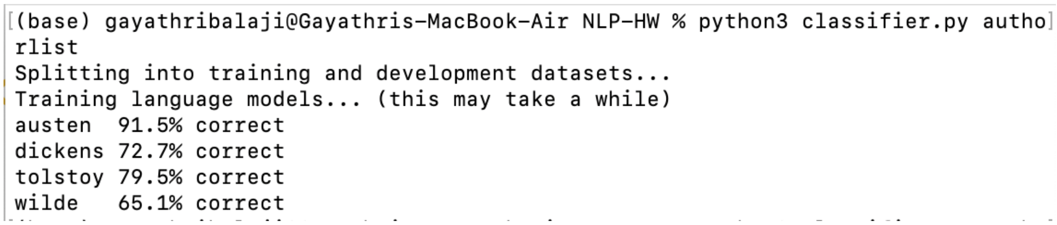
CSCI 5541 (F23) HW2: Building n-gram language models from scratch
REPORT

Name: Gayathri Balaji

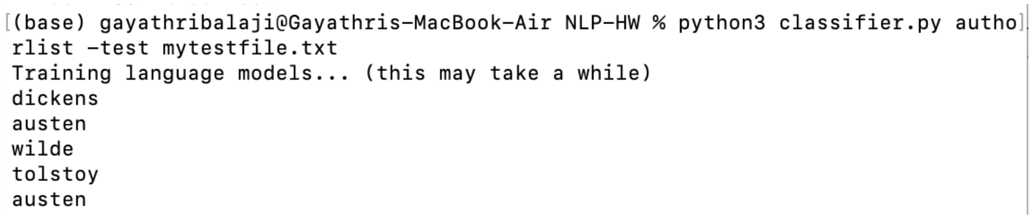
1. **How to run:**

   To check authorship classification: python3 classifier.py authorlist



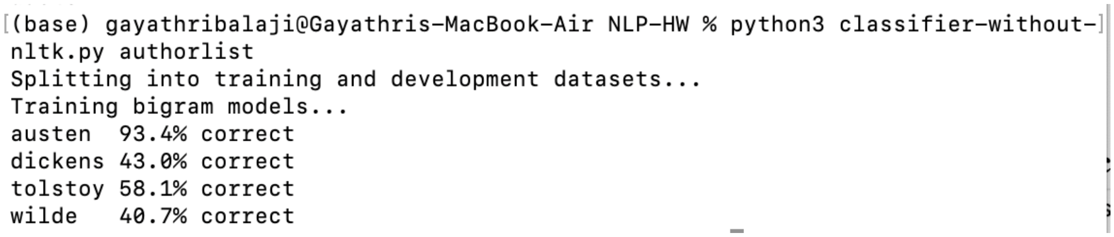
```
[(base) gayathribalaji@Gayathris-MacBook-Air NLP-HW % python3 classifier.py autho]
 rlist
 Splitting into training and development datasets...
 Training language models... (this may take a while)
 austen  91.5% correct
 dickens 72.7% correct
 tolstoy 79.5% correct
 wilde   65.1% correct
```

   To check classification for test file: python3 classifier.py authorlist -test mytestfile.txt



```
[(base) gayathribalaji@Gayathris-MacBook-Air NLP-HW % python3 classifier.py autho]
 rlist -test mytestfile.txt
 Training language models... (this may take a while)
 dickens
 austen
 wilde
 tolstoy
 austen
```

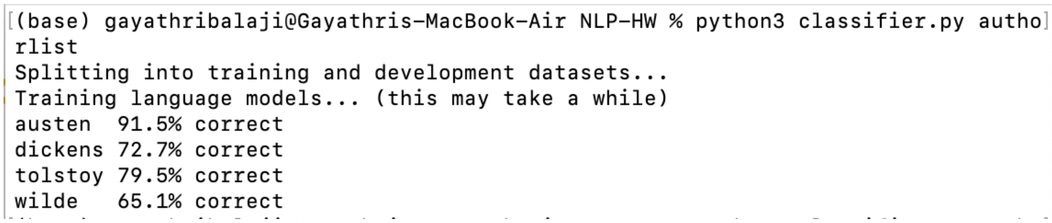
   To check classification n-gram language models without using NLTK: python3 classifier-without-nltk.py authorlist



```
[(base) gayathribalaji@Gayathris-MacBook-Air NLP-HW % python3 classifier-without-]
 nltk.py authorlist
 Splitting into training and development datasets...
 Training bigram models...
 austen  93.4% correct
 dickens 43.0% correct
 tolstoy 58.1% correct
 wilde   40.7% correct
```

2. **Type of Encoding:** UTF-8
3. **What information is in your Language Models (bigrams, trigrams, etc):** Used bigram as n-gram model
4. **What method of smoothing you are using:** Implemented many smoothing methods such as Maximum Likelihood Estimation (MLE) as the baseline, Stupid Backoff, Kneser-Ney Interpolated, Lidstone Smoothing and Witten-Bell Smoothing of which Lidstone was the one which I chose based on analysis. We want to avoid the occurrence of an infinite perplexity score, and only the Lidstone method gives all finite values. Additionally, it has a lower average perplexity in bigrams language model than in trigrams model and performs better with the development/validation sets. Therefore, we conclude that the Lidstone smoothing method applied on the bigram language model is the best option.

