V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1

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

What is Word Embedding? Give one advantage and one disadvantage of using word embedding in an application. **[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

One of the major advantage of this dense representation of low-dimensional vector that the representation is more generalized and computationally efficient to process.

One of the main limitations of static word embedding's or word vector space models is that they failed to represents out-of-vocabulary term for example a term that is not available on training time when these representations are learned. 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.2

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

Question No.3

Consider the following examples for the task of text classification [5+5]

	docID	words in document	in c = *China*?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

a. Using feature vectors as term frequency and the k-Nearest Neighbors (KNN) with k=3 identify the class of test instance docID=5?

Consider the features

V= <japan; macao; osaka; sapporo; shanghai; taipei; taiwan>

D1= <0,0,0,0,0,1,1> magnitude |D1| = √2
D2=<0,1,0,0,1,0,1> magnitude |D1| = √3

D3=<1,0,0,1,0,0,0> magnitude |D1| = √2
D4=<0,0,1,1,0,0,1> magnitude |D1| = √3
D5=<0,0,0,1,0,0,2> magnitude |D1| = √5

Now distance as (dot-prouct) with D5 for each document can be given as

Dot-Product (D1,D5) = 2
Dot-Product (D2,D5) = 2
Dot-Product (D3,D5) = 1
Dot-Product (D4,D5) = 3

For K=3 the three neighbors of D5 will be D1, D2 and D3 from these D1 and D2 majority belong to class c=china is YES

b. Using the same feature vectors and the Rocchio's algorithm, classify the test instance docID=5?

Consider the two mass vectors of the classes

μ(class=china=Yes) = ½ (D1+D2) = ½ { <0,0,0,0,0,1,1> + <0,1,0,0,1,0,1> }

= <0,1/2,0,0,1/2,1/2,1>

μ(class=china=No) = ½ (D3+D4) = ½ { <1,0,0,1,0,0,0> + <0,0,1,1,0,0,1>}

= <1/2,0,1/2, 1,0,0,1/2>

Now angle between vectors from μ(class=china=Yes) to D5 can be computed as using:

angle (μ(class=china=Yes), D5) = 0.8283

angle (μ(class=china=No), D5) = 0.8283

As the D5 has the same angle from the two center mass Rocchio's cannot be able to decide about this example.