Ref. No. : Ex/IT/PE/H/T/423B/2022

B.E. INFORMATION TECHNOLOGY FOURTH YEAR SECOND SEMESTER – 2022

NLP AND TEXT MINING (HONS.)

Time: 4 hours | Full Marks: 70

1. [CO1] Answer the following questions:
   a. Explain Levensthein's minimum edit distance algorithm.
   b. Compute minimum edit distance between "paceful" and "peaceful".

   6+6

2. [CO2]Answer the following questions:
   a. You are given the following data: vocabulary $V = \{w_1, w_2, w_3\}$ and the bigram probability distribution p on $V \times V$ given by: $p(w_1, w_1) = 0.25$, $p(w_2, w_2) = 0.0$, $p(w_3, w_3) = 0.25$, $p(w_2, w_1) = 0.125$, $p(w_1, w_3) = 0.25$, $p(w_1, *) = 0.5$ (that is w_1 as the first of a pair), $p(*, w_2) = 0.125$. Compute $p(w_1, w_2)$ and $p(w_2 \mid w_3)$.

   b. If the only choice you had is between a 3-gram and a 4-gram language model which was based on some corpus (say the Brown corpus) which will you prefer and why?

   5+5

3. [CO1] Answer the following questions:
   a. Consider the regular expression below which is supposed to describe some morphological facts about English adjectives (disregard any spelling changes that happen and interpret "*" as zero or once)
   (un-)* (adj-root)(-er| -est| -ly)*
   We see it works for an adjective like happy. What is the problem with the proposed regular expression model? And how will you go about correcting it - explain your approach?
   b. In a statistical named entity recognition (NER) system using the IOB (in, out, begin) labelling approach a model is learned using a suitably tagged corpus. What features are/can typically be used in tagging words and what are the labels (assume we use only 3 NER types: proper names, locations, organizations)? For the following fragment of some corpus indicate the labels (do not give the features). Note: natural phenomena like cyclones/typhoons/comets are often given names.
   Cyclone Trump in the Bay of Bengal moved towards Chennai.

   6+6

4. [CO2] Explain Laplace Smoothing. What is the problem with Laplace smoothing

7+7

5. [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 tfidf 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., $y = \text{softmax}(z)$. 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 back propagation.

3+3+7

6. [CO4] Name two types of summarization. Explain their differences. Draw a workflow for one type of text summarizer.

1+4+4