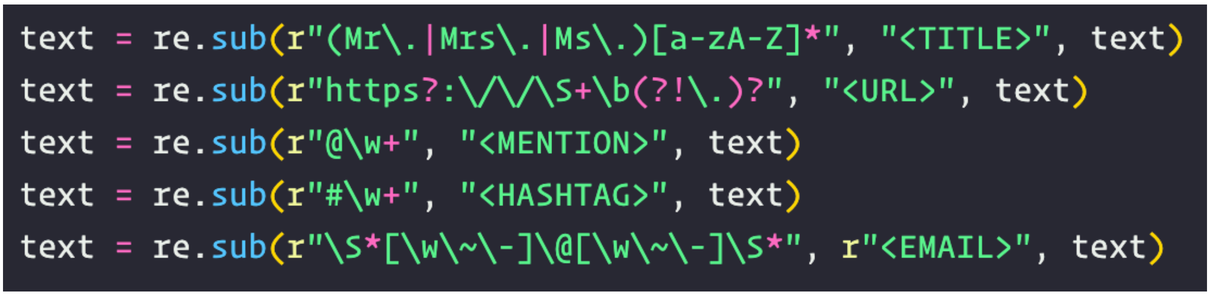
. (c) Mentions: <MENTION>

You are also encouraged to try other tokenisation and placeholder substitution schemes based on your observations from the corpora used for the smoothing task to achieve a better language model. You may find percentages, age values, expressions indicating time, time periods occurring in the data. You’re free to explore and add multiple such reasonable tokenization schemes in addition from the ones listed above. Specify any such schemes you use in the final README.

—

Observations:

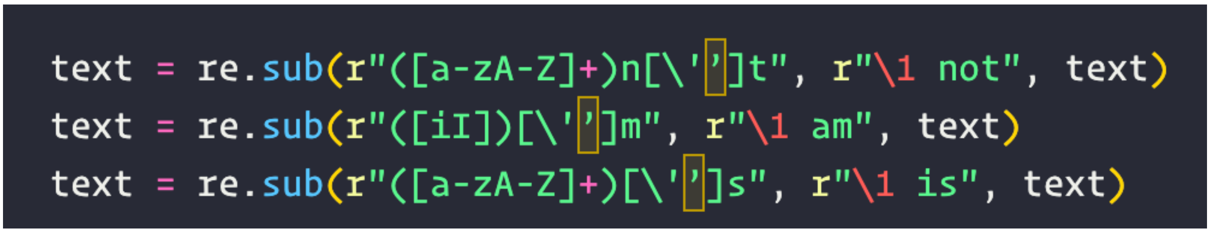
We have been given with two corpora, both of them are books. Both of them have various junk text such as front page of the book, index page, acknowledgement page, footnotes, etc. thus both of them needs cleaning of the text. As we need to use this tokenizer in n-gram language model, we need to capture limited context, so we can eliminate various tokens such as emails, URLs, mentions, etc. By carefully looking at the whole corpus, I have narrowed down to various tokens that will not help in training the n-gram language model and replaced them. This will help in capturing better context.



```
text = re.sub(r"(Mr\.|Mrs\.|Ms\.)[a-zA-Z]*", "<TITLE>", text)
text = re.sub(r"https?:\/\/\S+\b(?!\.)?", "<URL>", text)
text = re.sub(r"@\w+", "<MENTION>", text)
text = re.sub(r"#\w+", "<HASHTAG>", text)
text = re.sub(r"\S*[\w\~\-]\@[\w\~\-]\S*", r"<EMAIL>", text)
```

The above snippet handles following things

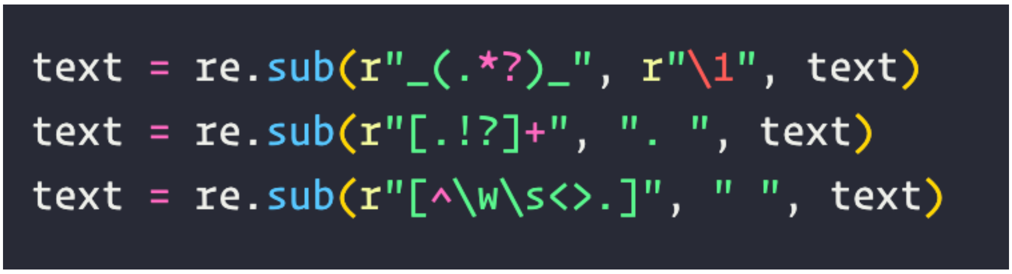
- Title
- URLs
- User Mentions
- Hashtags
- Emails



```
text = re.sub(r"([a-zA-Z]+)n[\'’]t", r"\1 not", text)
text = re.sub(r"([iI])[\'’]m", r"\1 am", text)
text = re.sub(r"([a-zA-Z]+)[\'’]s", r"\1 is", text)
```

In the above snippet, various cases where ‘ can occur are handles

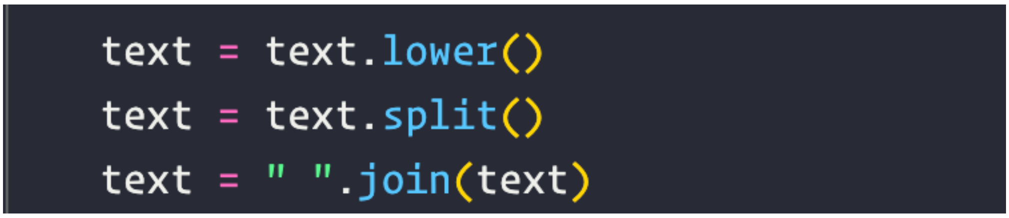
- Words like Couldn’t are transformed to Could not
- Words like I’m are transformed to I am
- Words like John’s are transformed to John is



```python
text = re.sub(r"_(.*?)_", r"\1", text)
text = re.sub(r"[.!?]+", ". ", text)
text = re.sub(r"[^\w\s<>.]", " ", text)
```

The above snippet handles cases like:

