COM6513, Lab2

Registration Number 170224545

22/02/2018

1. description

In this lab, the program successfully implements three language models (unigram, bigram, bigram with add-1 smoothing (Laplace)) to perform sentence completion.

The program lowercases the texts and removes punctuation. Then three models are built.

The program can be executable by running:

python3 lab2.py news-corpus-500k.txt questions.txt

And the result is as follows:

(1) The answers using the unigram model are：

[whether, through, peace, court, allowed, check, here, serial, choose, sell]

(2) The answers using the bigram model are：

[whether, through, piece, court, allowed, check, hear, ____, ____, ____]

(3) The answers using the bigram with add-1 smoothing model are：

[whether, through, piece, court, allowed, check, hear, cereal, choose, sell]

‘____’ in the (2) means that the two answers have the same zero probability, so the program cannot choose which is better.

2. Evaluation

(1) Accuracy:

The program prints the probabilities of each answer. As the probabilities show, no language model returns only 0 probabilities. no tie is with non-zero probabilities. So the program is correct.

(2) Discussion of result

As we can know, the correct answer is:

[whether, through, piece, court, allowed, check, hear, cereal, chews, sell].

The unigram model has 6 correct answers and 4 wrong answer. The accuracy rate is 60%.

The bigram model has 7 correct answers and 3 wrong answer. The accuracy rate is 70%.

The bigram with add-1 smoothing model has 9 correct answers and 1 wrong answer. The accuracy rate is 90%.

As we can see, the unigram model has the lowest accuracy rate, because it only calculates the probability of a single word. The bigram model is better but it also suffers from the zero probability. The bigram with add-1 smoothing model has the highest accuracy rate because it avoids zero probabilities.