THE UNIVERSITY OF WARWICK

Department: Computer Science

Module Code: CS918

Module Title: Natural Language Processing

Exam Paper Code: CS9180_A

Duration: 2 hours

Exam Paper Type: 24-hour window

STUDENT INSTRUCTIONS

1. Read all instructions carefully. We recommend you read through the entire paper at least once before writing.
2. There are **SIX** questions. All candidates should attempt **FOUR** questions.
3. You should not submit answers to more than the required number of questions.
4. You should handwrite your answers either with paper and pen or using an electronic device with a stylus (unless you have special arrangements for exams which allow the use of a computer). Start each question on a new page and clearly mark each page with the page number, your student id and the question number. Handwritten notes must be scanned or photographed and all individual solutions collated into a single PDF with pages in the correct order.
5. Please ensure that all your handwritten answers are written legibly, preferably in dark blue or black ink. If you use a pencil ensure that it is not too faint to be captured by a scan or photograph.
6. Please check for legibility before uploading. It is your responsibility to ensure your work can be read.
7. Add your student number to all uploaded files.
8. You are permitted to access module materials, notes, resources, references and the Internet during the online assessment.
9. You must not communicate with any other candidate during the assessment period or seek assistance from anyone else in completing your answers. The Computer Science Department expects the conduct of all students taking this assessment to conform to the stated requirements. Measures will be in operation to check for possible misconduct. These will include the use of similarity detection tools and the right to require live interviews with selected students following the assessment.
10. By starting this assessment, you are declaring yourself fit to undertake it. You are expected to make a reasonable attempt at the assessment by answering the questions in the paper.

IMPORTANT INFORMATION

- We strongly recommend you use Google Chrome or Mozilla Firefox to access the Alternative Exams Portal.
- You are granted an additional 45 minutes beyond the stated duration of this assessment to allow for downloading/uploading your assessment, your files and any technical delays.
- Students with approved Alternative Exam Arrangements (Reasonable Adjustments) that permit extra time and/or rest breaks will have this time added on to the stated duration.
- You must have completed and uploaded the assessment before the 24-hour assessment window closes.
- Late submissions are not accepted.
- If you are unable to submit your assessment, you must submit Mitigating Circumstances immediately, attaching supporting evidence and your assessment. The Mitigating Circumstances Panel will consider the case and make a recommendation based on the evidence to the Board of Examiners.

SUPPORT DURING THE ASSESSMENT

Operational Support:

- Use the Alternative Exams Portal to seek advice immediately if during the assessment period:
  - you cannot access the online assessment
  - you believe you have been given access to the wrong online assessment

Technical Support:

- If you experience any technical difficulties with the Alternative Exam Portal please contact *helpdesk@warwick.ac.uk*.
- Technical support will be available between 09:00 and 17:00 BST for each examination (excluding Sunday).

Academic Support:

- If you have an academic query, contact the invigilator (using the 'Contact an Invigilator' tool in AEP) to raise your issue. Please be aware that two-way communication in AEP is not currently possible.
- Academic support will normally be provided for the duration of the examination (i.e. for a 2 hour exam starting at 09:00 BST, academic support will normally be provided between 09:00 and 11:45 BST). Academic support beyond this time is at the discretion of the department.

Other Support:

- If you can not complete your assessment for the following reasons submit Mitigating Circumstances immediately:
  - you lose your internet connection
  - your device fails
  - you become unwell and are unable to continue
  - you are affected by circumstances beyond your control

CS918

THE UNIVERSITY OF WARWICK

Examination: Summer 2021

Natural Language Processing

---

1. (a) Write **regular expressions** for the following languages.

   i. The set of all upper case alphabetic strings starting with `K`; [1]

   ii. the set of all alphanumeric words with one or more digits either at the beginning or at the end. It should match `5exam`, `exam123`, but not `exa7m`, `2exam99`, or `exam`; [3]

   iii. The set of all strings that denote prices in dollars. Assume these money expressions are numeric, start with a dollar sign and capture up to two decimal places. [3]

   (b) For the following sentence: [6]
   `Two planes are flying high in the sky.`
   What are the results of applying lemmatisation and stemming, respectively?

