NLP HW2
Name: Naga Hemachand Chinta

1) **Code link**: https://github.com/Nagahemachand/Author-attribution.git
2) **Type of encoding**: ASCII
3) **Information of the language model:** It's a bigram model
4) **Smoothing technique used at the end**: LIDSTONE
5) **Dealing with OOV**: Lidstone smoothing, while primarily designed to handle the problem of zero probabilities for n-grams that appear in the training data, indirectly helps in dealing with out-of-vocabulary (OOV) words during runtime in language modeling. OOV words can be considered as n-grams of length one (unigrams). Lidstone smoothing assigns a non-zero probability to all unigrams, including OOV words, even if they were not observed in the training data.
6) **Tweaks to improve results**: Smoothing techniques are the most important tweaks and then the data preprocessing also is affecting the accuracy.
7) **Accuracies of the authors for development set**:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist
splitting into training and development datasets
training LMs...(this may take a while)
austen  91.6% correct
dickens 63.5% correct
tolstoy 73.9% correct
wilde   64.9% correct
```

8) **Samples**:
Austen:
n=2:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample
training LMs...(this may take a while)
['i', 'know', 'he', 'would', 'be', 'hurt', 'by', 'my', 'failing', 'in', 'such', 'a', 'mark', 'of', 'respect', 'to', 'him
', 'on', 'the', 'present', 'occasion']: austen
perplexity: 161.21224085638684
['if', 'he', 'would', 'act', 'in', 'this', 'sort', 'of', 'manner', 'on', 'principle', 'consistently', 'regularly', 'thei
r', 'little', 'minds', 'would', 'bend', 'to', 'his']: austen
perplexity: 304.15572038490603
['depend', 'upon', 'it', 'emma', 'a', 'sensible', 'man', 'would', 'find', 'no', 'difficulty', 'in', 'it']: austen
perplexity: 233.4018210004582
['to', 'him', 'who', 'has', 'it', 'might', 'not', 'be', 'so', 'easy', 'to', 'burst', 'forth', 'at', 'once', 'into', 'per
fect', 'independence', 'and', 'set', 'all', 'their', 'claims', 'on', 'his', 'gratitude', 'and', 'regard', 'at', 'nought'
]: austen
perplexity: 337.3626662195726
['our', 'amiable', 'young', 'man', 'is', 'a', 'very', 'weak', 'young', 'man', 'if', 'this', 'be', 'the', 'first', 'occas
ion', 'of', 'his', 'carrying', 'through', 'a', 'resolution', 'to', 'do', 'right', 'against', 'the', 'will', 'of', 'other
s']: austen
perplexity: 215.83613049391144
```

n=3:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample_Austen
ngram_n: 3
training LMs...(this may take a while)
['i', 'know', 'he', 'would', 'be', 'hurt', 'by', 'my', 'failing', 'in', 'such', 'a', 'mark', 'of', 'respect', 'to', 'him
', 'on', 'the', 'present', 'occasion']: austen
perplexity: 303.58348450740147
['if', 'he', 'would', 'act', 'in', 'this', 'sort', 'of', 'manner', 'on', 'principle', 'consistently', 'regularly', 'thei
r', 'little', 'minds', 'would', 'bend', 'to', 'his']: austen
perplexity: 472.0287632760828
['depend', 'upon', 'it', 'emma', 'a', 'sensible', 'man', 'would', 'find', 'no', 'difficulty', 'in', 'it']: austen
perplexity: 298.3902498627951
['to', 'him', 'who', 'has', 'it', 'might', 'not', 'be', 'so', 'easy', 'to', 'burst', 'forth', 'at', 'once', 'into', 'per
fect', 'independence', 'and', 'set', 'all', 'their', 'claims', 'on', 'his', 'gratitude', 'and', 'regard', 'at', 'nought'
]: austen
perplexity: 551.5086816886751
['our', 'amiable', 'young', 'man', 'is', 'a', 'very', 'weak', 'young', 'man', 'if', 'this', 'be', 'the', 'first', 'occas
ion', 'of', 'his', 'carrying', 'through', 'a', 'resolution', 'to', 'do', 'right', 'against', 'the', 'will', 'of', 'other
s']: austen
perplexity: 442.8944843896188
```

Dickens:

n=2:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample
training LMs...(this may take a while)
['so', 'long', 'as', 'a', 'servant', 'was', 'present', 'no', 'other', 'words', 'passed', 'between', 'them']: dickens
perplexity: 246.24083262461372
['if', 'he', 'would', 'act', 'in', 'this', 'sort', 'of', 'manner', 'on', 'principle', 'consistently', 'regularly', 'thei
r', 'little', 'minds', 'would', 'bend', 'to', 'his']: austen
perplexity: 304.15572038490603
['no', 'no', 'no', 'said', 'the', 'uncle', 'pleasantly']: dickens
perplexity: 181.76878231747457
['thank', 'you', 'said', 'the', 'marquis', 'very', 'sweetly', 'indeed']: dickens
perplexity: 198.45450942619695
['in', 'effect', 'sir', 'pursued', 'the', 'nephew', 'i', 'believe', 'it', 'to', 'be', 'at', 'once', 'your', 'bad', 'fort
une', 'and', 'my', 'good', 'fortune', 'that', 'has', 'kept', 'me', 'out', 'of', 'a', 'prison', 'in', 'france', 'here']:
dickens
perplexity: 250.2801155219249
```

