Y/D/TY

Reg. No:

Final Assessment Test – April 2025

VIT Vellore Institute of Technology

Course: CSI4001 . Natural Language Processing and Computational Linguistics

Class NBR(s): 2122/2130 Slot: C1+TC1

Time: Three Hours Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

Answer ALL Questions

(10 X 10 = 100 Marks)

1. a) Assume you are the CTO (Chief Technology Officer) of an NLP Company. You have a task at hand of analysing reviews of a Logistics service. You are handed over a GB of text data – all are logistics service reviews. Explain precisely on what kind of NLP stage-wise outcomes and insights can you determine from this data? [7]

   b) Design an NLP pipeline for text processing involved in the above Question No. 1 a). using NLTK (Natural Language Tool Kit). Augment your NLTK code with comments. [3]

2. a) Given below are some excerpts from Tweets.txt dataset: [5]

   Im reading the Wall Street Journal.
   Just tryin' out twitter, thanks for asking
   OMG, having fun with work:-)
   Im wasting time setting up twitter 2010 on my mobile phone when I ought to be working instead ☹
   Working on some invoices rather than my Ruby on Rails project!!!
   Just finished cheese pizza and cola now back to worrying...
   Goodnight Helsinki, coordinating a workshop about fore sighting on 20.04.2007 ☺

   Disambiguate the punctuations and contractions wherever applicable in the above text. Present your analysis of emojis in text processing. Do you have to retain them or remove them? Why?

   b) *Characters from different languages are encoded with minimum overlapping* - Discuss this point with the motivation and history of encoding, right from the 7-bit ASCII to the recent UNICODE encodings. [5]

3. a) Determine the minimum edits required between the words ***'tutor'*** and ***'rumor'***. Consider uniform cost for the update, insert and delete operations. Apply backtracking to find the final cost of edits. [5]

   b) Assume that your company has won a contract to develop the Corpus of a Tribal language. What method you will employ to gather this data? What are the parameters and copy-right clauses that should be considered in creating a corpus? Explain with necessary corpus protocols and examples as applicable. [5]

4. Bigram counts for 6 words from a Speech corpus is given below:

	How	good	is	oriental	cuisine	here
How	6	1027	0	9	0	0
is	2	0	608	1	6	6
good	2	0	4	686	2	0
continental	0	0	2	0	16	2
cuisine	1	0	0	0	0	82
here	15	0	15	0	1	4

Calculate the probability that implies the prediction of the following sentence:

How good is oriental cuisine here?

Use Laplacian smoothing technique and determine the Reconstituted counts of the words and Bigram Probabilities in a better probability space.

Unigram counts of each of the words:

How – 3437; *good* – 1215; is – 3256; oriental – 938; *cuisine* – 213; *here* – 1506.

5. Design an FSA for Adjectives in English demonstrating Derivational Morphology. Provide a supporting lexicon and show the transitions for words in the lexicon. Identify the exceptions, if any that could not be addressed by the FSA.

6. An excerpt of an article published in the Journal of American Medical Informatics Association in given below:

Ferraro et al in their research work reported that "POS tagging is an important syntactic process whose performance can greatly affect subsequent downstream processes such as syntactic parsing and semantic inference. We demonstrate that through domain adaptation, we can reduce the residual error in POS tagging in a cost-effective manner leveraging current out-of-domain algorithms with a modest amount of in-domain (clinical) annotated data. We confirm that clinical narratives have different linguistics characteristics than those of general English and biomedical texts. We show that state-of-the-art POS taggers with accuracies upward of 97% quickly drop to accuracies in the 80%s when applied to clinical narratives using their general English or biomedical source domain models."

(i) Comment on the observation by the authors. **[3]**

(ii) Design a HMM PoS Tagger and present the mathematical model. **[4]**

(iii) What should be taken care to address the authors' concerns as determined in (i). **[3]**

7. Gathering Emission Probability is sometimes seen as a drawback while performing PoS Tagging. How is Maximum Entropy a better option in such cases?

Perform the Part-of-Tagging using the Maximum Entropy approach for the sentence:

An inspirational sunset

Use the following Features derived from Language Rules and Neighbourhood properties:

Features	DET	NOUN	VERB	ADJ	VERB	NOUN
F1: Ti-1 = DET, Ti = ADJ				1		
F2: Ti-1= NOUN, Ti=VERB			1			
F3: Ti-1 = ADJ, Ti = NOUN						1
F4: Wi-1 = AN, Ti = ADJ				1		
F5: WI-1 =AN, Wi+1 = SUNSET, Ti = ADJ				1		
F6: Wi-1 = inspirational, Ti= NOUN						1
F7: Wi+1 = inspirational, Ti = DET	1					
F8: Wi-1 = NULL, Ti = NOUN		1				

8. Assume the following corpus:

Tesseract is an optical character recognition engine for various operating systems. It is a free software, released under the Apache License. Originally developed by Hewlett-Packard as proprietary software in the 1980s, it was released as open source in 2005 and development was sponsored by Google in 2006. Tessaract is good optical character recognition engine. Optical character recognition is significant in many applications.

Design an NLP pipeline to find the similarity between the following two sentences S1 and S2 based on the corpus given above:

S1: ***Google Cloud Vision is a character recognition engine.***
S2: ***OCR is an optical character recognition engine***

Include notes on the similarity metric used and mention its pros and cons.

9.a) Elucidate on how Random walk is used for knowledge-based Word Sense Disambiguation.

OR

9.b) Using Lesk's Algorithm design the Word Sense Disambiguation technique for the ambiguous words in the following two sentences:
The barn was reduced to ashes in forest fire.
The study table is made of ash wood.

10.a) Explain a Case study on Information Extraction and its application in Text Summarization.

OR

10.b) Develop a Sentiment Analysis architecture using a Machine Learning Technique. Augment your answer with python code and comments on Importing Data, transformation to feature vectors, using transformers/vectorisation techniques, tokenization, classification and accuracy metric.

⇔⇔⇔ **Y/D/TY** ⇔⇔⇔