KENYATTA UNIVERSITY

EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE

(COMPUTER SCIENCE)

SCO 409: NATURAL LANGUAGE PROCESSING

2ND SEMESTER 2023/2024 TIME: 2HOURS

INSTRUCTIONS: Attempt question ONE and ANY other TWO. (Total marks 70)

QUESTION 1 (30 MARKS)

A) Distinguish between stochastic grammars and deterministic grammars (5 Marks)

B) Explain the technical role of each of the following tasks in NLP (5 Marks)

i. Stemming
ii. Lemmatization
iii. Tokenization
iv. Sentiment Analysis
v. Word Sense Disambiguation
vi. Part of Speech (POS) Tagging

C) Justify the significance of machine translation (4 Marks)

D) Demonstrate the use of TFIDF in vectorization of textual data for carrying out NLP tasks. Support your answer with appropriate equations / formulae (6 Marks)

E) Distinguish between data standardization and data normalization techniques used in text preprocessing in NLP. Provide an example in each case. (6 Marks)

F) Describe the two main Python libraries / packages for NLP tasks (4 Marks)

QUESTION 2 (20 MARKS)

A) Discuss any four challenges of NLP in spite of advancements in this field (12 Marks)

B) Evaluate Word2Vec model of representing text in NLP projects. Be sure to include its extensions and limitations / limits (8 Marks)

QUESTION 3 (20 MARKS)

A) Briefly discuss any four applications of NLP (12 Marks)

B) Describe speech recognition technology, indicating how the technology is (could be) used in supporting business functions of large corporations. In addition provide examples of popular applications of this technology (8 Marks)

QUESTION 4 (20 MARKS)

A) Language is infinite but a corpus has a finite 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. Evaluate three important elements for corpus design (9 Marks)

B) Write a Python code snippet for carrying out Named Entity Recognition using the sample text below (11 Marks)

```
text = "Robert Downey Jr. is an American actor and producer.
He is best known for his roles in films such as Iron Man, The
Avengers, and Sherlock Holmes. Downey has won several awards
for his acting, including two Screen Actors Guild Awards and a
Golden Globe Award. He has also been nominated for an Academy
Award."
```

The expected output should as shown below

```
Robert Downey Jr.: PERSON
Iron Man: PERSON
Avengers: ORG
Sherlock Holmes: PERSON
Downey: ORG
Screen Actors Guild Awards: ORG
```

QUESTION FIVE (20 MARKS)

A) Distinguish between Deep Parsing and Shallow Parsing as used in NLP (8 Marks)

B) Describe the process of evaluating two language models A and B. Assume you are carrying out an extrinsic evaluation of a N-gram language model. (12 Marks)

END