A37938

Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Third Year Undergraduate

06-37810

LH Natural Language Processing

Resit Examinations 2023

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 80, which will be rescaled to a mark out of 100.

Question 1

(a) What is a Markov Chain? Illustrate your answer with an example with respect to Natural Language Processing. **[5 marks]**

(b) Propose an area of Natural Language Processing that makes use of Hidden Markov Models. Ensure you fully describe, with appropriate definitions, how Hidden Markov Models are applicable to that area. **[5 marks]**

(c) One of the assumptions made with Hidden Markov Models is the 'bigram' assumption. If this was changed to a 'unigram' assumption, how would this affect the performance of the model? You may want to refer to your previous answer for part (b) to help exemplify this. **[5 marks]**

(d) Suppose the transition and emission probabilities are as exemplified in the following tables:

a_{ij}	s_1 = verb	s_2 = noun	s_3 = adjective	s_f = FINISH
s_0 = START	$a_{01} = 0.6$	$a_{02} = 0.4$	$a_{03} = 0.0$	$a_{0f} = 0.0$
s_1 = verb	$a_{11} = 0.01$	$a_{12} = 0.84$	$a_{13} = 0.0$	$a_{1f} = 0.15$
s_2 = noun	$a_{21} = 0.6$	$a_{22} = 0.1$	$a_2 3 = 0.0$	$a_{2f} = 0.3$
s_3 = adjective	$a_{31} = 0.0$	$a_{32} = 0.79$	$a_{33} = 0.01$	$a_{3f} = 0.2$

Table 1: Transition probability matrix A.

b_{it}	fly
s_1 = verb	$b_1(fly) = 0.8$
s_2 = noun	$b_2(fly) = 0.7$
s_3 = adjective	$b_3(fly) = 0.4$

Table 2: Emission probability matrix B (assume more data, hence probabilities in each row are not 1).

Using the Viterbi algorithm, find the most probable tag sequence that generated the observation sequence "Fly!" and estimate its probability. Show your workings. **[5 marks]**

Question 2

(a) What makes Logistic Regression a discriminative classifier for NLP tasks? Compare this to another type of model used by supervised machine learning classifiers in your answer. **[5 marks]**

(b) Design and give a rationale for your choice of 5 features for a Logistic Regression classifier that would be used for authorship attribution. **[5 marks]**

(c) For a given input to a binary Logistic Regression classifier for Sentiment Classification, the feature vector has the following values: $[3, 2, 1, 3, 0, 4.19]$, the weights are $[-5.0, 2.5, -1.2, 0.5, 2.0, 0.7]$, there is a bias of 0.1 and the class should be '1' if $P(y = 1 \mid x) > 0.5$, else it is '0'. By showing your calculations give the class that the input will be categorized as. **[5 marks]**

(d) The authorship classifier above has an accuracy of 0.95. Discuss the extent to which this is an appropriate method of classifier evaluation for authorship attribution. Ensure you include a comparison to other metrics in your discussion. **[5 marks]**

Question 3

Consider the following sentence:

Julie loves the exciting section of The New York Times.

(a) Given the above sentence, identify all tokens. **[4 marks]**

(b) Given the above sentence, assign the correct part-of-speech tag to each token. **[4 marks]**

(c) Describe three different ways of representing syntactic structure? What are the advantages and disadvantages of each of them? **[6 marks]**

(d) Can the following Context Free Grammar parse the above sentence? N.B the grammar is missing the lexicon part, you should have that if you did part of speech tagging. Punctuation is not part of the grammar. If the grammar can parse the sentence, then draw the parse tree. Otherwise, suggest a grammar that can parse the sentence.

$$
\begin{aligned}
S &\rightarrow NP \mid VP \\
VP &\rightarrow V \mid V\ NP \mid V\ PP \mid V\ PP\ NP \\
PP &\rightarrow P\ NP \\
NP &\rightarrow N \mid Det\ N \mid Adj\ N \mid Det\ Adj\ N
\end{aligned}
$$

[6 marks]

Question 4

Toxicity Detection on Social Media is the task of detecting hate speech, abuse and other toxic language.

Intelligently discuss the components required to build a system for Toxicity Detection on Social Media. This discussion should be no more than 2 pages in length. The discussion should be structured around the NLP pipeline and how this framework could be applied to the task of Toxicity Detection on Social Media. **[20 marks]**

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.