Ref. No. : Ex/IT/PE/H/T/423B/2023

B.E. INFORMATION TECHNOLOGY FOURTH YEAR SECOND SEMESTEREXAM 2023

NLP AND TEXT MINING (HONS,)

Time: 3 hours Full Marks: 100

ANSWER ALL QUESTIONS

1. Answer the following questions.

(i) [CO1] When you create a password on a website often there are constraints on the nature of the password.
Suppose the constraints on the password are:
a) Password should be at least 8 characters long.
b) It must have at least one special character from !%@*.
c) It must have some letters and some digits - that is at least one letter and one digit.
d) It should not have characters not listed above.

Assume you had a function called match(regexp,str) that returns True or False depending on whetherthe regular expression regexp matches the string str or not.

Write an expression using match, egrep/any other regular expression language and boolean operations like OR, AND, NOT that will return True if the password satisfies the given constraints andFalse otherwise.

[Hint: egrep has defined character classes: [[:alnum:]] - all alpha-numeric characters, [[:digit:]] -all digits, [[:upper:]] - all upper case characters. Use the character classes to express your answerconcisely.]

12

(ii) [CO1] Design a finite state transducer with E-insertion orthographic rule that parses from surface level “foxes” to lexical level “fox+N+PL” using FST.

8

2. Answer the following questions.

(i) [CO2] consider only 3 POS tags that are noun, model and verb. Assume that the training data is as follows.

Mary Jane can see Will
Spot will see Mary
Will Jane spot Mary?
Mary will pat Spot

Create a table and fill it with the co-occurrence counts of the tags.

Let the test sentence be 'Will can spot Mary'.

Calculate the probability of this sequence be <s>, M, V, N, N, </s>.

Show the run of Viterbi algorithm on this test sentence.

15

(ii) [CO2] For a news corpus, the number of unique words were found to be 200,000. The number of unique bigrams were found to be 4,000,000, out of which 50% occurred only once, while 20% occurred twice. Suppose you are using Good-Turing smoothing. Estimate the effective bigram count for c = 0 and c = 1 using Good-Turing smoothing, where c denotes the count observed in the news corpus.

5

3. [CO3] Answer the following questions:

a. In a corpus of 10000 documents you randomly pick a document, say D, which has a total of 250 words and the word 'data' occurs 20 times. Also, the word 'data' occurs in 2500 (out of 10000) documents. What will be the tf-idf entry for the term 'data' in a bag of words vector representation for D?

b. Suppose you have the following two 4-dimensional word vectors for two words w_1 and w_2 respectively: w_1 = (0.2, 0.1, 0.3, 0.4) and w_2 = (0.3, 0, 0.2, 0.5) What is the cosine similarity between w_1 and w_2? Are the words w_1 and w_2 similar or dissimilar?

c. In word2vec (skipgram) we compute the predicted probability distribution vector ($\hat{y}$) on the vocabulary via a softmax over the output vector z i.e.,y = softmax(z). Given that the desired output is y and the error function E is the cross entropy function $E = \Sigma_i -y_i \ln(y_i)$ derive the gradient for the first step in the backpropagation.

d. What is the IDF of a term that occurs in every document? Justify

e. Explain the significance of negative sampling in Word2Vec.

3+3+7+3+4

4. [CO4] Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in the following Table.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

(i) Compute the tf-idf weights for the terms car, auto, insurance, best, for each document, using the df_t values from the following table.

	df_t
car	18165
auto	6723
insurance	19,241
best	25235

N = 806,791

(ii) Compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms.

(iii) Rank the three documents by computed score for the query car insurance, for each of the following cases of term weighting in the query:
a) The weight of a term is 1 if present in the query, 0 otherwise.
b) Euclidean normalized idf

7+5+(4x2)

5. Answer the following questions:

i. [CO5] An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

3

B) [CO5] Consider the following matrix for a classifier for spam filtering.

	Reality:1 (Really a spam)	Reaity:0 (Not a spam)
Prediction:1 (Predicted as a spam)	10	55
Prediction:0 (Predicted as a non-spam)	10	25

a) Calculate accuracy, precision, recall, F1-score.
b) Which measure should be improved in this case for a better classifier?

3+2

C) Discuss one method by which we can say that one classifier is better than another classifier for a given metric.

12