5 Formal Models of Language (pjb48)

c, d, e are melodic elements. A birdsong is composed of a sequence of these elements as follows:

$c\ c\ d\ c\ c\ e\ c\ c\ d\ c\ c\ d\ c\ c\ e$

(a) (i) Using byte pair encoding induce a context-free grammar for this sequence of birdsong. Show your workings and state what you have decided to do in the case of a tie. [6 marks]

(ii) Draw the derivation tree that parses the birdsong using your induced grammar. [1 mark]

(iii) What are the shortcomings of this method of grammar induction for natural languages? [3 marks]

(b) Assuming that c, d and e are the only melodic elements available in the birdsong, and that the excerpt we are given is probabilistically representative of the birdsong in the wild, what is the average information produced per element? Provide relevant equations. [3 marks]

(c) In Part (b) we assumed a 1st-order model of the birdsong. What assumption does this make about the sequence of elements? [1 mark]

(d) Consider a 2nd-order model of the birdsong and calculate the conditional entropy. Provide relevant equations. [4 marks]

(e) How can we calculate the entropy rate of birdsong? Provide relevant equations. [2 marks]