n=3:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample_Dickens
ngram_n: 3
training LMs...(this may take a while)
['so', 'long', 'as', 'a', 'servant', 'was', 'present', 'no', 'other', 'words', 'passed', 'between', 'them']: dickens
perplexity: 565.57899355506
['no', 'no', 'no', 'said', 'the', 'uncle', 'pleasantly']: dickens
perplexity: 359.18147012984565
['thank', 'you', 'said', 'the', 'marquis', 'very', 'sweetly', 'indeed']: dickens
perplexity: 321.76366908610174
['in', 'effect', 'sir', 'pursued', 'the', 'nephew', 'i', 'believe', 'it', 'to', 'be', 'at', 'once', 'your', 'bad', 'fort
une', 'and', 'my', 'good', 'fortune', 'that', 'has', 'kept', 'me', 'out', 'of', 'a', 'prison', 'in', 'france', 'here']:
dickens
perplexity: 508.9773984143765
['his', 'tone', 'lingered', 'in', 'the', 'air', 'almost', 'like', 'the', 'tone', 'of', 'a', 'musical', 'instrument']: di
ckens
perplexity: 701.2681923493127
```

Tolstoy:

n=2:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample
training LMs...(this may take a while)
['to', 'be', 'able', 'to', 'crush', 'it', 'absolutely', 'he', 'awaited', 'the', 'arrival', 'of', 'the', 'rest', 'of', 't
he', 'troops', 'who', 'were', 'on', 'their', 'way', 'from', 'vienna', 'and', 'with', 'this', 'object', 'offered', 'a', '
three', 'days', 'truce', 'on', 'condition', 'that', 'both', 'armies', 'should', 'remain', 'in', 'position', 'without', '
moving']: tolstoy
perplexity: 386.77892679174363
['marching', 'thirty', 'miles', 'that', 'stormy', 'night', 'across', 'roadless', 'hills', 'with', 'his', 'hungry', 'ill'
, 'shod', 'soldiers', 'and', 'losing', 'a', 'third', 'of', 'his', 'men', 'as', 'stragglers', 'by', 'the', 'way', 'bagrat
in', 'came', 'out', 'on', 'the', 'vienna', 'znaim', 'road', 'at', 'hollabrnn', 'a', 'few', 'hours', 'ahead', 'of', 'the'
, 'french', 'who', 'were', 'approaching', 'hollabrnn', 'from', 'vienna']: tolstoy
perplexity: 633.2441595179863
['bagratins', 'exhausted', 'and', 'hungry', 'detachment', 'which', 'alone', 'covered', 'this', 'movement', 'of', 'the',
'transport', 'and', 'of', 'the', 'whole', 'army', 'had', 'to', 'remain', 'stationary', 'in', 'face', 'of', 'an', 'enemy'
, 'eight', 'times', 'as', 'strong', 'as', 'itself']: tolstoy
perplexity: 548.7894551104067
['the', 'russian', 'emperors', 'aide', 'de', 'camp', 'is', 'an', 'impostor']: tolstoy
perplexity: 371.9610826548146
['bonapartes', 'adjutant', 'rode', 'full', 'gallop', 'with', 'this', 'menacing', 'letter', 'to', 'murat']: tolstoy
perplexity: 821.1559239221157
```

