Autumn Examinations 2021/2022

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1, 1MAO2, 1MAI1
Exam(s)	MSc in Computer Science (Data Analytics), MSc in Computer Science (Artificial Intelligence), MSc in Computer Science (Artificial Intelligence) - Online
Module Code(s)	CT5120, CT5146
Module(s)	Introduction to Natural Language Processing, Introduction to Natural Language Processing - Online
Paper No.	1
External Examiner(s)	Dr John Woodward
Internal Examiner(s)	Dr. Michael Madden *Dr. John McCrae Dr Bharathi Raja Chakravarthi Dr Omnia Zayed

Instructions: Answer 4 sections out of 5; each section is worth 25 marks (100 marks total). **Use a separate answer book for each section answered.**

Duration	2 hours
No. of Pages	6
Discipline(s)	Computer Science
Course Co-ordinator(s)	Dr. Frank Glavin Dr. Matthias Nickles Dr. James McDermott

Requirements:

Release in Exam Venue	Yes
MCQ	No
Handout	None
Statistical/ Log Tables	None
Cambridge Tables	None
Graph Paper	None
Log Graph Paper	None
Other Materials	None
Graphic material in colour	No

Introduction to Natural Language Processing

Exam Duration: 2 Hours

You must answer 4 of the following sections

Section 1: Text Classification

Question 1A: **5 Marks**

What does it mean when a word is **out of vocabulary**? List two solutions to this problem and describe how they solve this issue.

Question 1B: **10 Marks**

State the formula for TF-IDF (Term Frequency-Inverse Document Frequency). How can TF-IDF be used with Naive Bayes to perform text classification?

Question 1C: **10 Marks**

Consider the following sentences with sentiment labels.

- This house room was good [POS]
- The roof was not as good as expected [NEG]
- The garden was good for the kids [POS]
- The kids loved the flowers [POS]

$$p(POS \mid good) = \frac{c(POS \cap good)}{c(good)}$$

$$\frac{2}{3} \qquad = \frac{2}{3}.$$

Using Bayes' Law, calculate the probability of the labels POS and NEG given a single feature that considers whether the word 'good' occurs in the text.

$$p(A \mid B) = \frac{c(A \cap B)}{c(B)}$$

$$p(NEG \mid good) = \frac{1}{3}$$

PTO

1A - vocabularies that does not appear in ~~training~~ set and hence model couldn't not recognize it in testing and giving 0 probabilities to the word.

- add-one smoothing, add 1 to all the words in train set.

- laplace smoothing. - add α amount to all the words in train set & normalize with expected number of OOV.

1B $\text{TF-IDF} = tf \times \left(\left(\log \frac{N}{N_w} \right) + 1 \right)$

tf = term frequency in the doc

N = number of docs

N_w = term frequency across all docs.

↳ can be used in NB by assuming it follows a probability distribution

$$\text{TF-IDF}_w \sim N(\mu_w, \sigma_w).$$

TF-IDF is used as vectorizer for feature extraction as input for NB.

Section 2: Linguistic Concept and Parsing

break a corpus to small unit.

Question 2A: **5 Marks**

Explain what task a tokenizer performs and how it deals with issues such as punctuation. How many types and tokens in the following sentence?

treated as a single token

'the cat sat on the mat'

6 tokens = the, cat, sat, on, the, mat

5 types = the, cat, sat, on, mat.

Question 2B: **10 Marks**

Consider the probabilistic context-free grammar below. Draw **one** parse tree and calculate the probability of that parse for the following sentence: "Connor Murphy and Ciara Byrne dance"

Rule	Probability	Rule	Probability
S → NP V	0.9	NN → Connor	0.2
S → CL CONJ CL	0.1	NN → Murphy	0.2
CL → NP V	1.0	NN → Ciara	0.2
NP → NP CONJ NP	0.2	NN → Byrne	0.2
NP → NN NP	0.3	NN → dance	0.2
NP → NN	0.5	V → Murphy	0.1
CONJ → and	1.0	V → dance	0.9