5. **How do you deal with out-of-vocabulary words during run time when you build a language model? :** When an OOV word is encountered in a context where no n-gram with that word exists in the training data, Lidstone smoothing allows for a non-zero probability to be assigned. Additionally, padded_everygram_pipeline automatically assigns <UNK> token to any out of vocabulary words
6. **Preprocessing done:** Tokenize sentences, remove empty spaces, extra lines
7. **Any other tweaks you made to improve results (backoff, etc.):** Apart from Smoothing and preprocessing, backoff to bigram was implemented
8. **The results (i.e., accuracy for each author) you get with the given data with an automatically extracted development set (i.e. the output from running it without the -test flag)**:



```
[(base) gayathribalaji@Gayathris-MacBook-Air NLP-HW % python3 classifier.py autho]
rlist
Splitting into training and development datasets...
Training language models... (this may take a while)
austen  91.5% correct
dickens 72.7% correct
tolstoy 79.5% correct
wilde   65.1% correct
```

9. **For each of your language models, generate five samples of each author given the same prompt you specify with their perplexity scores (i.e., a total of five samples and perplexity scores by four different language models)**:

Author	Sentence	Austen	Dickens	Tolstoy	Wilde
Austen	['thing', 'that', 'is', 'felt', 'more']	483.75	1555.82	1748.79	1260.12
Austen	['and', 'they', 'were', 'equally', 'in']	124.08	435.55	433.36	440.14
Austen	['who', 'will', 'be', 'almost', 'sure']	185.40	1383.36	1877.73	1046.56
Austen	['coming', 'to', 'go', 'over', 'the']	590.87	1062.41	561.83	957.88
Austen	['come', 'and', 'in', 'berkeley', 'street']	186.66	733.03	956.09	623.44
Dickens	['there', 'was', 'in', 'every', 'muscle']	306.01	238.36	239.02	293.66
Dickens	['and', 'this', 'sort', 'of', 'my']	238.19	296.34	362.78	593.46

Dickens	['which', 'were', 'all', 'and', 'there']	280.48	234.06	244.32	351.23
Dickens	['coin', 'from', 'me', 'or', 'semblance']	4445.96	1455.67	6062.22	4635.81
Dickens	['client', 'and', 'interest', '</s>', '<s>']	11245.62	4567.14	17628.44	11234.98
Tolstoy	['then', 'sviazhsky', 'keeping', 'based', 'on']	2364.26	2225.63	482.31	2464.53
Tolstoy	['and', 'this', 'struggle', 'between', 'the']	834.83	1864.59	661.27	1925.92
Tolstoy	['which', 'was', 'able', 'to', 'the']	410.39	506.00	524.29	842.24
Tolstoy	['calm', 'i', 'going', 'to', 'prince']	6549.22	7177.81	433.43	4310.08
Tolstoy	['by', 'and', 'imploringly', 'at', 'all']	706.67	658.33	535.07	781.01
Wilde	['them', 'said', 'lord', 'francis', 'he']	6797.25	4503.50	4551.54	695.84
Wilde	['agitation', 'and', 'the', 'fifteenth', 'century']	2871.21	6071.31	3417.04	964.14
Wilde	['whom', 'was', 'a', 'cigarette', 'i']	4169.19	1109.89	1331.17	821.33
Wilde	['by', 'saracen', 'cards', 'painted', 'of']	3630.46	3761.45	4799.71	826.94
Wilde	['but', 'dont', 'know', 'everybody', '</s>']	1502.59	390.18	553.88	294.63

10. **Analysis of Different Smoothing methods:**
Tried different smoothing methods with combinations of n-grams.

Training Method	Austen	Dickens	Tolstoy	Wilde
SB Bigram	89.0%	43.7%	55.7%	47.3%
Lidstone Bigram	90.4%	66.5%	77.8%	63.5%
Lidstone Bigram	87.6%	52.8%	53.8%	62.1%
WBI Bigram	88.1%	72.7%	79.3%	67.5%
WBI trigram	87.8%	55.2%	56.1%	63.1%

While accuracy of models using WBI were high for some authors, Lidstone was the smoothening which gave the least perplexity. WBI and SB were giving inf perplexity for some values. That is why the Lidstone smoothing method was chosen based on both perplexity and accuracy.

11. **Classification of Sample:** Below are the samples taken from authors famous books and placed in mytestfile.txt

- **Dickens**: Darkness shrouded the city streets, and the biting winter wind swept through the cobbled lanes, as a lonely orphan boy named Oliver Twist ventured forth, his empty stomach a constant companion on this bleak and unforgiving night\n
- **Austen**: It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.\n
- **Wilde**: The only way to get rid of a temptation is to yield to it. Resist it, and your soul grows sick with longing for the things it has forbidden to itself.\n
- **Tolstoy**: He sought all his life for a good place and a good position, and always wore a sword as an officer of state and kissed up to people. But the moment he got drunk, he bared his head and began cursing.\n
- **Austen**: The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property, where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance.\n