- Markdown
- End Sentence Punctuation
- Removes Punctuation

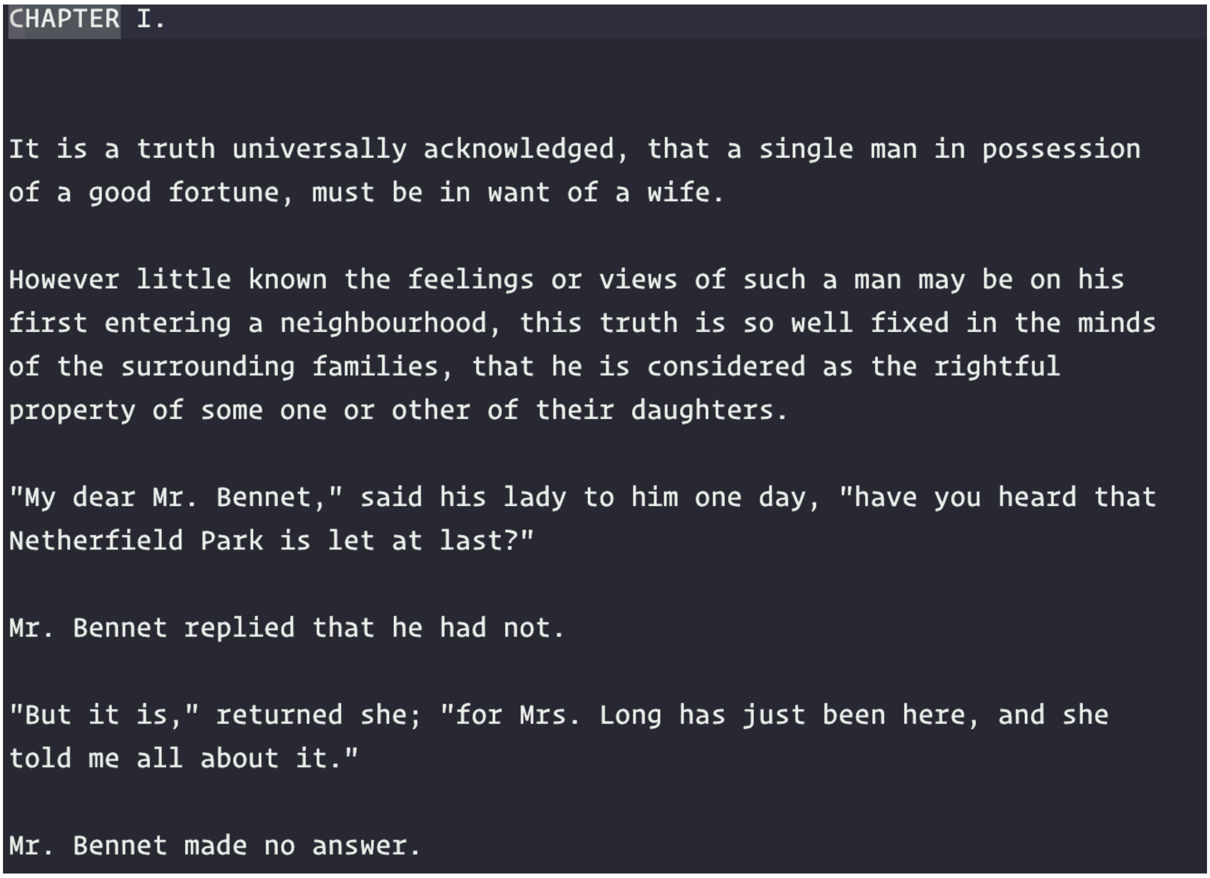


```python
    text = text.lower()
    text = text.split()
    text = " ".join(text)
```

The above snippet handles the extra spaces that are added while removing the above mentioned tokens.

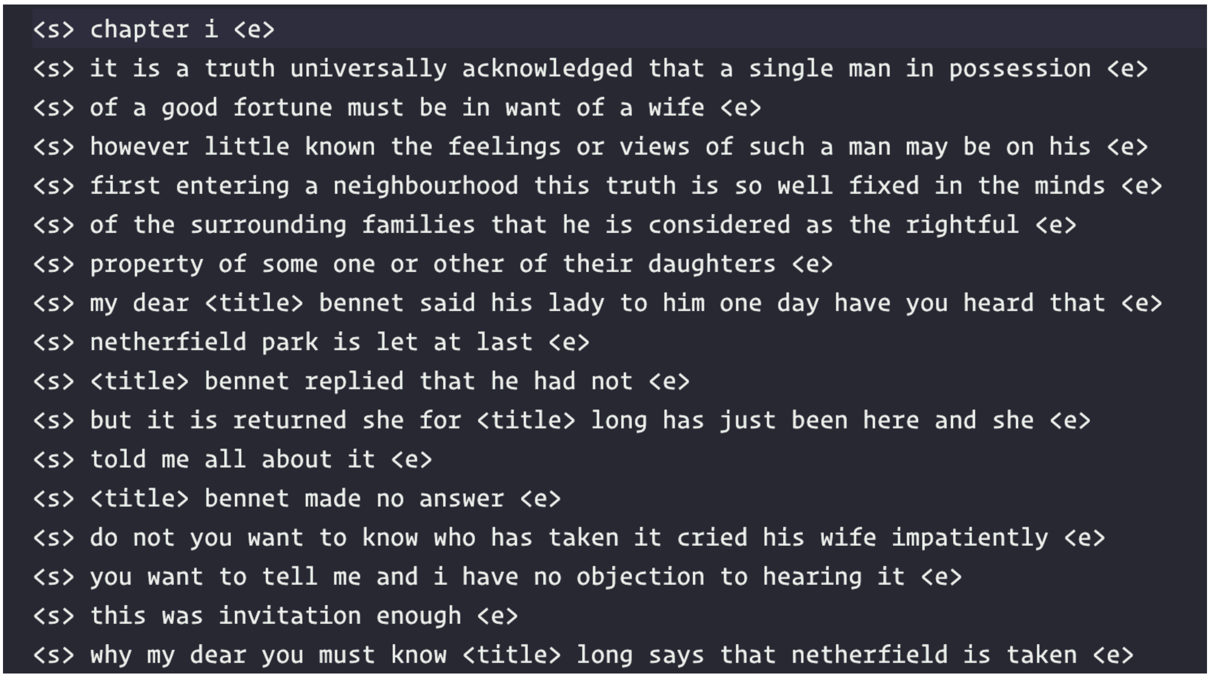
Results:

Original Text



```
CHAPTER I.


It is a truth universally acknowledged, that a single man in possession
of a good fortune, must be in want of a wife.

However little known the feelings or views of such a man may be on his
first entering a neighbourhood, this truth is so well fixed in the minds
of the surrounding families, that he is considered as the rightful
property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that
Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she
told me all about it."

Mr. Bennet made no answer.
```

Clean Text



```
<s> chapter i <e>
<s> it is a truth universally acknowledged that a single man in possession <e>
<s> of a good fortune must be in want of a wife <e>
<s> however little known the feelings or views of such a man may be on his <e>
<s> first entering a neighbourhood this truth is so well fixed in the minds <e>
<s> of the surrounding families that he is considered as the rightful <e>
<s> property of some one or other of their daughters <e>
<s> my dear <title> bennet said his lady to him one day have you heard that <e>
<s> netherfield park is let at last <e>
<s> <title> bennet replied that he had not <e>
<s> but it is returned she for <title> long has just been here and she <e>
<s> told me all about it <e>
<s> <title> bennet made no answer <e>
<s> do not you want to know who has taken it cried his wife impatiently <e>
<s> you want to tell me and i have no objection to hearing it <e>
<s> this was invitation enough <e>
<s> why my dear you must know <title> long says that netherfield is taken <e>
```

Q2. Smoothing

You have been given two corpus: “Pride and Prejudice” corpus, and “Ulysses” corpus. Your task is to design Language Models for both corpora using smoothing. Ensure that you use the tokenizer created in task 1 for this task.

1. Create language models with the following parameters:

   1. (a) On “Pride and Prejudice” corpus:
      - i. LM 1: tokenization + 4-gram LM + Kneser-Ney smoothing.
      - ii. LM 2: tokenization + 4-gram LM + Witten-Bell smoothing.

      (b) On “Ulysses” corpus:

      - iii. LM 3: tokenization + 4-gram LM + Kneser-Ney smoothing.
      - iv. LM 4: tokenization + 4-gram LM + Witten-Bell smoothing.
   2. For each of these corpora, create a test set by randomly selecting 1000 sentences. This set will not be used for training the LM.

   a) Calculate perplexity score for each sentence of “Pride and Prejudice” corpus and “Ulysses” corpus for each of the above models and also get average perplexity score on the train corpus.

   b) Report the perplexity scores for all the sentences in the training set. Report the perplexity score on the test sentences as well, in the same manner above.