   (c) What is the vocabulary size ($|V|$) for the following sentence? [3]
   `The four hobbits make their way back to the Shire, only to find that it has been taken over by men directed by "Sharkey".`

   (d) Consider the word "`jam`". Three different senses drawn from WordNet, along with their glosses and examples are shown below: [5]

Sense jam-1
Gloss: a crowded mass that impedes or blocks <a traffic jam>
Example: Trucks sat in a jam for ten hours waiting to cross the bridge.

Sense jam-2
Gloss: an often impromptu performance by a group especially of jazz musicians that is characterized by improvisation
Example: The saxophone players took part in a free-form jazz jam

Sense jam-3
Gloss: a food made by boiling fruit and sugar to a thick consistency
Example: He spread home-made jam on his toast.

For the word “`preserve`” and its gloss in the WordNet: “`fruit preserved by cooking with sugar`", compute its similarity with each of the senses of “`jam`” using the Lesk algorithm. Describe the Lesk algorithm first and then show how to calculate the word pair similarity in this instance.

(e) Describe how to segment sentences in running text. What are the challenges involved in sentence segmentation? [4]

---

2. This question is concerned with **Language Modelling**.

   (a) What is smoothing and why do we use it in language modelling? [2]

   (b) Describe Interpolated Kneser-Ney smoothing for bigrams. [6]

   (c) For a text corpus containing a total of 5,000 word tokens and 2,000 unique words, we have the following counts of unigrams and bigrams:

n-**gram**	beautiful	flower	sky	beautiful flower	beautiful sky
Count	120	140	360	45	0

   i. Estimate the probabilities $P(flower)$ and $P(flower|beautiful)$ using Maximum Likelihood estimation. [4]

   ii. Estimate the bigram probability $P(sky|beautiful)$ using Maximum Likelihood estimation and add-k smoothing with $k = 0.5$. [4]

   (d) i. Calculate the probability of the sentence '`i want to eat chinese food`'. Give two probabilities, one using the bigram probabilities in Figure 1, and another using the add-one smoothed bigram probabilities in Figure 2. Assume $P(i|\text{<s>}) = 0.25$ and $P(\text{</s>}|food) = 0.68$, where <s> and </s> denote the sentence start and end respectively. [6]

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 1: Bigram probabilities for eight words.

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Figure 2: Add-one smoothed bigram probabilities for eight word.

   ii. Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why. [3]

3. (a) The evaluation of a text classifier produced the following confusion matrix.

Predicted Class	**Actual Class**		
	A	B	C
A	45	3	2
B	7	21	6
C	1	10	72

Based on this confusion matrix, compute the following values:

i. Macro-averaged F1. [4]

ii. Micro-averaged F1. [3]

(b) Consider the following Grammar: [12]

S$\rightarrow$NP VP	NP$\rightarrow$Alice	Verb$\rightarrow$drove
VP$\rightarrow$Verb NP	NP$\rightarrow$London	P$\rightarrow$to
VP$\rightarrow$Verb PP	NP$\rightarrow$Birmingham	P$\rightarrow$in
VP$\rightarrow$VP PP	NP$\rightarrow$December	CONJ$\rightarrow$and
NP$\rightarrow$NP PP		
NP$\rightarrow$NP and NP		
PP$\rightarrow$ P NP		

Show three possible parse trees for the sentence below by applying the given grammar.

`Alice drove to London and Birmingham in November.`

(c) Name two groups of methods for obtaining word embeddings. Give one advantage and one disadvantage of each of them. [6]

4. (a) Say we have a training set consisting of two tagged sentences: [8]

```
the/DT can/NN is/VB in/IN the/DT shed/NN
the/DT dog/NN can/VB see/VB the/DT cat/NN
```

