Title: Language Modeling
Module: COM6513 Natural Language Processing
Name: Ziyuan XIA
Registration Number: 180216466

Description:

In this lab, three language models (unigram, bigram and bigram with add-one smooth). A question file is applied to these models to test their performance. The script is edited and compiled under Windows OS.

The unigram model resulted 60% in accuracy, the bigram model resulted 80% and the bigram model with add-one smoothing resulted 90%. The detailed information is shown in Figure 1.

```
'The result of unigram model is: '
["<s> i don't know whether to go out or not </s> ",
 '<s> we went through the door to get inside </s> ',
 '<s> they all had a peace of the cake </s> ',
 '<s> she had to go to court to prove she was innocent </s> ',
 '<s> we were only allowed to visit at certain times </s> ',
 '<s> she went back to check she had locked the door </s> ',
 '<s> can you here me </s> ',
 '<s> do you usually eat serial for breakfast </s> ',
 '<s> she normally choose with her mouth closed </s> ',
 "<s> i'm going to sell it on the internet </s> "]
------------------------------------------------------------
'The result of bigram model is: '
["<s> i don't know whether to go out or not </s> ",
 '<s> we went through the door to get inside </s> ',
 '<s> they all had a piece of the cake </s> ',
 '<s> she had to go to court to prove she was innocent </s> ',
 '<s> we were only allowed to visit at certain times </s> ',
 '<s> she went back to check she had locked the door </s> ',
 '<s> can you hear me </s> ',
 'zero probability, cannot choose',
 'zero probability, cannot choose',
 "<s> i'm going to sell it on the internet </s> "]
------------------------------------------------------------
'The result of bigram model with add-1 smooth is: '
["<s> i don't know whether to go out or not </s> ",
 '<s> we went through the door to get inside </s> ',
 '<s> they all had a piece of the cake </s> ',
 '<s> she had to go to court to prove she was innocent </s> ',
 '<s> we were only allowed to visit at certain times </s> ',
 '<s> she went back to check she had locked the door </s> ',
 '<s> can you hear me </s> ',
 '<s> do you usually eat cereal for breakfast </s> ',
 '<s> she normally choose with her mouth closed </s> ',
 "<s> i'm going to sell it on the internet </s> "]
```

Figure 1. Result of all models

The unigram model is built up by single words in the corpus. The probability of each word is stored as a value corresponding to each word in a dictionary. There is a unigram counting dictionary built as the count of context for bigram language model.

The bigram model is built up by tuples which consist of every two words in unigram model. The probability of each tuple represents the conditional probability of a word after a one-word context.

Considering zero probability problem might occur, each bigram is assumed to be seen once more (add-one smoothing) so that the model can be more robust. This is called bigram model with add-one smoothing.

Observation and discussion:

The unigram model has the poorest performance among three models. The reason is that it has not considered any context. The answer is chosen by its probability in the corpus which means the answer is always the most frequent word among choices no matter it makes sense or not (i.e. word “the” will always be chosen if it is in the choice set).

The bigram model has an acceptable performance which every successfully returned answer is correct. This proved that considering context can improve the quality of the language model. However, considering longer context make zero probability problem more frequent. There are two zero probability bigrams detected in the used question set which the model cannot choose a answer.

The bigram model with add-one smoothing has the best performance among three implemented models. There is only one incorrectly answered question out of ten questions which is “normally choose” and “choose with”, the correct answer should be “normally chews” and “chews with”. This happened because “choose” and “chews” are both verbs and both frequently seen before “with” or after “normally” but “choose” is more frequent than “chews”. This problem can be solved by using a model which considers longer context. For example, “normally chews with her mouth” should be more frequent than “normally choose with her mouth” does.

In conclusion, the results are quite close to they are expected: considering longer context gives higher accuracy but more likely to have zero probability problems.