Observations

Kneser-Ney smoothing [KN95] is a modified version of absolute discounting. The idea is to optimize the calculation of the lower-order n-gram probabilities in case the higher- order n-gram was unseen in the corpus. It is thereby originally a backoff smoothing algorithm

[KN95]. The high-level motivation is that, using the backoff version of abso- lute discounting, the information that the higher-order n-gram was unseen is not taken into account when backing-off and calculating the lower-order probability [KN95].

Steps

1. Clean the text using above created tokeniser
2. Create *backward dictionary, forward dictionary, and middle dictionary* in order to find various continuation count in the formula
3. Calculate recursively

Kneser-Ney:

Base Case:

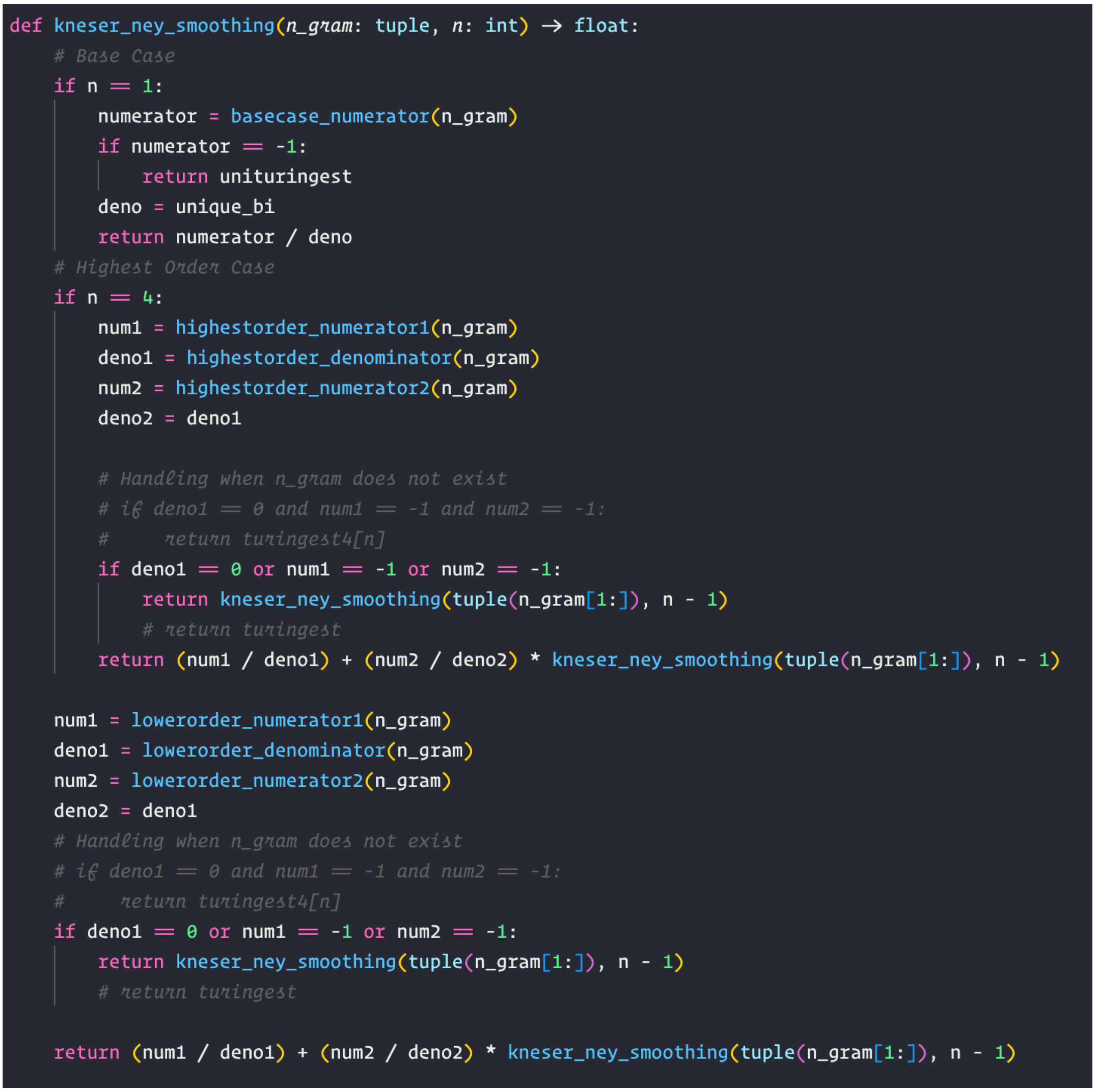
$$P_{\mathrm{KN}}(w_i) = \frac{N_{1+}(\bullet w_i)}{N_{1+}(\bullet\bullet)}$$

Lower n-grams

$$\begin{aligned} P_{\mathrm{KN}}(w_i|w_{i-n+1}^{i-1}) = & \frac{\max\{N_{1+}(\bullet w_{i-n+1}^{i}) - D, 0\}}{N_{1+}(\bullet w_{i-n+1}^{i-1}\bullet)} \\ & + \frac{D}{N_{1+}(\bullet w_{i-n+1}^{i-1}\bullet)} N_{1+}(w_{i-n+1}^{i-1}\bullet) P_{\mathrm{KN}}(w_i|w_{i-n+2}^{i-1}) \end{aligned}$$

