CSCI 5541 Assignment 2

Building ngram language models from scratch

BH's Group
Member(s):
Bin Hu

March 1, 2023

1 Abstract

In this project we implement trigram language models to learn the text of four writers by nltk package. A text classifier is made based on these models which is able to estimate which writer is the author of the given sentence. And sentences are generated by each language models given the same prompt.

2 Text Pre-processing

It is assumed that the author list and the text files are in the same folder as the classifier.py. The program firstly read the authors' names in the authorlist.txt, and then read the for text files line by line. The encoder we used is ASCII because the difference in LM perplexities of the two encoding types (the other one is UTF-8) is negligible in our experiment. When reading the lines, the newline marker (/n) is removed and all lines of texts are concatenate together because they are not splitted into sentences correctly. The nltk function sent_tokenize is used to break the sentences and the sentences are all turned to be lower case.

The tokenizer we used is the wordpunct_tokenize function which is native to nltk, becuase it is observed that it can contribute to 2-3% higher accuracy for each language model we generated that those are tokenized by nltk word_tokenize tokenizer. Such a better tokenizer can help solve the problem of out-of-vocabulary words. And all sentences are padded with the sentence starting/ending markers at the last step.

Another good tokenizer is the word piece tokenizer (as is applied in BERT) because it can break the complex word into small pieces which are likely to be included in the vocabulary. However, we did not used the word piece tokenizer because it is slower.

3 Developing Set Creation

It is required that in the cases that the test file is not given, 10% of the provided sentences should be separated as the developing set. Here we used the random.shuffle() function to shuffle the splitted sentences to ensure the sentences are splitted fairly. The splitting is done by list slicing.

4 Trigram Model and Optimization

We build the trigram language models with the KneserNeyInterpolated class which is native to nltk. Interpolated version of Kneser-Ney smoothing with absolute discount is applied in this model thus it shows a much better performance than the most likelihood estimation (MLE) model. By adjusting the discount as a hyperparameter, we found the best perplexity shows when then discount is 0.75.

5 Evaluation

The evaluation is based both on perplexity and the classification accuracy. Perplexity can be calculated simply by the function of the language model class. Classification is calculated based on the for trigram models' perplexity of a given sentences and the model of the lowest perplexity is considered as the estimation.

6 Appendices

6.1 Output without given the test file

```
python classifier.py authorlist

>>>
### HW2 for CSCI 5541 23S ###
### Author: Bin Hu (BH's Group) ###
splitting into training and development...
training LMs... (this may take a while)
Results on dev set:
austen:         79.07% correct  350.44 perplexity
dickens:        73.49% correct  388.38 perplexity
tolstoy:        74.80% correct  401.16 perplexity
wilde:          69.17% correct  460.86 perplexity
```

6.2 Generated samples

Two first words are given as the prompt:

austen:

i would ' . ' ' ' . . . ' . . ' ' . ' ' . ' ' . .

it is mind hands character , letter great opinion , curiosity mother house - mind visit style side feelings - feelings creation .

i do '

that is kindness affection post first marrying attentions , mind situation , marriage for education feelings letter civilities reception thoughts end countenance name

how about think let tell do thinkˆ let signifytalking do let t̂alk well think let they get k̂now let

dickens:

i would ' . ' ' ' . . . ' . . ' ' . ' ' . ' ' . . crownsresign think as our .!' ' then , ' – have have be i it be have have it .!' '

it is mind hands character , letter great opinion , curiosity mother house - mind visit style side feelings - feelings creation . agent a the great time , only , my time many bosom bench a right not step the kite houseas ,

i do ' miss ' ' ' ' ' ' my ' ' me ' ' ' me me ' ' ' ' '

that is kindness affection post first marrying attentions , mind situation , marriage for education feelings letter civilities reception thoughts end countenance name bench putting the such – sake name sake lips all aunt a a so face quite society quite the head walk

how about think let tell do thinkˆ let signifytalking do let t̂alk well think let they get k̂now let it know like you know remember matter you do know theer know the that you mind ought do you made ought

tolstoy:

i would ' . ' ' ' . . . ' . . ' ' . ' ' . ' ' . . crownsresign think as our .!' ' then , ' – have have be i it be have have it .!' ' ' ' ' ' ' ' ' ' ' ' ' long ' ' ' ' ' ' ' ' ' '

it is mind hands character , letter great opinion , curiosity mother house - mind visit style side feelings - feelings creation . agent a the great time , only , my time many bosom bench a right not step the kite houseas , voice remark character success suite face getting the theonly an how horse there ' the head that right ' the not i do ’ miss ’ ’ ’ ’ ’ ’ my ’ ’ me ’ ’ ’ me me ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ god ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’

that is kindness affection post first marrying attentions , mind situation , marriage for education feelings letter civilities reception thoughts end countenance name bench putting the such – sake name sake lips all aunt a a so face quite society quite the head walk your that meeting not presence what camp hand room guns the sorrow , side . battery servant our ’ hand so

how about think let tell do thinkˆ let signifytalking do let ˆtalk well think let they get ˆknow let it know like you know remember matter you do know theer know the that you mind ought do you made ought it be i know tell ’ ’ remember see care you it always ’ get be tell be gush ’ like

wilde:

i would ’ . ’ ’ ’ . . . ’ . . ’ ’ . ’ ’ . ’ ’ . . crownsresign think as our .!’ ’ then , ’ – have have be i it be have have it .!’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ long ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ i ’ ’ - ’ i ’ ’ ’ have i with ’ for with , ’ with

it is mind hands character , letter great opinion , curiosity mother house - mind visit style side feelings - feelings creation . agent a the great time , only , my time many bosom bench a right not step the kite houseas , voice remark character success suite face getting the theonly an how horse there ’ the head that right ’ the not verse brother house plays as theory feet age ripening divinity views four court own . eyes hand life influence company hand

i do ’ miss ’ ’ ’ ’ ’ ’ my ’ ’ me ’ ’ ’ me me ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ god ’ , ’

that is kindness affection post first marrying attentions , mind situation , marriage for education feelings letter civilities reception thoughts end countenance name bench putting the such – sake name sake lips all aunt a a so face quite society quite the head walk your that meeting not presence what camp hand room guns the sorrow , side . battery servant our ’ hand so , secretary boyhood all . fault sake . absence plays lips heart head hand lips own death admirer lips face absurd

how about think let tell do thinkˆ let signifytalking do let ˆtalk well think let they get ˆknow let it know like you know remember matter you do know theer know the that you mind ought do you made ought it be i know tell ’ ’ remember see care you it always ’ get be tell be gush ’ like - ever quarrel talk know wantme bear go know sneer like want you like mean want go like you changed go