XXday, XX XXX XXXX
Available from XX:XX BST
Expected Duration: 1 hour 30 minutes
Time Allowed: 3 hours
Timed exam within 24 hours

DEGREES of MSci, MEng, BEng, BSc, MA and MA (Social Sciences)

TEXT AS DATA H
COMPSCI 4074

(Answer all 4 questions)

This examination paper is an open book, online assessment and is worth a total of 60 marks.

1. **(a)** (i) Consider three documents, and a total vocabulary of 4, with frequencies as follows: $D1 = [2,2,2,0]$, $D2 = [0,3,0,3]$, $D3 = [1,1,1,1]$ Calculate the cosine similarity between each pair of documents and find which pair of documents is most similar.

[4]

(ii) Using log with base 2 within the IDF function, calculate the IDF of the following words. Assume that there are 4096 documents in the collection.

Word	Document Frequency
the	4096
theory	256
theology	8

[2]

(iii) Thinking about what these IDF values mean, which of these three words is most discriminative.

[1]

(b) (i) Given the table of collection counts below, provide the definition of the unigram probability, P(w). Complete the table with the probability for each term in the collection below.

w	count	P(w)
euro	75	
championship	50	
scotland	25	
wembley	20	
ronaldo	10	
referee	10	
goal	5	
defeat	5	

[3]

(ii) For $D1$ and $D2$ below, what is the KL-Divergence, $KL(D1||D2)$, without smoothing? How did you recognise this without doing any calculations?

D1

t	P(t—D1)
euro	0.2
championship	0.05

D2

t	P(t—D2)
wembley	0.1
goal	0.2
championship	0.08

[2]

(iii) What is the value for $KL(D1||D2)$ using Jelenik-Mercer smoothing with $\lambda = 0.7$ for the calculation. Use the collection probabilities from (i) for the background and the document probabilities from (ii). Show your workings (you do not need evaluate your arithmetic).

[3]

2. (a) In the table below are unigram and bigram counts taken from a corpus of 2400 documents. Use these values with a bigram model to calculate the probability of the pangram: ”the five boxing wizards jump quickly”. Show the probabilities of each bigram as a fraction then calculate the combined probability of the sentence.

unigram	unigram count		bigram	bigram count	P(x_i|x_{i-1})
			$\langle s \rangle$ the	1096	
the	3147		the five	227	
five	821		five boxing	17	
boxing	536		boxing wizards	1	
wizards	7		wizards jump	3	
jump	692		jump quickly	420	
quickly	587		quickly $\langle e \rangle$	500	

[5]

(b) A Logistic Regression classifier for a binary classification problem has an accuracy of 49% on the training data, 47% on the validation data, and 51% on the test data. A dummy classifier has an accuracy of 50% on all three data subsets. Describe the model ‘fit’. Suggest a way to improve the fit and identify a further problem that could arise from improper application of your approach.

[5]

(c) You are analyzing a large collection of textual descriptions of cars. Your task is to automatically discover non-overlapping groups of cars from the data using clustering.

(i) Name and describe an algorithm you would use to perform this task and explain why it is suitable.

[2]

(ii) Name and describe a method for automatically selecting the value of *K*, the number of clusters.

[3]

3. **(a)** You are working as a researcher and discussing embeddings with your colleague. For an upcoming project, your colleague argues that transforming the values of an embedding to be one-hot encoded would be a good representation of its semantics and much better in terms of computational efficiency.

Do you agree with your colleague's statement? Please motivate your answer.

[5]

(b) In a Bavarian attic, a series of tapes have been found containing old speeches from the famous WW2 dictator. An attempt to transcribe them was made, however the quality is so deteriorated that there are missing words here and there. As we know there is a full catalog of already transcribed speeches and our engineers are wondering how we could utilise those in an attempt to fill the missing gaps of the newly found material.

(i) Describe the process for pre-training BERT on previously transcribed speeches. Briefly discuss any considerations and training objectives. [5]

(ii) There is a suspicion that some of these speeches may not be from the dictator himself, but rather some "imitator". Describe how you would use BERT to detect the impostor.

[5]

4. (a) Using your knowledge of Part of Speech (PoS) Tagging by means of HMMs, build emission and transition tables from the training sentences S1 and S2. Then utilise the Viterbi algorithm to figure out the most likely PoS tagging sequence for test sentence T1.

To simplify your computations smoothing is not required. I.e. let frequencies of 0 result in probability of 0.

Training sentences: [

S1: ⟨s⟩ May (NN) Is (VB) Tomorrow (NN) ⟨e⟩
S2: ⟨s⟩ Tomorrow (NN) May (MD) Rain (VB) ⟨e⟩

]

Testing sentence: T1: ⟨s⟩ May Rain Tomorrow ⟨e⟩

(i) Build a HMM emissions table [5]

(ii) Build a HMM transitions table [6]

(iii) Compute the most likely PoS tagging. Show your workings and the result of your computation for the most likely sequence. [4]