Higher n-grams

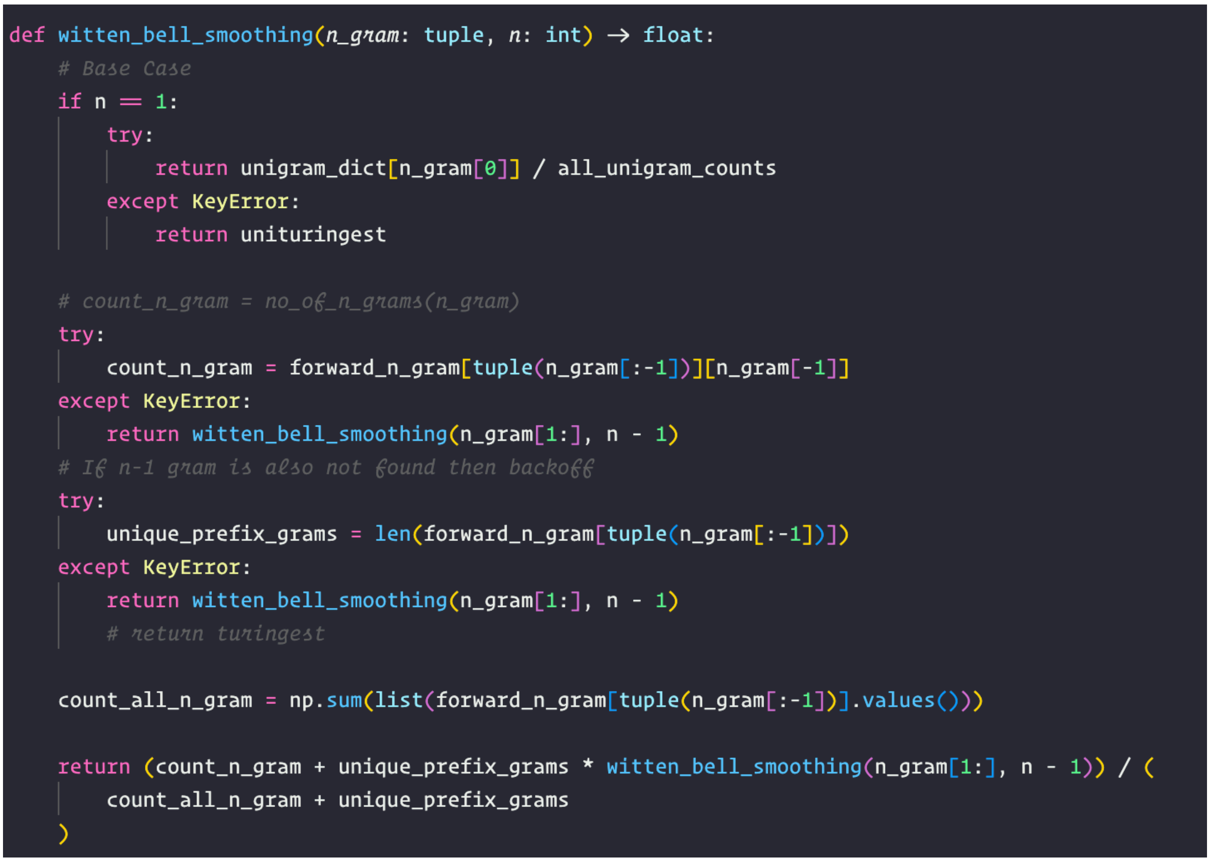
$$
\begin{aligned}
P_{\mathrm{KN}}(w_i|w_{i-n+1}^{i-1}) = & \frac{\max\{N_{1+}(\bullet w_{i-n+1}^{i}) - D, 0\}}{N_{1+}(\bullet w_{i-n+1}^{i-1}\bullet)} \\
& + \frac{D}{N_{1+}(\bullet w_{i-n+1}^{i-1}\bullet)} N_{1+}(w_{i-n+1}^{i-1}\bullet) P_{\mathrm{KN}}(w_i|w_{i-n+2}^{i-1})
\end{aligned}
$$



```python
def kneser_ney_smoothing(n_gram: tuple, n: int) -> float:
    # Base Case
    if n == 1:
        numerator = basecase_numerator(n_gram)
        if numerator == -1:
            return unituringest
        deno = unique_bi
        return numerator / deno
    # Highest Order Case
    if n == 4:
        num1 = highestorder_numerator1(n_gram)
        deno1 = highestorder_denominator(n_gram)
        num2 = highestorder_numerator2(n_gram)
        deno2 = deno1

        # Handling when n_gram does not exist
        # if deno1 == 0 and num1 == -1 and num2 == -1:
        #     return turingest4[n]
        if deno1 == 0 or num1 == -1 or num2 == -1:
            return kneser_ney_smoothing(tuple(n_gram[1:]), n - 1)
            # return turingest
        return (num1 / deno1) + (num2 / deno2) * kneser_ney_smoothing(tuple(n_gram[1:]), n - 1)

    num1 = lowerorder_numerator1(n_gram)
    deno1 = lowerorder_denominator(n_gram)
    num2 = lowerorder_numerator2(n_gram)
    deno2 = deno1
    # Handling when n_gram does not exist
    # if deno1 == 0 and num1 == -1 and num2 == -1:
    #     return turingest4[n]
    if deno1 == 0 or num1 == -1 or num2 == -1:
        return kneser_ney_smoothing(tuple(n_gram[1:]), n - 1)
        # return turingest

    return (num1 / deno1) + (num2 / deno2) * kneser_ney_smoothing(tuple(n_gram[1:]), n - 1)
```

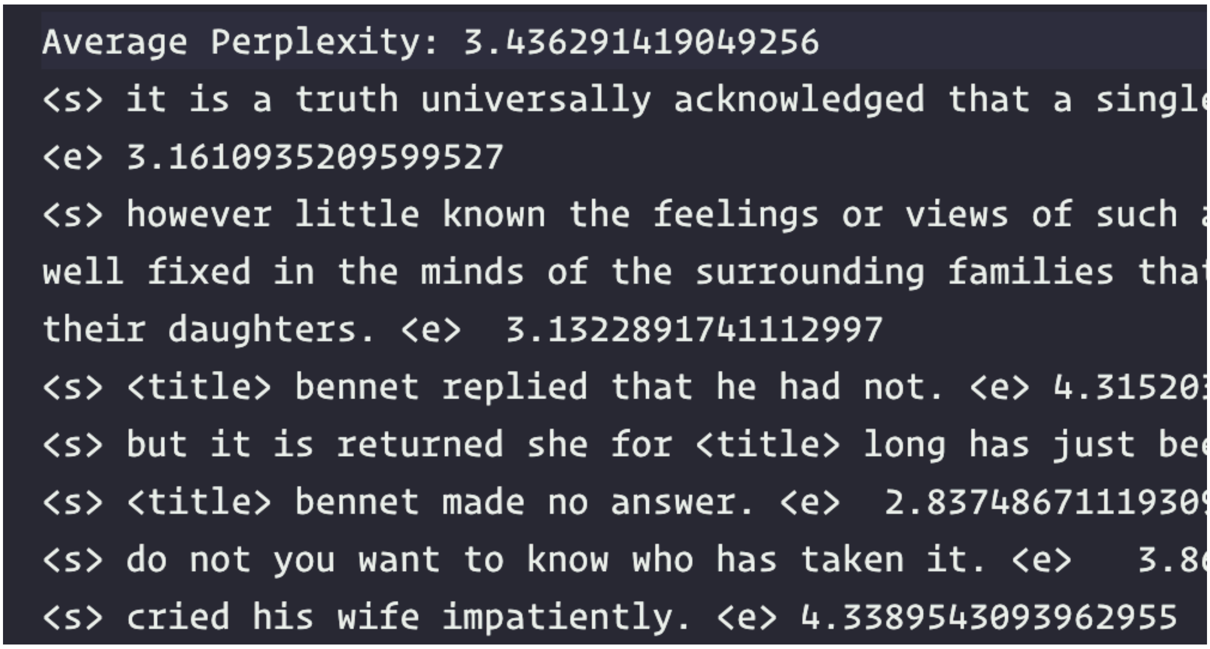
Witten Bell:

$$p_{\mathrm{WB}}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) + N_{1+}(w_{i-n+1}^{i-1}\bullet)p_{\mathrm{WB}}(w_i|w_{i-n+2}^{i-1})}{\sum_{w_i} c(w_{i-n+1}^{i}) + N_{1+}(w_{i-n+1}^{i-1}\bullet)}$$



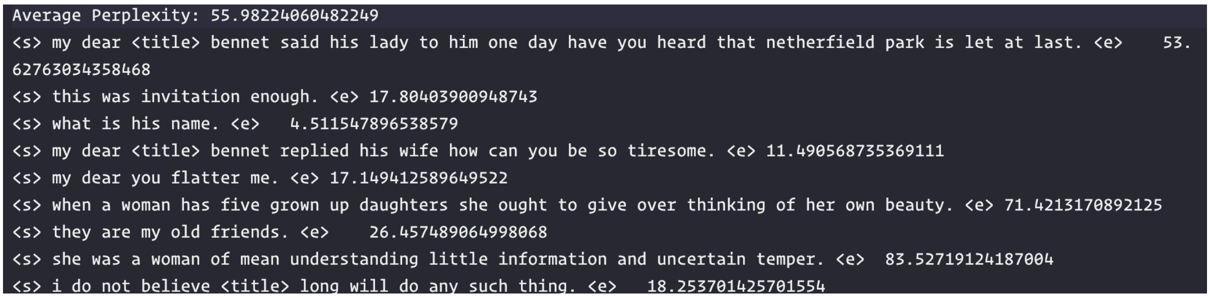
```
def witten_bell_smoothing(n_gram: tuple, n: int) -> float:
    # Base Case
    if n == 1:
        try:
            return unigram_dict[n_gram[0]] / all_unigram_counts
        except KeyError:
            return unituringest

    # count_n_gram = no_of_n_grams(n_gram)
    try:
        count_n_gram = forward_n_gram[tuple(n_gram[:-1])][n_gram[-1]]
    except KeyError:
        return witten_bell_smoothing(n_gram[1:], n - 1)
    # If n-1 gram is also not found then backoff
    try:
        unique_prefix_grams = len(forward_n_gram[tuple(n_gram[:-1])])
    except KeyError:
        return witten_bell_smoothing(n_gram[1:], n - 1)
        # return turingest

    count_all_n_gram = np.sum(list(forward_n_gram[tuple(n_gram[:-1])].values()))

    return (count_n_gram + unique_prefix_grams * witten_bell_smoothing(n_gram[1:], n - 1)) / (
        count_all_n_gram + unique_prefix_grams
    )
```

