V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2V2

National University of Computer & Emerging Sciences
FAST-Karachi Campus
CS4051- Information Retrieval
Quiz#4 (Double Weightage)

Dated: April 26, 2023 Marks: 20
Time: 20 min. Std-ID: _____Sol______

Question No.1

Consider the following examples for the task of text classification

Dataset	**DocID**	**Features- Words in documents**	**Class Fruit=Yes/No**
Training set	1	Orange, Orange, Lemon, Red	No
	2	Orange, Red, Blue, Yellow	No
	3	Apricot, Apple, Mango	Yes
	4	Apple, Banana , Orange	Yes
	5	Blue, Orange, Yellow	No
Test set	6	Orange, Mango, Melon	?
	7	Orange, Red, Lemon, Yellow	?

a. Using the training data calculate the class prior probabilities?

Prior of Class Fruit=Yes = 2/5
Prior of Class Fruit = No = 3/5

b. Using Multinomial Naïve Bayes to estimate the probabilities of each term (feature) that that are given in the problem.

P(Orange/Fruit=Yes) = (1+1)/ (6+10) =1/8
P(Mango/Fruit=Yes) = (1+1)/(6+10)=1/8
P(Red/Fruit=Yes) = (0+1)/(6+10)= 1/16
P(Lemon/Fruit=Yes) = 1/16
P(Yellow/Fruit=Yes) = 1/16

P(Orange/Fruit=No) = (4+1)/(11+10)=5/21
P(Mango/Fruit=No) = (0+1)/(11+10)=1/21
P(Red/Fruit=No) = (2+1)/(11+10)=1/7
P(Lemon/Fruit=No) = (1+1)/(11+10)=2/21
P(Yellow/Fruit=No) = (2+1)/(11+10)=1/7

c. Predict the class labels for the two instances in test set?

P (D6/ Fruit=Yes) = P(Fruit=Yes) * P (Orange/ Fruit=Yes) * P (Mango/ Fruit=Yes) * P (Melon/ Fruit=Yes)

P (D6/ Fruit=Yes) = 2/5* 1/8* 1/8* P (Melon/ Fruit=Yes)

As Melon is out of vocabulary term so we can assign it a possible probability of token that is 1/11

So, P (D6/ Fruit=Yes) = 2/5* 1/8* 1/8* 1/11 = 0.00056

Similarly, for P (D6/ Fruit=No) = 3/5* 5/21* 1/21* 1/11 = 0.00061

D6 belong to class Fruit=No

P (D7/ Fruit=Yes) = 2/5 * 1/8* 1/16 * 1/16 * 1/16 = 0.000012

P (D7/ Fruit=No) = 3/5 * 5/21* 1/7 * 2/21 * 1/7 = 0.000227

D6 belong to class Fruit=No

Question No. 2

What is Word Embedding? How embedding's are learned? Illustrate what kind of text feature it fails to capture in representation? **[5]**

In Natural Language Processing (NLP), a word embedding is a representation of a word. This representation is learned using Neural Network and an individual word is represented as real-valued vectors in a predefined fixed length of vocabulary used to represent the similar words. For example, consider if we want to represents <man, woman, boy, girl, prince, princess, queen, king>, we can use a fixed set of embedding for the terms like <femininity, youth, royalty> so the given words are represented as follow:

Words /Embedding's	femininity	youth	royalty
man	0	0	0
woman	1	0	0
boy	0	1	0
girl	1	1	0
prince	0	1	1
princess	1	1	1
queen	1	0	1
king	0	0	1

Word-Embedding's are learned by using a Neural Network, the task is to learn similar word embedding for words that appear many times in similar contexts by guessing missing words in a huge corpus of text sentences. There are two broad categoies of architecture that can be used to learn embedding's (i) Bag-of-Word model and (ii) Skip-gram Model

From the application prospective – Embedding are dense – low dimensional vector that enable efficient computation of semantic based processing. Relationship between words are captured and maintain.

Word Embedding's are unable to handle out of training vocabulary – terms that not part of the training data. Moreover, words with multiple meanings are conflated into a single representation (a single vector in the semantic space). In other words, polysemy and homonymy are not handled properly.

Question No. 3

How Neural Information Retrieval different from traditional Information Retrieval? Outline some of the benefits of Neural IR? [5]

Traditional IR model are based on assumptions about common representation scheme for both document and query. The retrieval process is very much dependent on these assumptions and does not support contextual retrieval based on semantics of the natural language text generally used in documents.

Neural Information Retrieval (NeurIR) is based on Neural Network. The representation is autonomously learned as per task setting. NeurIR model consists of at least one input and one output layers and possibly zero or more hidden layers with different numbers of neurons in each layers. The connection and weight of these neurons automatically set to optimized through supervised or unsupervised learning for the retrieval task. Hence it is more generalized model.

Benefits:

- Context based word representation can be autonomously learning for any NLP specific task like Ad hoc Information Retrieval.
- More generalized model and It can be quite robust to support long/complex queries