Language modelling.

Registration Number : **180128033**

In this report, we are implementing three language models to preform sentence completion.We are considering possible candidate words and then asking language model to suggest which version of sentence is the most possible one.

1 Unigram Language Model

In this model,we are calculating probability of 2 given words and then returning the one with higher probability.After applying the above approach to the testing data, we are getting accuracy of **0.6** (6 sentence out of 10).

Test Sentences:

- I don't know ____ to go out or not. : weather/whether **correct**
- We went ____ the door to get inside. : through/threw **correct**
- They all had a ____ of the cake. : piece/peace **wrong**
- She had to go ____ to prove she was innocent. : caught/court **correct**
- We were only ____ to visit at certain times. : aloud/allowed **correct**
- She went back to ____ she had locked the door. : cheque/check **correct**
- Can you ____ me ? : hear/here **wrong**
- Do you usually eat ____ for breakfast ? : serial/cereal **wrong**
- She normally ____ with her mouth closed . : chews/choose **wrong**
- I'm going to ____ it on the internet . : cell/sell **correct**

2 Bigram Language Model

In bigram model, we have used words not whole sentence to predict the word because for eg: Ajay is going/went to India. Let 'going' and 'went' be the word to be predicted.So, we consider word before and after the given words to find the word with highest probability neglecting the other words in the sentence because they don't effect the result.

- We are adding the probability of the word to be predict with the previous word and then dividing the whole with the probability of the previous word for the both word given for prediction.
- We are adding the probability of the word to be predict with the next word to it and then dividing the whole with the probability of the given words for the both word given for prediction.
- After implementing above steps, we are multiplying the results from above 2 steps for both the predicted word and then comparing them to find the word with higher probability.
- While applying above approach, we have assigned '0' to the word that have not occured.
- After implementation and testing, we got 8 correct sentence out of 10 sentences i.e accuracy of **0.8** which greater **Unigram language model.**

Test Sentences:

- I don't know ____ to go out or not. : weather/whether **correct**
- We went ____ the door to get inside. : through/threw **correct**
- They all had a ____ of the cake. : piece/peace **correct**
- She had to go ____ to prove she was innocent. : caught/court **correct**
- We were only ____ to visit at certain times. : aloud/allowed **correct**
- She went back to ____ she had locked the door. : cheque/check **correct**
- Can you ____ me ? : hear/here **correct**
- Do you usually eat ____ for breakfast ? : serial/cereal **wrong**
- She normally ____ with her mouth closed . : chews/choose **wrong**
- I'm going to ____ it on the internet . : cell/sell **correct**

3 Bigram Language Model with smoothing

In bigram model with smoothing, we have used words not whole sentence to predict the word because for eg: Ajay is going/went to India. Let 'going' and 'went' be the word to be predicted.So, we consider word before and after the given words to find the word with highest probability neglecting the other words in the sentence because they don't effect the result.

- We are adding the probability of the word to be predict, the probability previous word and 1, then dividing the total of the probability of the previous word and the count of bigram dictionary for the both word given for prediction.
- We are adding the probability of the word to be predict, the probability next word and 1, then dividing the total of the probability of the given words and the count of bigram dictionary for the both word given for prediction.
- After implementing above steps, we are multiplying the results from above 2 steps for both the predicted word and then comparing them to find the word with higher probability.
- After implementation and testing, we got 10 correct sentence out of 10 sentences i.e accuracy of **1** which greater than the than the **Bigram language model** and **Unigram language model.**

e **Test Sentences:**

- I don't know ____ to go out or not. : weather/whether **correct**
- We went ____ the door to get inside. : through/threw **correct**
- They all had a ____ of the cake. : piece/peace **correct**
- She had to go ____ to prove she was innocent. : caught/court **correct**
- We were only ____ to visit at certain times. : aloud/allowed **correct**
- She went back to ____ she had locked the door. : cheque/check **correct**
- Can you ____ me ? : hear/here **correct**
- Do you usually eat ____ for breakfast ? : serial/cereal **correct**
- She normally ____ with her mouth closed . : chews/choose **correct**
- I'm going to ____ it on the internet . : cell/sell **correct**

4 Discussion

We are getting higher accuracy in bigram with smoothing as compared to the unigram and bigram language model because we have non-zero probability values in bigram with smoothing language model, whereas in bigram and unigram we have zero probabilility values.