Sindhuja Yerramalla
U00839259

Data Mining

Assignment - 7

Q1.

* Bag-of-words refers to what kind of information you can exctrapt from a document. Vector space model refers to the data structure for each document. Both aspects complement each other

Vector Space Model (VSM)

Given the bag of words that you extracted from the document, you create a feature vector for the document, where each feature is a word and the feature's value is a term weight. The term weight might be:

→ A binary value. (with 1 indicating that the term occured in the document, and 0 indicating that it did not);

→ A term frequency value (indicating how many times the term occured in the document). or

→ A TF-IDF value. (small floating point number like 1.23)

Bag-of-Words :- For a given document, you extract only the terms to create an unordered list of words. No POS tag, no syntax, no semantics, no position, no bigrams, no trigrams. Only the unigram words to represent the document.

Reference :- stackexchange.com.

Q2

Given,

No of words in document = 200
(C)

No of occurences of words apple (Ca) = 4

Total No. of documents (N) = 10,00,000

No. of documents in which the word apple appear (Na) = 100

* Raw term frequency of apple t(f,d) = count of word apple in document

$\boxed{t(f,d) = 4}$

* Inverse document frequency of apple, $IDF(t) = 1 + \log_{10}\left(\frac{\text{Total Docs in collection}}{\text{No. of Docs containing apple}}\right)$

$$= 1 + \log_{10}\left(\frac{N}{Na}\right)$$

$$= 1 + \log_{10}\left(\frac{10,00,000}{100}\right)$$

$$= 1 + \log_{10} 10,000$$

$$= 1 + 4$$

$\boxed{IDF(t) = 5}$

* tf-idf weight of apple w(t,d)

$$= TF(t,d) \times IDF(t)$$

$$= 4 \times 5$$

IDF $\boxed{W(t,d) = 20}$

Q3

Given

Document D = "I like apple and banana".

Vocabulary V = {I, you, she, he, like, dislike, apple, orange, strawberry, banana, and, or}

a) Maximum likelihood of Apples:-

No of occurences of word apple in document = 1 (c apple)

Sum of No of occurences of each word in document

$\sum_{i=1}^{N} c(w_i) = 5.$

$$\text{Probability (apple} \mid D) = \frac{C(\text{apple})}{\sum_{i=1}^{N} c(w_i)} = 1/5$$

$\boxed{P(\text{apple} \mid D) = 1/5 = 0.2}$

b) Probability (apple(D) after laplace smoothing

Count of apple in document $c(w,d) = 1$

laplace smoothing constant $\delta = 1$

Length of document $|d| = 5$

Vocabulary size $\delta(V) = 12.$

$$P(\text{apple} \mid D) = \frac{c(w,d) + \delta}{|d| + \delta|V|} = \frac{1+1}{5+12}$$

$\boxed{P(\text{apple} \mid D) = 0.12}$

Q4

Need for Smoothening in statistical Model:-

* Smoothing is mainly used to eliminate unseen events.

* Consider a N-gram model. If a word is not seen in training data then it's probability becomes zero.

* This means that future documents should not contain that word. But this ~~will~~ is not what it is required

* To make sure that probability of word doesn't become zero, we use smoothing.

* By using smoothing probability of word never becomes zero

* Smoothing assigns non-zero probabilities to words that are not seen.

Smoothing methods

* Additive smoothing → Adds constant δ to count of each word.

Reference :- Senior's Assignment (Past Semester).