1. LM1 (Train): **Average Perplexity: 3.43**



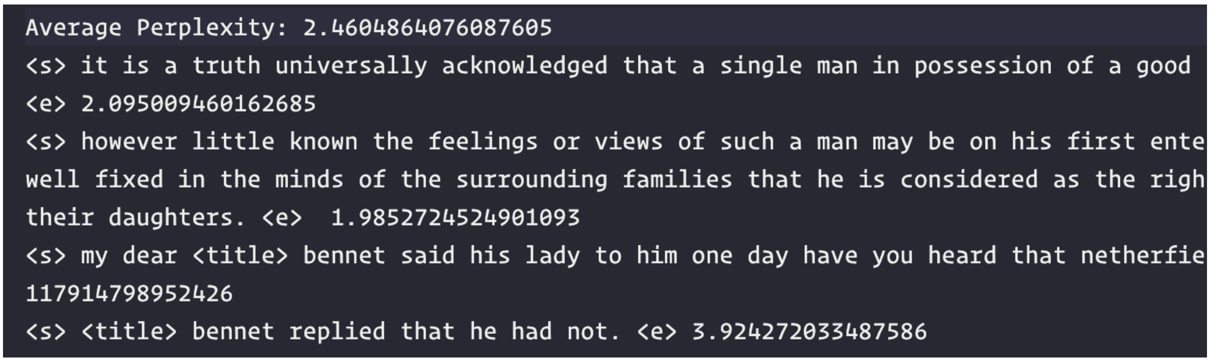
```
Average Perplexity: 3.436291419049256
<s> it is a truth universally acknowledged that a singl
<e> 3.1610935209599527
<s> however little known the feelings or views of such
well fixed in the minds of the surrounding families tha
their daughters. <e>  3.1322891741112997
<s> <title> bennet replied that he had not. <e> 4.31520
<s> but it is returned she for <title> long has just be
<s> <title> bennet made no answer. <e>  2.8374867111930
<s> do not you want to know who has taken it. <e>   3.8
<s> cried his wife impatiently. <e> 4.3389543093962955
```

2. LM1 (Test): **Average Perplexity: 55.98**



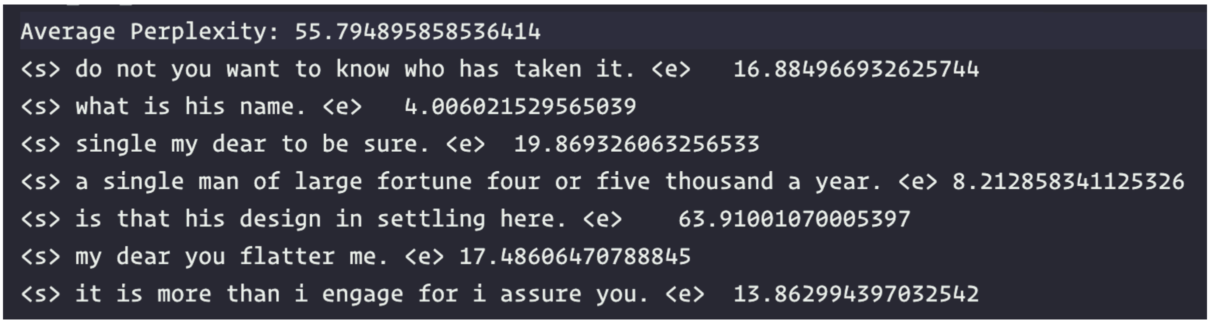
```
Average Perplexity: 55.98224060482249
<s> my dear <title> bennet said his lady to him one day have you heard that netherfield park is let at last. <e>    53.
62763034358468
<s> this was invitation enough. <e> 17.80403900948743
<s> what is his name. <e>   4.511547896538579
<s> my dear <title> bennet replied his wife how can you be so tiresome. <e> 11.490568735369111
<s> my dear you flatter me. <e> 17.149412589649522
<s> when a woman has five grown up daughters she ought to give over thinking of her own beauty. <e> 71.4213170892125
<s> they are my old friends. <e>    26.457489064998068
<s> she was a woman of mean understanding little information and uncertain temper. <e>  83.52719124187004
<s> i do not believe <title> long will do any such thing. <e>   18.253701425701554
```

3. LM2 (Train): **Average Perplexity: 2.46**



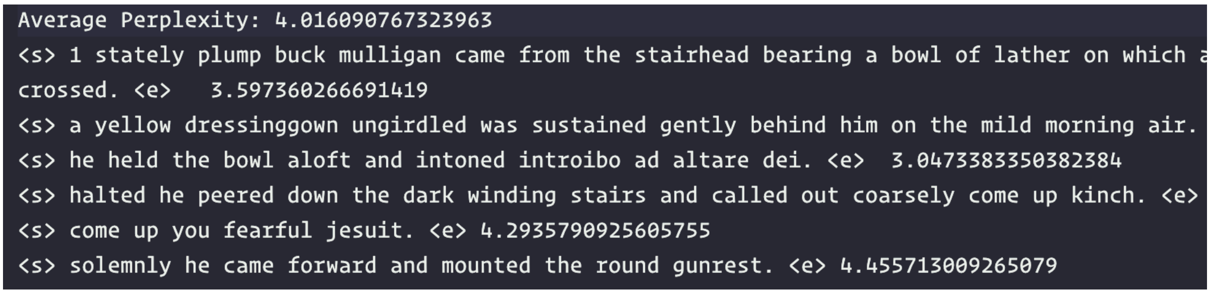
```
Average Perplexity: 2.4604864076087605
<s> it is a truth universally acknowledged that a single man in possession of a good
<e> 2.095009460162685
<s> however little known the feelings or views of such a man may be on his first ente
well fixed in the minds of the surrounding families that he is considered as the righ
their daughters. <e>  1.9852724524901093
<s> my dear <title> bennet said his lady to him one day have you heard that netherfie
117914798952426
<s> <title> bennet replied that he had not. <e> 3.924272033487586
```

