Language Model in NLP

What is language model in NLP?

- It is a model which knows our language.
- A model of the probability of a sequence of words.
- It estimates the relative likelihood of different phases and are useful in many NLP applications.
- The goal of probabilistic language model is to calculate the probability of a sentence of sequence of words.

$$P(w)=P(w_1, w_{2,} w_{3,} w_{4, \ldots,} w_n)$$

- It can be used to find the probability of the next word in the sentence.

$$P(w_s)=P(w_s|w1, w_{2,} w_{3,} w_{4, \ldots,} w_{s-1})$$

Language Model

I am the king of the ……world……

Input Predicted Word

I am the king of the world ……of……

Input Predicted Word

Example: keyboard of Mobile phone

Advantages of LM

✓It can predict what words are likely to come next in a text.

Ex: Suggest completions for an email o text message.

✓Capable to compute more probable alterations to a text

Ex: Suggest spelling or Grammar corrections

✓With a pair of models, we can compute the most probable translation of sentences.

✓With some example questions/answer pairs as training data, we can compute the most likely answer to a question.

Corpus

- **corpus** is a collection of texts, on which we can perform various natural language processing (NLP) functions.

- In simplest terms, a corpus is a folder of text files on your computer, and corpus readers process all these text files at once, though each file can be called on individually.

Feature, Document and Corpus

Feature

Every unique word in the corpus is considered as a feature.

Document

A document is a single text data point (a text file, book, blog, article, webpage).

Tokenization

It is the process of breaking text into pieces (called tokens).

Corpus

It a collection of all the documents present in our dataset.

Example :

Dog hates a cat. It loves to go out and play. Cat loves to play with a ball.

Corpus = "Dog hates a cat. It loves to go out and play. Cat loves to play with a ball."
Documents = [1. dog hates a cat.
2. it loves to go out and play.
3. cat loves to play with a ball.]

Features= ['and', 'ball', 'cat', 'dog', 'go', 'hates', 'it', 'loves', 'out', 'play', 'to', 'with']

Types of Language Model

- The bag-of-words model
- N-gram word models
- Other n-gram models
- Smoothing n-gram models
- Word representations
- Parts-of-speech (POS) tagging
- Grammar based Language modelling
- Statistical language modelling

N-gram Language Modelling (Probabilistic Model)

- n-gram : sequence of n words
- 1-gram (Unigram)(having no history word)
- 2-gram (Bigram) (having one history word)
- 3-gram(Trigram) (having three history word)
- N-gram (having n-1 history word)

Example: I am the king

[I] [am] [the] [king]

[I am] [am the] [the king]

[I am the] [am the king]

N-gram Language Modelling (Probabilistic Model)

- Unigram Probability $P(w) = \frac{count(w)}{N}$
- Bayes Rule $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Bigram Probability $P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$
- Trigram Probability $P(w_i|w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2} w_{i-1})}$

N-gram Language Modelling (Probabilistic Model)

- Corpus

The girl bought a chocolate

The boy ate the chocolate

The girl bought a toy

The girl played with the toy

Example : For Unigram

P(the)=0.6
P(girl)=0.3
P(bought)=0.2
P(a)=0.2
P(chocolate)=0.2
P(boy)=0.1
P(ate)=0.1
P(toy)=0.2
P(played)=0.1
P(with)=0.1

- Vocabulary/Feature

{the, girl, bought, a, chocolate, boy, ate, toy, played, with}

N=No of features = 10

N-gram Language Modelling (Probabilistic Model)

- Input : The
- Output: The girl
- P(girl|the) = $\frac{count(the,girl)}{count(the)} = \frac{3}{6}$ = 0. 5
- P(boy|the)= $\frac{count(the,boy)}{count(the)} = \frac{1}{6}$=0.166
- Probabilities for our vocabulary for input: The ……

the = 0
girl = 0.5
bought=0
a=0
chocolate =0.166
boy=0.166
ate=0.166
toy=0
played=0
with=0

N-gram Language Modelling (Probabilistic Model)

- Input : The girl
- Output: The girl bought
- P(bought|the,girl) = $\frac{count(the,girl,bought)}{count(the\ girl)} = \frac{2}{3} = 0.67$
- P(played|the ,girl) = $\frac{count(the,girl,played)}{count(the,girl)} = \frac{1}{3} = 0.33$
- Probabilities for our vocabulary for input: The girl ……

The = 0
Girl = 0
Bought=o.67
A=0
Chocolate =0
Boy=0
Ate=0
Toy=0
Played=0.33
With=0

N-gram Language Modelling (Probabilistic Model)

- Input : The boy
- Output: The boy ate
- P(ate|the,boy) = $\frac{count(the,boy,ate)}{count(the\ boy)} = \frac{1}{1} = 1$
- Probabilities for our vocabulary for input: The boy ……

the = 0
girl = 0
bought=0
a=0
chocolate =0
boy=0
ate=1
toy=0
played=0
with=0

N-gram Language Modelling (Probabilistic Model)

Disadvantages of N-Grams

1. It has too many features.

2. Due to too many features, the feature set becomes too dense and is computationally expensive.

3. Choose the optimal value of N is not that easy task.

THANK YOU