University of Glasgow

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]

Solution:

$cos(\theta_{A,B}) = \frac{A \cdot B}{||A|| \times ||B||}$

$||D1|| = \sqrt{12} = 2\sqrt{3}$

$||D2|| = \sqrt{18} = 3\sqrt{2}$

$||D3|| = \sqrt{4} = 2$

$D1 \cdot D2 = 6$

$D2 \cdot D3 = 6$

$D1 \cdot D3 = 6$

$cos(\theta_{1,2}) = \frac{D1 \cdot D2}{||D1|| \times ||D2||} = \frac{6}{\sqrt{12}\sqrt{18}} = \frac{1}{\sqrt{6}} = 0.4082$

$cos(\theta_{2,3}) = \frac{D2 \cdot D3}{||D2|| \times ||D3||} = \frac{6}{\sqrt{18} \times 2} = \frac{3}{3\sqrt{2}} = \frac{1}{\sqrt{2}} = 0.7071$

$cos(\theta_{1,3}) = \frac{D1 \cdot D3}{||D1|| \times ||D3||} = \frac{6}{\sqrt{12} \times 2} = \frac{3}{2\sqrt{3}} = \frac{3\sqrt{3}}{6} = \frac{\sqrt{3}}{2} = 0.8660$

The most similar pair of documents is D1 & D3

[1] each for correct cos for each pair

[1] for identifying most similar pair

(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]

Solution:

$$\texttt{idf}_t = \log_2(\frac{N}{\texttt{df}_t}) = \log_2(\frac{4096}{\texttt{df}_t})$$

$\texttt{idf}_{the} = \log_2(\frac{4096}{4096}) = \log_2(1) = \log_2(2^0) = 0$

$\texttt{idf}_{theory} = \log_2(\frac{4096}{256}) = \log_2(16) = \log_2(2^4) = 4$

$\texttt{idf}_{theology} = \log_2(\frac{4096}{8}) = \log_2(512) = \log_2(2^9) = 9$

[2] for correctly calculating the IDF values

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

[1]

Solution: Theology is the most discriminative of the three words [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]

Solution:

$$P(w) = \frac{count(w,C)}{\sum_w count(w,C)}$$

[1] for correct definition

$\sum_w count(w,C) = 200$

w	count	P(w)
euro	75	0.375
championship	50	0.25
scotland	25	0.125
wembley	20	0.1
ronaldo	10	0.05
referee	10	0.05
goal	5	0.025
defeat	5	0.025

[2] for correct probabilities

(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]

Solution: $KL(D1||D2) = \infty$ [1]
because e.g. euro (in D1) is not in D2 [1]

(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]

Solution:
$KL(D1||D2) = 0.2 \times \log((0.7 \times 0.2 + 0.3 \times 0.375)/(0.7 \times 0 + 0.3 \times 0.375)) +$
$0.05 \times \log((0.7 \times 0.05 + 0.3 \times 0.25)/(0.7 \times 0.08 + 0.3 \times 0.25))$

[1] each of the two correct components of the KL formula
[1] for demonstration of correct use of Jelinek-Mercer smoothing
though they're not needed it could be helpful to know the value that comes from the answer: depends on base for log
for $\log_2$: $KL(D1||D2) = 0.22067$
for $\log_e$: $KL(D1||D2) = 0.15296$
for $\log_{10}$: $KL(D1||D2) = 0.06643$

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]

Solution:

unigram	unigram count		bigram	bigram count	P($x_i|x_{i-1}$)
			$\langle s \rangle$ the	1096	1096/2400=137/300
the	3147		the five	227	227/3147
five	821		five boxing	17	17/821
boxing	536		boxing wizards	1	1/536
wizards	7		wizards jump	3	3/7
jump	692		jump quickly	420	420/692 = 105/173
quickly	587		quickly $\langle e \rangle$	500	500/587

$$P(sentence) = \frac{137}{300} \times \frac{227}{3147} \times \frac{17}{821} \times \frac{1}{536} \times \frac{3}{7} \times \frac{105}{173} \times \frac{500}{587}$$
$$= 2.81947 \times 10^{-7}$$

