MACHAKOS UNIVERSITY

University Examinations for 2019/2020 Academic Year

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (MATHEMATICS AND COMPUTER SCIENCE)

SCO 409: NATURAL LANGUAGE PROCESSING

DATE: 27/10/2020 **TIME: 8.30-10.30 AM**

INSTRUCTIONS

Answer question ONE and any other TWO questions.

QUESTION ONE (30 MARKS)

a) Define Natural Language Processing (2 marks)

b) The history of NLP is divided into four distinct phases, state and explain. (8 marks)

c) Ambiguity, generally used in natural language processing, can be referred as the ability of being understood in more than one way. Explain the different types of ambiguities found in NLP (5 marks)

d) State and explain the NLP phases (8 marks)

e) Natural Language Processing basically can be classified into two parts i.e. Natural Language Understanding and Natural Language Generation which evolves the task to understand and generate the text. Expound on the statement using NLU and NLG (7 marks)

QUESTION TWO (20 MARKS)

a) Language is infinite but a corpus has to be finite in size. For the corpus to be finite in size, we need to sample and proportionally include a wide range of text types to ensure a good corpus design. What is corpus and corpora in light of the underlined statement. (6 marks)

b) In constructing a corpus, concern is on the overall design: for example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples. Each of these involves a sampling decision, either conscious or not. While obtaining a representative sample, we need to consider the following: Sampling unit, Sampling frame and Population. Describe what is a Sampling unit, Sampling frame and Population (6 marks)

c) A regular expression (RE) is a language for specifying text search strings**.** Outline the properties of regular expressions (8 marks)

QUESTION THREE (20 MARKS)

a) Explain the relation between Finite Automata, Regular Grammars and Regular Expressions (6 marks)

b) At the lexical level, Semantic representations can be replaced by the words that have one meaning. In NLP system, the nature of the representation varies according to the semantic theory deployed. Expound (4 marks)

c) Natural Language Processing can be applied into various areas. List and explain (10 marks)

QUESTION FOUR (20 MARKS)

a) The regular expression language is a powerful tool for pattern-matching. Outline the basic operations in regular expressions (6 marks)

b) Word tokenization and normalization are generally done by cascades of simple regular expressions substitutions or finite automata. Expound (6 marks)

c) What is the Porter algorithm and how does it help in stemming and stripping off affixes? (8 marks)

QUESTION FIVE (20 MARKS)

Syntactic information about the text can be important to assist in resolving ambiguities and in establishing the appropriate relations among the words in a text. At the most basic level, determining whether a word is a noun or a verb (or some other part of speech) can be useful. This is accomplished through tools that perform **part of speech (POS) tagging**. Then, identification of phrases in the text can be important, such as recognizing that a sequence of words forms a single conceptual unit (e.g. breast cancer, NF kappa beta inhibitor). A commonly used strategy for this is **shallow parsing,** which involves identifying coarse phrasal structures, such as noun phrases, without identifying the specific grammatical relationships among them. In contrast, **Deep parsing** determines the full set of grammatical relations among words in a sentence, producing a complete parse tree to represent these relations.

Expound further using concrete examples what is speech tagging , shallow and deep parsing