4. LM2 (Test): **Average Perplexity: 55.79**



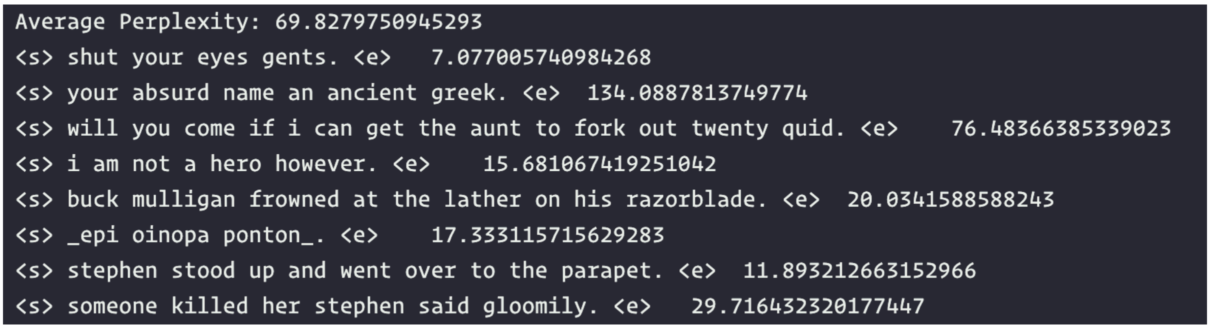
```
Average Perplexity: 55.794895858536414
<s> do not you want to know who has taken it. <e>   16.884966932625744
<s> what is his name. <e>   4.006021529565039
<s> single my dear to be sure. <e>  19.869326063256533
<s> a single man of large fortune four or five thousand a year. <e> 8.212858341125326
<s> is that his design in settling here. <e>    63.91001070005397
<s> my dear you flatter me. <e> 17.48606470788845
<s> it is more than i engage for i assure you. <e>  13.862994397032542
```

5. LM3 (Train): **Average Perplexity: 4.01**



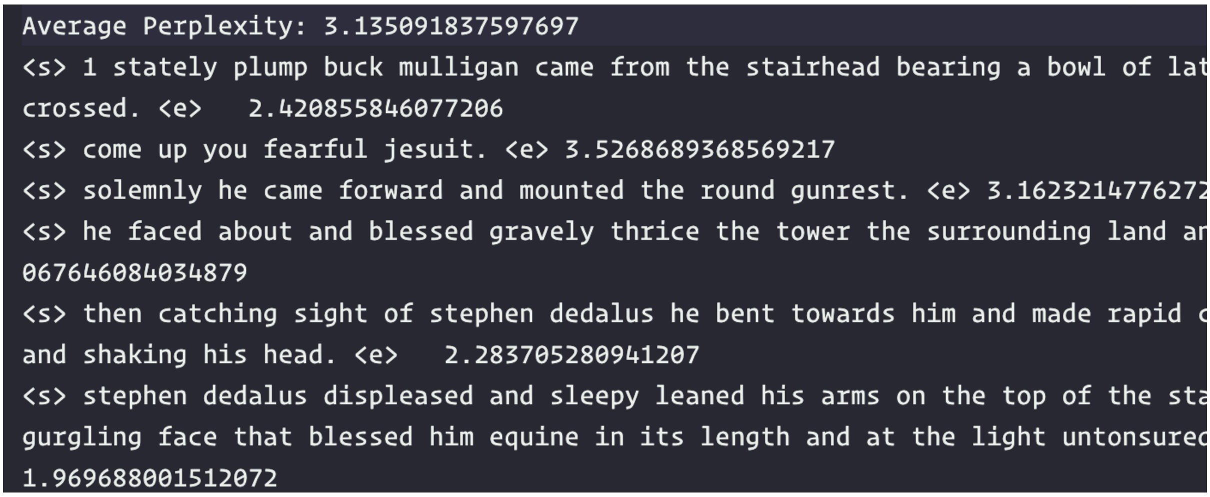
```
Average Perplexity: 4.016090767323963
<s> 1 stately plump buck mulligan came from the stairhead bearing a bowl of lather on which a
crossed. <e>   3.597360266691419
<s> a yellow dressinggown ungirdled was sustained gently behind him on the mild morning air.
<s> he held the bowl aloft and intoned introibo ad altare dei. <e>  3.0473383350382384
<s> halted he peered down the dark winding stairs and called out coarsely come up kinch. <e>
<s> come up you fearful jesuit. <e> 4.2935790925605755
<s> solemnly he came forward and mounted the round gunrest. <e> 4.455713009265079
```

6. LM3 (Test): **Average Perplexity: 69.82**



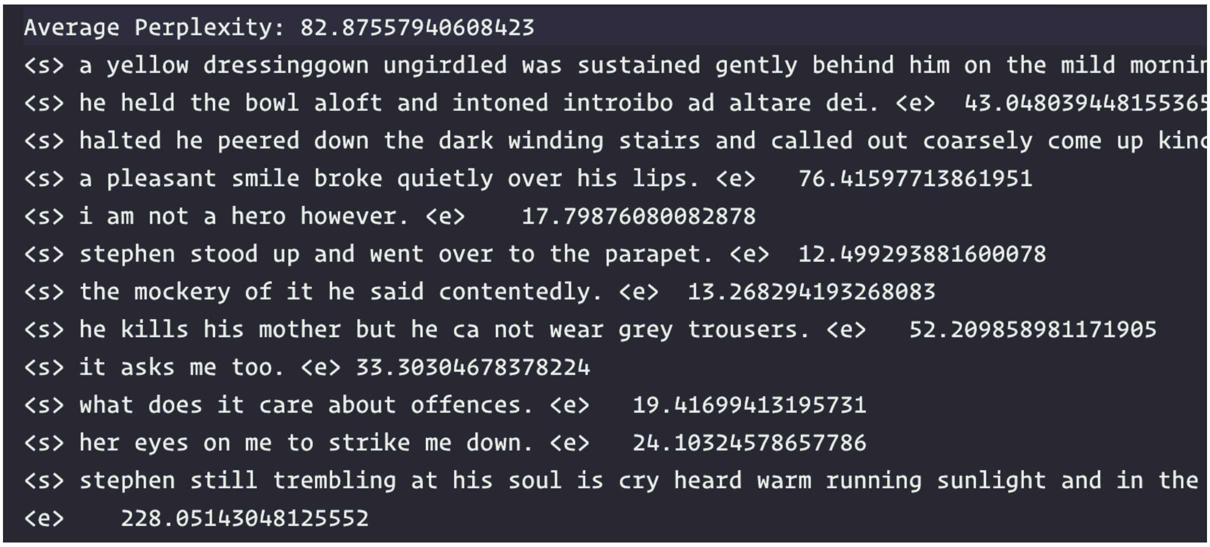
```
Average Perplexity: 69.8279750945293
<s> shut your eyes gents. <e>   7.077005740984268
<s> your absurd name an ancient greek. <e>  134.0887813749774
<s> will you come if i can get the aunt to fork out twenty quid. <e>    76.48366385339023
<s> i am not a hero however. <e>    15.681067419251042
<s> buck mulligan frowned at the lather on his razorblade. <e>  20.0341588588243
<s> _epi oinopa ponton_. <e>    17.333115715629283
<s> stephen stood up and went over to the parapet. <e>  11.893212663152966
<s> someone killed her stephen said gloomily. <e>   29.716432320177447
```

7. LM4 (Train): **Average Perplexity: 3.13**



```
Average Perplexity: 3.135091837597697
<s> 1 stately plump buck mulligan came from the stairhead bearing a bowl of lat
crossed. <e>   2.420855846077206
<s> come up you fearful jesuit. <e> 3.5268689368569217
<s> solemnly he came forward and mounted the round gunrest. <e> 3.16232147762722
<s> he faced about and blessed gravely thrice the tower the surrounding land ar
067646084034879
<s> then catching sight of stephen dedalus he bent towards him and made rapid c
and shaking his head. <e>   2.283705280941207
<s> stephen dedalus displeased and sleepy leaned his arms on the top of the sta
gurgling face that blessed him equine in its length and at the light untonsured
1.969688001512072
```

