NLP

Language Modeling

- → N-gram
- → Bag of words (BOW)
- → TF - IDF

→ predicting the next word

→ A language model is a statistical model that predicts the probability of a sequence of words in a given context

→ It helps generate coherent and grammatically correct sentences.

Example

"The cat is on the ______"

↓

mat or matress

→ A language model can predict that the missing word is likely mat or matress based on the context

Application: Machine Transalation, Speech Recognition, text generation.

I) N-grams → contigous sequence of N items (words)

Sentence: "I love natural language processing."

N-grams → singe words (unigrams)
pairs of adjacent word (biagrams)
trigrams (three-word sequence)

unigrams   biagrams   trigrams

["I", "love", "natural, language", processing"]
→ unigrams

[I love", "love natural", "natural language",
"language processing"]
→ biagrams

["I love natural", "love natural language",
natural lang. processing"]

↳ trigrams

→ N-grams help capture local context and improve language modeling.

II) Bag of words (BOW):

Basic Terminologies

(i) CORPUS → Paragraph

(ii) Documents → sentence

(iii) Vocabulary → unique words

(iv) words → vector

{ D1 → "The cat chased the mouse."
D2 → "The mouse ran away." } CORPUS

→ The: 4, cat: 1, chased: 1, mouse: 2, ran: 1, away: 1

→ The BoW model represents a document as a collection of words, ignoring their order. It creates a vector where each dimension corresponds to a unique word, and the value represents the word's frequency in the document.

BOW:

Vocabulary: ["The", "cat", "chased", "mouse", "ran", "away"]

→ BOW is simple but loss word order and context.

III) TF - IDF (Term Frequency - Inverse Document Frequency)

→ TF-IDF is a numerical statistic used to evaluate the importance of a term within a document relative to a collection of documents (corpus).

Term Frequency (TF)

Measures how often a term appears in a document.

Inverse Document Frequency (IDF)

Measures how relevant a term is across the entire corpus

Example

D1 : The cat sat on the mat.

D2 : the dog lay on the mat.

For the term "cat"

- TF (in document 1): 1
- IDF (across corpus): log(Total doc. / Doc. with cat)

$$= \log(2/1)$$

$$= \log(2) = 0.301$$

$$\therefore \; TFIDF = TF \times IDF$$

$$= 1 \times (0.301) = 0.301$$

score

→ For the term "the"

$$\therefore \; \overset{\downarrow}{TF(\text{Doc}1)} = \boxed{2}$$

total DOC

$$IDF = \log(\overset{\downarrow}{\cancel{2}}/\cancel{2})$$

$$= \log(1) = 0$$

$$TFIDF = TF \times IDF = 2 \times 0 = 0$$

Words that are common across documents (like "the") have a lower TF-IDF score, while words unique to a document (like "cat") have a higher score.

Application: Information Retrieval

Search Engine

document clustering