Week 6

Monday, June 29, 2020 3:42 PM

Natural language processing

Language has :

- Syntax
- Semantics

Formal Grammar :

A system of rules for generating sentences in a language

Example : context-free Grammar

N-gram :

A contiguous sequence of n items from a sample of text

Examples :

Character n-gram

Word n-gram

Types :

Unigram : $n = 1$

Bigram : $n = 2$

Trigram : $n = 3$

Tokenization :

The task of splitting a sequence of characters into pieces (tokens)

- Word tokenization : …splitting words ………

Text Categorization :

Bag-of-words model :

Model that represents text as an unordered collection of words

Naïve Bayes :

Bayed Rule :

$P(b \mid a) = P(a \mid b) * p(b) / p(a)$

$P(B \mid A)$ proportional to $P(A \mid B) * p(B)$

$P(B \mid A)$ proportional to $P(B, A)$

Naïve : assume that the words are independent from one another

Additive smoothing :

Adding a value (a)to each value in our distribution to smooth the data

Example :

Laplace Smoothing :

Adding 1 to each value in our distribution :

Pretending we have seen each value one more time than we have

----

Information retrieval :

The task of finding relevant documents in response to a user query

Topic Modeling :

Models for discovering the topics for a set of documents

Term frequency :

Number of times a term appears in a document

Function words :

Words that have little meaning on their own, but are used to Grammatically connect other words

Content words :

Words that carry meaning independently

Inverse document frequency :

Measure of how common or rare a word is across documents

Equals to : Log $\frac{TotalDocuments}{documents\ Containing\ (\boldsymbol{word})}$

Tf-idf :

Ranking of what words are important in a document by multiplying term frequency (TF) by inverse document frequency (IDF)

- **Multiply it by the count of word in the document to get the probability**

Information Extraction :

The task of extracting Knowledge from the document

Word representation :

One-hot representation :

Representation of meaning as a vector with a single 1, and with other value as 0

Distribution representation :

Representation of a meaning distributed across multiple values

Example :

Word2vec :

Model for generating word vectors

Skip-gram architecture :

Neural network architecture for predicting context words given a target word