n=3:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample_Tolstoy
ngram_n: 3
training LMs...(this may take a while)
['to', 'be', 'able', 'to', 'crush', 'it', 'absolutely', 'he', 'awaited', 'the', 'arrival', 'of', 'the', 'rest', 'of', 't
he', 'troops', 'who', 'were', 'on', 'their', 'way', 'from', 'vienna', 'and', 'with', 'this', 'object', 'offered', 'a', '
three', 'days', 'truce', 'on', 'condition', 'that', 'both', 'armies', 'should', 'remain', 'in', 'position', 'without', '
moving']: tolstoy
perplexity: 684.4097061239222
['marching', 'thirty', 'miles', 'that', 'stormy', 'night', 'across', 'roadless', 'hills', 'with', 'his', 'hungry', 'ill'
, 'shod', 'soldiers', 'and', 'losing', 'a', 'third', 'of', 'his', 'men', 'as', 'stragglers', 'by', 'the', 'way', 'bagrat
in', 'came', 'out', 'on', 'the', 'vienna', 'znaim', 'road', 'at', 'hollabrnn', 'a', 'few', 'hours', 'ahead', 'of', 'the'
, 'french', 'who', 'were', 'approaching', 'hollabrnn', 'from', 'vienna']: tolstoy
perplexity: 1282.7936734387906
['bagratins', 'exhausted', 'and', 'hungry', 'detachment', 'which', 'alone', 'covered', 'this', 'movement', 'of', 'the',
'transport', 'and', 'of', 'the', 'whole', 'army', 'had', 'to', 'remain', 'stationary', 'in', 'face', 'of', 'an', 'enemy'
, 'eight', 'times', 'as', 'strong', 'as', 'itself']: tolstoy
perplexity: 930.5469570927794
['the', 'russian', 'emperors', 'aide', 'de', 'camp', 'is', 'an', 'impostor']: tolstoy
perplexity: 554.0280652573053
['bonapartes', 'adjutant', 'rode', 'full', 'gallop', 'with', 'this', 'menacing', 'letter', 'to', 'murat']: tolstoy
perplexity: 1085.101594364378
```

Wilde:

n=2:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample
training LMs...(this may take a while)
['as', 'he', 'thought', 'of', 'it', 'a', 'sharp', 'pang', 'of', 'pain', 'struck', 'through', 'him', 'like', 'a', 'knife'
, 'and', 'made', 'each', 'delicate', 'fibre', 'of', 'his', 'nature', 'quiver']: wilde
perplexity: 263.69464742469376
['he', 'stood', 'there', 'motionless', 'and', 'in', 'wonder', 'dimly', 'conscious', 'that', 'hallward', 'was', 'speaking
', 'to', 'him', 'but', 'not', 'catching', 'the', 'meaning', 'of', 'his', 'words']: wilde
perplexity: 352.47919497028244
['he', 'would', 'become', 'dreadful', 'hideous', 'and', 'uncouth']: wilde
perplexity: 327.33271303397555
['as', 'he', 'thought', 'of', 'it', 'a', 'sharp', 'pang', 'of', 'pain', 'struck', 'through', 'him', 'like', 'a', 'knife'
, 'and', 'made', 'each', 'delicate', 'fibre', 'of', 'his', 'nature', 'quiver']: wilde
perplexity: 263.69464742469376
['who', 'wouldnt', 'like', 'it']: wilde
perplexity: 192.93482478538834
['it', 'is', 'one', 'of', 'the', 'greatest', 'things', 'in', 'modern', 'art']: wilde
perplexity: 97.05751942985485
```

n=3:

```
C:\Users\nagah\Downloads\NLP_hw2>python classifier.py authorlist -test Sample_Wilde
ngram_n: 3
training LMs...(this may take a while)
['as', 'he', 'thought', 'of', 'it', 'a', 'sharp', 'pang', 'of', 'pain', 'struck', 'through', 'him', 'like', 'a', 'knife'
, 'and', 'made', 'each', 'delicate', 'fibre', 'of', 'his', 'nature', 'quiver']: wilde
perplexity: 532.6274830621715
['he', 'stood', 'there', 'motionless', 'and', 'in', 'wonder', 'dimly', 'conscious', 'that', 'hallward', 'was', 'speaking
', 'to', 'him', 'but', 'not', 'catching', 'the', 'meaning', 'of', 'his', 'words']: wilde
perplexity: 589.4664356464593
['he', 'would', 'become', 'dreadful', 'hideous', 'and', 'uncouth']: wilde
perplexity: 370.082834603288
['as', 'he', 'thought', 'of', 'it', 'a', 'sharp', 'pang', 'of', 'pain', 'struck', 'through', 'him', 'like', 'a', 'knife'
, 'and', 'made', 'each', 'delicate', 'fibre', 'of', 'his', 'nature', 'quiver']: wilde
perplexity: 532.6274830621715
['who', 'wouldnt', 'like', 'it']: wilde
perplexity: 248.15954331475027
['it', 'is', 'one', 'of', 'the', 'greatest', 'things', 'in', 'modern', 'art']: wilde
perplexity: 166.86064821243482
```

9) **Steps on understanding the procedural development for the code**:

```
#Step-1: Define help description for input arguments from cmd line
and create argument parser to store the cmd line arguments.
#Step-2: Load the text data from the needed files based on cmd line
argument.
#Step-3: Preprocess the loaded text data.
##Sub step-3a: Lower case
##Sub step-3b: Remove newline (Other inf error will occur)
##Sub step-3c: Remove punctuations (Other inf error may occur)
##Sub step-3d: Remove extra space (Other inf error may occur)
##Sub step-3e: Split sentence into word lists
##Sub step-3f: Remove empty lists (Other inf error may occur)
#Step-4: Split the data as per the criteria defined as per the
arguments from cmd line.
#Step-5: Build a function for training model
#Step-6: Build a function for development
#Step-7: Build a function for testfile testing
```

10) **Contributions**:

a) ChatGPT (for code skeletal structure and conceptual understanding)
b) Stack overflow
c) Andrej Karpathy tutorials
d) Github repositories of ngram model applications.