8. LM4 (Test): **Average Perplexity: 82.87**



```
Average Perplexity: 82.87557940608423
<s> a yellow dressinggown ungirdled was sustained gently behind him on the mild mornin
<s> he held the bowl aloft and intoned introibo ad altare dei. <e>  43.048039448155365
<s> halted he peered down the dark winding stairs and called out coarsely come up kin
<s> a pleasant smile broke quietly over his lips. <e>   76.41597713861951
<s> i am not a hero however. <e>    17.79876080082878
<s> stephen stood up and went over to the parapet. <e>  12.499293881600078
<s> the mockery of it he said contentedly. <e>  13.268294193268083
<s> he kills his mother but he ca not wear grey trousers. <e>   52.209858981171905
<s> it asks me too. <e> 33.30304678378224
<s> what does it care about offences. <e>   19.41699413195731
<s> her eyes on me to strike me down. <e>   24.10324578657786
<s> stephen still trembling at his soul is cry heard warm running sunlight and in the
<e>    228.05143048125552
```

Conclusion

As we have used modified Kneser Ney Smoothing, we can see that it outperforms Witten Bell Smoothing in many forms. Although it largely depends on the test-train partition and the quality of sentences in each split. As we have used random split, each time perplexity is different but it is observed that Modified Kneser Ney outperforms Witten Bell.

Q3. Neural Language Model

You have been given two datasets. Make your own data splits (for eg: 70%, 15%, 15%). Train separates models on each corpus.

1. Create a neural language model using a recurrent architecture such as LSTMs for both the corpora provided.

   (a) Use the train split for training the model parameters.

   (b) Dev split for hyperparameter optimization as well saving the best model checkpoint.

   (c) Test split for a final check on the performance.

2. Perplexity computation:

   (a) Calculate the perplexity score for each sentence of the train splits for both the corpora only using the trained model and get the averaged perplexity score on it.

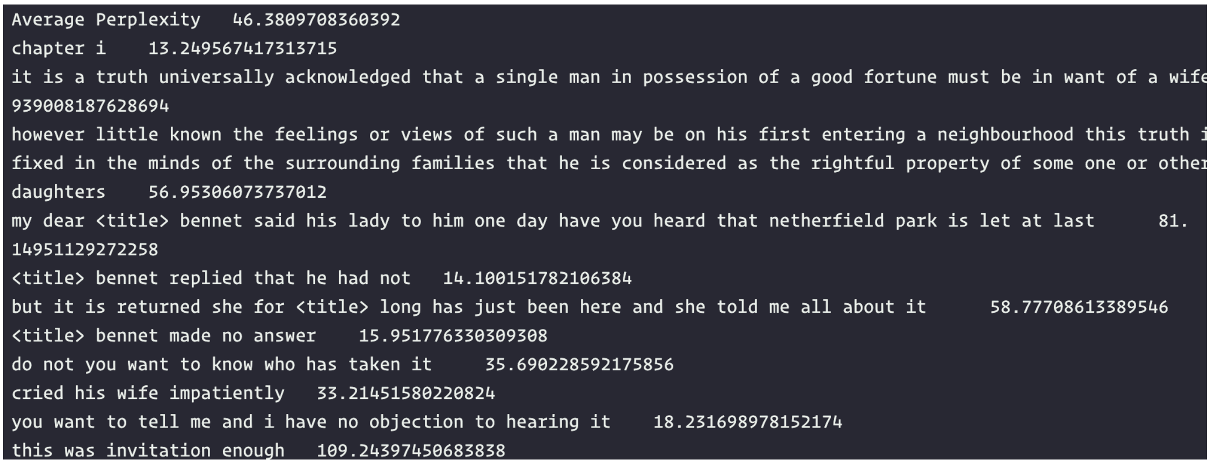
   (b) Report the perplexity score on the test sentences as well, in the same manner as above {LM 5 (Pride and Prejudice), LM 6 (Ulysses)}.

   (c) Compare and analyse the behaviour of the different LMs and put your analysis and visualisation in report.

Observations

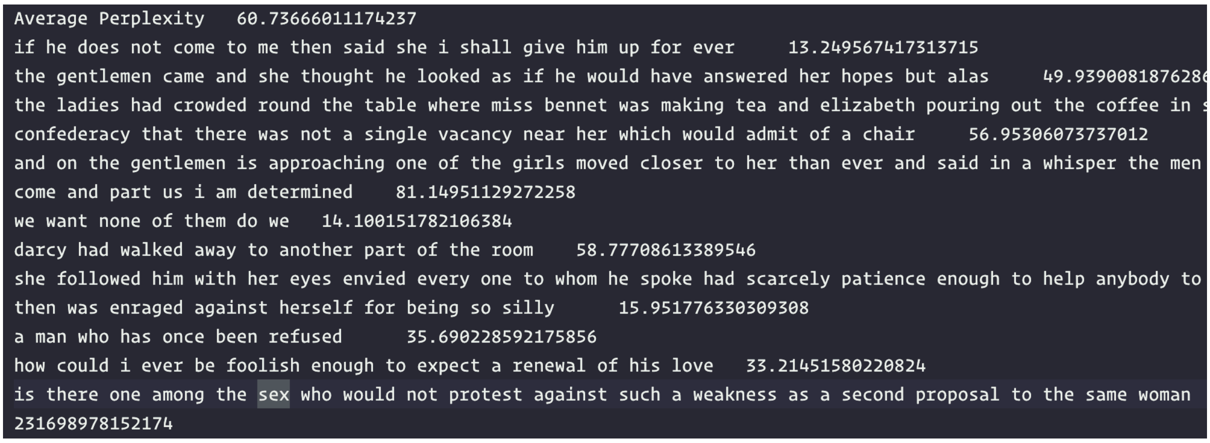
Since the copora given are books, they must have large context in them. Due to n-gram capturing only limited context, we have used LSTM neural language model. Larger context is feeded and based on that weights are set.

1. LM 7 (Train): **Average Perplexity: 46.38**



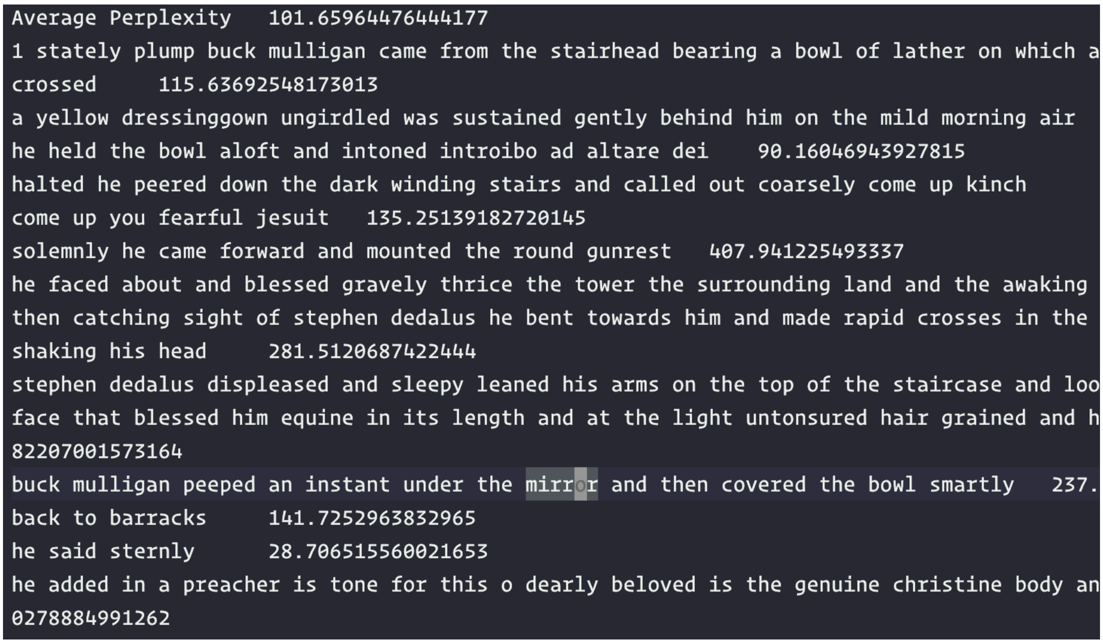
```
Average Perplexity   46.3809708360392
chapter i    13.249567417313715
it is a truth universally acknowledged that a single man in possession of a good fortune must be in want of a wife
939008187628694
however little known the feelings or views of such a man may be on his first entering a neighbourhood this truth i
fixed in the minds of the surrounding families that he is considered as the rightful property of some one or other
daughters    56.95306073737012
my dear <title> bennet said his lady to him one day have you heard that netherfield park is let at last      81.
14951129272258
<title> bennet replied that he had not   14.100151782106384
but it is returned she for <title> long has just been here and she told me all about it      58.77708613389546
<title> bennet made no answer    15.951776330309308
do not you want to know who has taken it     35.690228592175856
cried his wife impatiently   33.21451580220824
you want to tell me and i have no objection to hearing it    18.231698978152174
this was invitation enough   109.24397450683838
```

