Cairo University
Faculty of Computers & Information
Natural Language Processing CS462

Mid-term Exam
2018

Student Department:

Student Name:

Student ID:

Marks:

Q1	***Q2***	***Q3***	***Total(15)***

Question 1 (4 Marks)

B is a corpus which only contains one single bit string:

1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0

1.1) Calculate the following bigram probabilities from the corpus **B** using MLE (Maximum Likelihood Estimation). Answer with a ratio ***p/q***, not a floating point number.

(a) **P**(0 | 1)

1 mark

C(10) / C(1) = 5/13

(b) **P**(0 | 0)

1 mark

C(00) / C(0) = 3/8

1.2) Assume a bigram language model created from corpus **B**. For each of the following bit strings, decide if it is more probable that x_1 resp x_2 is 0 or 1.

(c) 1 0 1 0 1 0 1 x_1

1 mark

x_1 = 1 is more probable
since P(0 | 1) = 5/13 < P(1 | 1) = 8/13

(d) 0 1 0 1 0 1 0 x_2

1 mark

x_2 = 1 is more probable
since P(0 | 0) = 3/8 < P(1 | 0) = 4/8

Question 2 (5 Marks)

1) Write regular expressions that recognize the following languages. (3 Marks)

(a) *Any string that contains at least three digits* 1 mark

.*\d.*\d.*\d.*

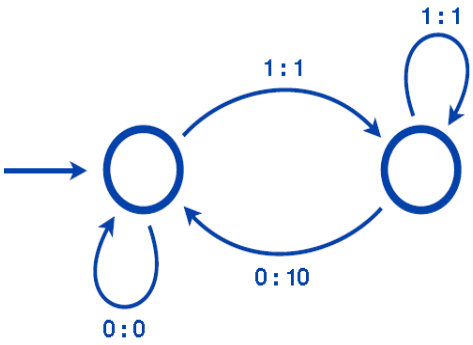
(b) Find a word ending in `ility` , *example* `accessibility` 1 mark

(\w*)ility or [0-9 A-Z a-z]*ility

(c) Any string that starts with one lowercase character, and either ends with two digits or with three vowels 1 mark

[a-z].*(\d\d|[aeiouAEIOU]{3})

2) Draw a finite state transducer from bitstrings to bitstrings, which doubles all 1's that are followed by a 0. This means that it should translate 110010011 to 11100110011, and 11001100 to 1110011100. (2 Marks)



(both states are final)

Question 3 (6 Marks)

1- What are the different types of morphologies that can be considered? Briefly describe the main differences between them.

2 mark

Solution: inflectional morphology: no change in the grammatical category (e.g. give, given, gave, gives) derivational morphology: change in category (e.g. process, processing, processable, processor, processabilty)

2- In the pair (blamed, blame+V+Past), what does "blamed" (resp. "blame+V+ Past") correspond to? What is each of the two forms useful for?

2 mark

They are surface form (i.e. word) and Lexical form (i.e. analysis).

Surface form is useful for NLP interface (input/output).

Lexical form is useful for internal representation, analysis or generation.

3- What is the problem addressed by a Part-of-Speech (PoS) tagger? Why isn't it trivial?

2 mark

The problem addressed by a PoS tagger is to assign part-of-speech tags (i.e. grammatical roles) to words within a given context (sentence, text).

This task is not trivial because of lexical ambiguity (words can have multiple grammatical roles, e.g. can/N can/V) and out-of-vocabulary forms (i.e. unknown words).