MACHAKOS UNIVERSITY

University Examinations for 2021/2022

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (COMPUTER SCIENCE)

SCO 409: NATURAL LANGUAGE PROCESSING

DATE: 14/12/2021 **TIME: 8.30-10.30 AM**

INSTRUCTIONS:

This paper consists of FIVE questions
*Answer **question one** and other **two questions** in this paper*

QUESTION ONE (30 MARKS)

a) Distinguish between Anaphoric and Pragmatic Ambiguity (4 marks)

b) Explain TWO factors that define Corpus Representativeness (4 marks)

c) Verbnet is organized into TWO verb classes extending Levin classes by refinement and addition of subclasses. Discuss what each of the TWO classes contain. (4 marks)

d) Discuss TWO Properties of Stochastic POST Tagging (4 marks)

e) Consider the following Python expression: *len(set(text4)).*

i. State the purpose of this expression. (2 marks)

ii. Explain the TWO steps involved in performing this computation. (2 marks)

f) Write a function that finds the 50 most frequently occurring words of a text that are not stop word (5 marks)

g) Explain FIVE applications of Natural Language Processing (5 marks)

QUESTION TWO (20 MARKS)

a) "Kim ran up a hill by the steepest path. Sandy had egged her on."

This narrative exemplifies some problems for natural language processing systems.

i. Identify these problems. (6 marks)

ii. Describe the stages of NLP that would be required to interpret the narrative. (6 marks)

iii. Evaluate how successful each stage is likely to be. (4 marks)

b) Write a function hedge(text) that processes a text and produces a new version with the word 'like' between every third word. (4 marks)

QUESTION THREE (20 MARKS)

a) Describe the TWO most common types of Treebanks in linguistics. (4 marks)

b) Explain the THREE elements of Semantic Analysis (6 marks)

c) Given a string *s = 'colorless'*. Write a Python statement that changes string s to "*colourless*" using only the slice and concatenation operations. (5 marks)

d) Describe the two-stage architecture of Rule-based POS tagging (5 marks)

QUESTION FOUR (20 MARKS)

a) Explain the FOUR components of Contex Free Grammar (4 marks)

b) Outline FOUR advantages of Transformation-based Learning (TBL) (4 marks)

c) Write regular expressions to match the following classes of strings:

i. A single determiner (assume that a, an, and the are the only determiners) (3 marks)

ii. An arithmetic expression using integers, addition, and multiplication, such as 2*3+8 (4 marks)

d) Given a trained classifier, how could you set up an experiment to evaluate its performance on some new text? (5 marks)

QUESTION FIVE (20 MARKS)

a) Use the following text to derive distributions for rat and chased.

The cat chased a rat. A big rat chased the big dog.

Use a five-word window, including open- and closed- class words, ignore case, punctuation and sentence boundaries and weight contexts by frequency. (5 marks)

i) Show unlabeled dependency links for each sentence and give distributions for rat and chased using contexts derived from single dependency links. (5 marks)

b) Discuss THREE design features of Information retrieval (IR) systems (6 marks)

c) Using an example, write a python code to define a parser (4 marks)