



Natural Language Processing COS4861

2024

Assignment 1

School of Computing

Lecturer: Dr M. Sibiya

This assignment consists of 5 questions that give a total of 100 marks.



Define tomorrow.

UNISA | 
university
of south africa

Important: This assignment will serve as your compass in navigating the intricate landscape of computation theory as it intersects with language. By immersing yourself in the study of regular expressions, you will harness their formidable capabilities to describe and manipulate patterns within textual data. Armed with this proficiency, you will then embark on a deeper exploration of automata, unraveling their inner workings and gaining insight into how computers process formal languages. Through this journey, you will not only acquire theoretical knowledge but also cultivate practical skills essential for proficient language processing in the realm of computer science.

Question 1

1. In Table 1 below underline single characters in the column “Pattern Matched” that are matched by the regular expressions in the column RE:

Table 1: Matching single letters using regular expressions.

RE	Pattern Matched
(i) /[A-Z]/	“peeew!! Running 5km in 30 minutes is not a child’s play.”
(ii) /[a-z]/	“1 2 3 4. Please do not tell me that you have not started learning from the NLP videos that were posted by Dr Sibiya.”
(iii) /[0-9]/	“Remove a 0 from the cellphone number and add a 1 instead.”
(iv) [^St]	“Students who use ChatGPT to answer the UNISA assignments of NLP are likely to fail the exams.”
(v) /[^a-z]/	“In 2025 UNISA will be launching an AI lab and plans to bring in a new qualification called MSc in Artificial Intelligence.”

1 Mark for each correct match (5)

2. In the context of computational language, how would you describe a formal language? (3)
3. What is the significant difference between a deterministic finite state automata (DFSA) and non-deterministic finite state automata (NFSA). Also use a diagram to explain your answer. (2)
4. Draw a complete Chomsky hierarchy and label its components ranging from the automaton involved, to languages and grammar. Briefly describe the expressive power of all the automaton involved in this hierarchy. (15)

Question 2

1. If $w = \text{'TOUCH'}$ is a word, derive its substrings including 'itself' and epsilon while excluding single character substrings of w . (11)
2. In the above question give the trivial substrings of w . (2)
3. If the alphabet, $\Sigma = \{0,1\}$ then compute: (7)
 \sum^0 , \sum^1 and \sum^2 .
4. Write the Kleene closure and positive closure of the above computations. (10)

Question 3

1. Assume that UNISA researchers developed a meaningful language to communicate with the bees as follows:
 $czz!$
 $czzz!$
 $czzzz!$
 $czzzzz!$
 $czzzzzz!$

If the strings are infinite, draw a model of the finite state machine for the above language using 5 states.

(10)

Question 4

1. Draw a model of a DFSA with the following state transitions:

$$\delta(q_0, 0) = q_0$$

$$\delta(q_0, 1) = q_1$$

$$\delta(q_1, 0) = q_1$$

$$\delta(q_1, 1) = q_2$$

$$\delta(q_2, 0) = q_1$$

$$\delta(q_2, 1) = q_2$$

(15)

Question 5

1. If $w_1=1010$ is a word, demonstrate all the steps whether the DFSA in Question 4 accepts the word w_1 . Give the reason for your answer. (15)
2. Give names and symbols of 5 tuples that define the DFSA. (5)

Total 100



Unisa 2024