2. LM7 (Test): **Average Perplexity: 60.73**



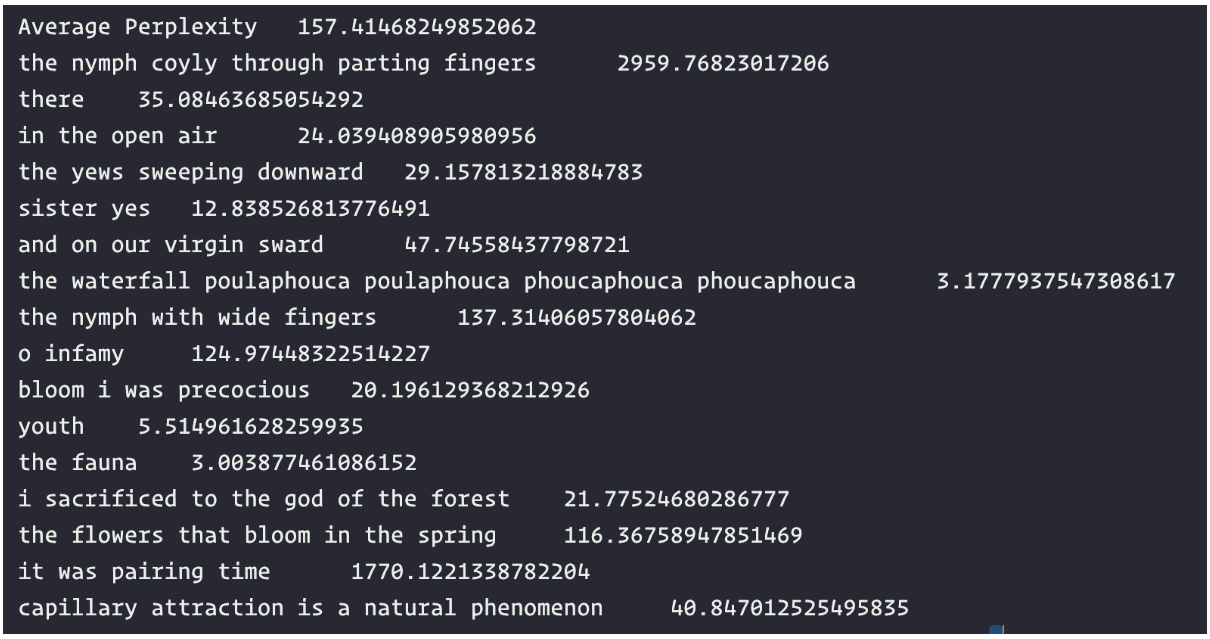
```
Average Perplexity   60.73666011174237
if he does not come to me then said she i shall give him up for ever     13.249567417313715
the gentlemen came and she thought he looked as if he would have answered her hopes but alas     49.9390081876286
the ladies had crowded round the table where miss bennet was making tea and elizabeth pouring out the coffee in s
confederacy that there was not a single vacancy near her which would admit of a chair     56.95306073737012
and on the gentlemen is approaching one of the girls moved closer to her than ever and said in a whisper the men
come and part us i am determined    81.14951129272258
we want none of them do we   14.100151782106384
darcy had walked away to another part of the room    58.77708613389546
she followed him with her eyes envied every one to whom he spoke had scarcely patience enough to help anybody to
then was enraged against herself for being so silly      15.951776330309308
a man who has once been refused      35.690228592175856
how could i ever be foolish enough to expect a renewal of his love   33.21451580220824
is there one among the sex who would not protest against such a weakness as a second proposal to the same woman
231698978152174
```

3. LM8 (Train): **Average Perplexity: 101.65**



```
Average Perplexity   101.65964476444177
1 stately plump buck mulligan came from the stairhead bearing a bowl of lather on which a
crossed     115.63692548173013
a yellow dressinggown ungirdled was sustained gently behind him on the mild morning air
he held the bowl aloft and intoned introibo ad altare dei    90.16046943927815
halted he peered down the dark winding stairs and called out coarsely come up kinch
come up you fearful jesuit   135.25139182720145
solemnly he came forward and mounted the round gunrest   407.941225493337
he faced about and blessed gravely thrice the tower the surrounding land and the awaking
then catching sight of stephen dedalus he bent towards him and made rapid crosses in the
shaking his head     281.5120687422444
stephen dedalus displeased and sleepy leaned his arms on the top of the staircase and loo
face that blessed him equine in its length and at the light untonsured hair grained and h
82207001573164
buck mulligan peeped an instant under the mirror and then covered the bowl smartly   237.
back to barracks     141.7252963832965
he said sternly      28.706515560021653
he added in a preacher is tone for this o dearly beloved is the genuine christine body an
0278884991262
```

LM8 (Test): **Average Perplexity: 157.41**



```
Average Perplexity   157.41468249852062
the nymph coyly through parting fingers      2959.76823017206
there    35.08463685054292
in the open air      24.039408905980956
the yews sweeping downward   29.157813218884783
sister yes   12.838526813776491
and on our virgin sward      47.74558437798721
the waterfall poulaphouca poulaphouca phoucaphouca phoucaphouca      3.1777937547308617
the nymph with wide fingers      137.31406057804062
o infamy     124.97448322514227
bloom i was precocious   20.196129368212926
youth    5.514961628259935
the fauna    3.003877461086152
i sacrificed to the god of the forest    21.77524680286777
the flowers that bloom in the spring     116.36758947851469
it was pairing time      1770.1221338782204
capillary attraction is a natural phenomenon     40.847012525495835
```

Conclusion

We can clearly see that even if the train perplexity of neural models in high but the test perplexity is improved. The most recently trained file on LM6 could not perform better than n-gram that can be because of the train-test split or the overfitting of the model as the corpus is huge and may not contain that much context. So it depends on the corpus as well weather neural networks can outperform n-gram models.

Files Link:

https://drive.google.com/drive/folders/13OgKTDrGJjAmrkbPy-T1LslRNveP-6Bk?usp=sharing