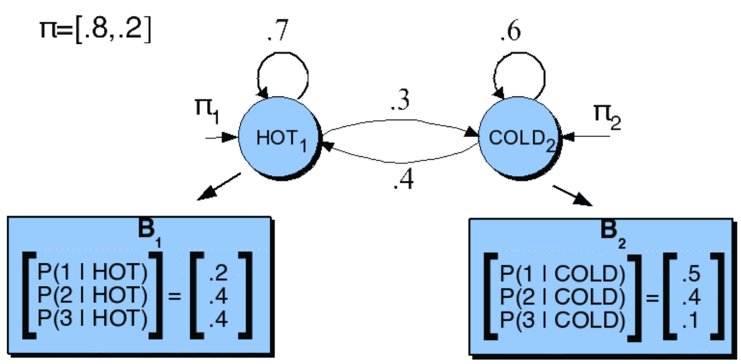
We train a bigram HMM tagger of the form

$$P(x_1, \cdots, x_n, y_0, y_1, \cdots, y_n) = \prod_{i=1}^{n} P(x_i|y_i)P(y_i|y_{i-1})$$

using simple maximum-likelihood estimates for the parameters, assuming y_0 is always assigned with a special `START` tag, and $P(y_1 = DT|y_0 = START) = 1$.

If we then use the Viterbi algorithm to find the maximum probability tag sequence for each of the training sentences, show that the tagger tags both sentences correctly. [**Hint**: check which word may have ambiguous tags and show that it is not possible to tag this word wrongly based on the training data.]

(b) Given the HMM in the diagram below, with weather conditions H and C as hidden states, and numbers of ice creams consumed by an individual as observations, compute using the Viterbi algorithm the most likely sequence of hidden states that would produce the observation sequence [2,2,1]. [9]



(c) Our recommendation engine has produced the ranking and relevance judgements output shown below for two queries. [8]

Ranking	Q1	Q2
1	Relevant	Not-relevant
2	Relevant	Relevant
3	Not-relevant	Relevant
4	Relevant	Not-relevant
5	Relevant	Relevant
6	Relevant	Relevant
7	Not-relevant	Not-relevant
8	Not-relevant	Relevant
9	Relevant	Not-relevant
10	Not-relevant	Relevant

Compute the Average Precision (AP) and Normalised Discounted Cumulative Gain (NDCG) for each query, as well as the Mean Average Precision (MAP) for both queries combined, up to the 5th element in all cases. Does our system perform better for Q1 or Q2?

5. A mobile network company receives many emails each day from its customers about fault reporting, bills enquiry, service complaint, upgrade request, etc. You have been hired as an NLP expert to help build an intelligent email handling software system that has the following functionality: filtering emails based on their priorities (e.g., some emails need to be attended immediately) and automatically directing the emails to the right person in the customer service department. In this context, answer the following questions:

   (a) What assumptions do you need to make about this NLP task? [4]

   (b) What are the set of features that you will extract from each email? What are the two most important? [6]

   (c) How would you represent each of the features extracted from emails? What pre-processing steps would you consider? [8]

   (d) How would you apply NLP approaches to this application? State the reasons for your approaches. [7]

6. (a) Explain the principle of the gradient descent algorithm for learning a neural network. You can use a diagram to illustrate your explanation. Explain the use of all the variables and constants that you introduce. [5]

   (b) Name two common choices for choosing the learning rate and briefly explain how they work. [6]

   (c) Suppose we want to classify movie review text as (1) either `positive` or `negative` sentiment; and (2) either `action`, `comedy` or `romance` movie genre. To perform these two related classification tasks, we use a neural network that shares the first and the hidden layers, but branches into two separate layers to compute the two classifications. The loss is a weighted sum of the two cross-entropy losses. [6]

   Assume in our training data, we have review documents each with the maximum length of 10 words and each document is labelled with its sentiment polarity as well as the genre. Roughly sketch a fully-connected single hidden layer neural network (hidden layer of size 5) that takes the review text as input using one-hot encoding and outputs sentiment polarity and genre. In your sketch, clearly mark the input layer, hidden layer and the outputs. You need not draw all the nodes or all edges between two layers, but make sure to indicate clearly the dimensionality of each layer.

   (d) For the neural network architecture in Part (c), assume the hidden layer uses ReLU as the activation function and the output layer uses the softmax function to generate the outputs. Write the equations of the neural network which map from the input to the output (you can use multiple equations), clearly indicating the dimension of each vector/matrix used in the equations. [6]

   (e) Write down the loss function of the neural architecture you designed in Part (c). [2]