[3] for correct bigram probabilities: in particular [1] for correctly inferring the count of $\langle s \rangle$

[2] for correctly combining probabilities

(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]

Solution: The model is underfitting.[1] We see this because it's accuracy is no better than a random dummy baseline for the all 3 sections of the data.[1] The fit can be improved by training the model for longer.[1]

The risk of more training is that "too much" fitting leads to a model that is overfitted.[1] Overfitting can be avoided by stopping the fit process when the test fit begins to decrease even if the training fit is improving. Cross-validation is another approach that can help prevent overfitting.[1 for either of these or any other appropriate approach to avoid overfitting]

What about checking the dataset for representativeness? Proper splits? that kind of stuff.

(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]

Solution: K-Means or hierarchical clustering are the obvious choices. [0.5] Further [0.5] for a brief description of how the chose clustering method is implemented. These are suitable because they are unsupervised ways to discover clusters. They partition the data into non-overlapping clusters. [1]

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

[3]

Solution: The Elbow method is most appropriate.[1] This selects the value of K by performing the clustering for a range of values of K.[1] The Sum of Squared Errors (SSE) is plotted against K over a range of values of K. Increasing K reduces SSE (until it is perfect when K=N, the number of data points) but at some point the rate of improvement from one K to the next is reduced. The value of K immediately before this happens should be chosen. [1]

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]

> **Solution:** While transforming an embedding into a one-hot encoding representation could lead to more efficient computations overall, in doing so, most of topical/semantic information would be disregarded thus making it, "most likely", a pointless endeavour.
>
> [2 marks for the right answer]
> [3 marks for a good motivation/explanation]

(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]

> **Solution:** The previously annotated material can be used to fine-tune BERT in order to encode the style and most likely words that the dictator would have used. The default Masked Language Model used by BERT's training process would be ideal to help in this task, as given the context of a missing word/s, a most likely candidate can be proposed attending both to the extensive LM already learnt by a pre-loaded BERT model, and what was learnt through the fine-tuning process on previously existing transcribed content.
>
> [2 marks for involving BERT's MLM]
> [3 marks for a well motivated solution/explanation]

(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]

> **Solution:** MLM can also be used. In this case we are looking at which words may be "out of place" i.e. not likely to have been pronounced by the dictator. By randomly hiding chunks of text we can compare the estimations of our learnt model in predicting those words against the actual text. Assuming our fine-tuned model has a good understanding of how the dictator communicates and word

choices, the impostor text should be statistically dissimilar to the estimations and proposed words given by the model. Next sentence prediction could also be a good answer here, as we can measure how the estimated sentence differs from the actual sentence. Higher differences could be a good indicator that the "imitator" is the author of that content.

[2 marks for involving BERT's MLM or Next Sentence Prediction]
[3 marks for a well motivated solution]

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]

Solution:

	May	Is	Tomorrow	Rain
NN	0.5	0	1	0
VB	0	1	0	1
MD	0.5	0	0	0

Each word in the training set has exactly one PoS assignment except for May which can be either MD or NN thus giving a 0.5 probability for each option.

[5 marks for the right values in the Table]

(ii) Build a HMM transitions table [6]

Solution:

	NN	VB	MD	⟨e⟩
⟨s⟩	1	0	0	0
NN	0	1/3	1/3	1/3
VB	1/2	0	0	1/2
MD	0	1	0	0

Note how NN can either be followed by a VB, an MD or the end of the sentence with an equal probability of $1/3$ each. Similarly a VB can be followed by either a NN or the end of sentence.

[6 marks for the right values in the Table]

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

Solution: The resulting sequence will be: $\langle s \rangle \Rightarrow$ (NN) $\Rightarrow$ (VB) $\Rightarrow$ (NN) $\Rightarrow \langle e \rangle$. The values in the tables were designed to result in 0's for all but the right sequence, so computations are simplified, and only one combination will prevail. The right and only sequence has a probability of 0.0277.

Note that given the ambiguity of the word "May" the final sequence assigns NN to "may" for the testing sequence.

[3 mark for proof of the computation to obtain the right sequence]
[1 marks for the total probability of the sequence]

END OF QUESTION PAPER