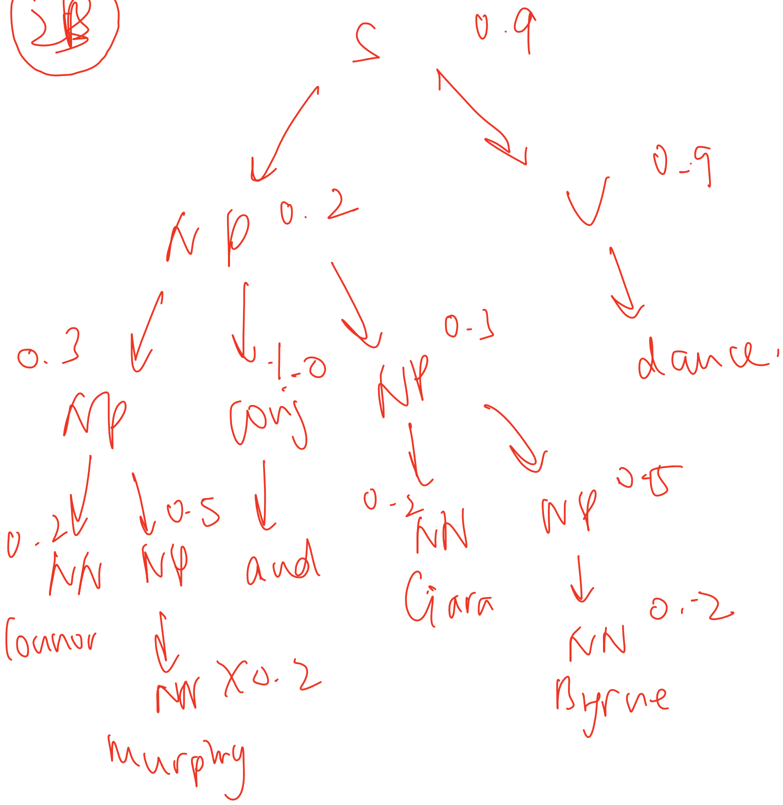
Question 2C: **10 Marks**

it's snowing it's falling
the old lady is snoring
she went to roof
and she bumped her head
and she couldn't get up in the morning.

For the above calculate all unigram and bigram probabilities. You should treat "it's" and "couldn't" as single tokens. Treat the whole corpus as a single sentence.

PTO

3B



$$= 0.9 \times 0.2 \times 0.9 \times 0.3 \times 1.0 \times 0.3 \times 0.2 \times 0.5 \times$$
$$0.2 \times 0.5 \times 0.2 \times 0.2$$

$$= 0.000005832$$

2c

it's	$\frac{2}{26}$	root	$\frac{1}{26}$
snowing	$\frac{1}{26}$	and	$\frac{2}{26}$
falling	$\frac{1}{26}$	jumped	$\frac{1}{26}$
the	$\frac{2}{26}$	her	$\frac{1}{26}$
old	$\frac{1}{26}$	head	$\frac{1}{26}$
lady	$\frac{1}{26}$	couldn't	$\frac{1}{26}$
is	$\frac{1}{26}$	get	$\frac{1}{26}$
snoring	$\frac{1}{26}$	up	$\frac{1}{26}$
she	$\frac{3}{26}$	in	$\frac{1}{26}$
went	$\frac{1}{26}$	morning	$\frac{1}{26}$
to	$\frac{1}{26}$		

unigram

sum = 26

2C Bigram.

it's snowing	1/2
snowing it's	1
it's falling	1/2
falling the	1
the old	1/2
old lady	1
lady is	1
is coming	1
coming the	1
she went	1/3
went to	1
to root	1

root and	1
and she	2/2 = 1
she bumped	1/3
bumped her	1
her head	1
head and	1
she couldn't	1/3
couldn't get	1
get up	1
up in	1
in the	1
the morning	1/2

Sum 2.25

Section 3: Vector Space Models

Question 3A: **10 Marks**

Give **two** reasons that we may create a vector representation of a word. Explain how a vector representation solves these problems.

Question 3B: **10 Marks**

*…its name stands for "**language** model for dialogue applications"...*
*…an example of a very large **language** model, or a computer program…*
*…I wrote recently, using **language** models in place of search engines …*
*…with the help of a **language** model. It has…*
…human-like interfaces such as **language.** For any automated system…

Source: The Guardian

Create a context vector for the word 'language' from the text above using a context window of two words either side (ignore all punctuation and case).

Question 3C: **5 Marks**

Suggest **one** change you could make in how the context vector is constructed that may improve performance.

PTO

3A

1. To detect similarity of words - based on context
   could calculate similarity btwn vectors
   cosine.

distributional semantics.

2. reduce dimensionality.
   optimize sparse vectors to shorter dense vectors (word embeddings)
   ① initialize with one-hot encoding
   ② count / predict co-occurrence for all words in vocab V within a given context with co-occurrence as positive instance or negative instance if otherwise.

3B.

context - frequency vector (bag-of-words)

target: language

vocab :

stands	1
for	3
model	3
very	1
large	1
or	1
recently	1
using	1
models	1
in	1
of	1
a	1
it	1
such	1
as	1
any	1

sum = 20

3C.

instead of vectorizing with frequency, we could compute the weight of each word to express how relevant or specific context c is for word w. We could use tf-idf to vectorize with weight.

Section 4: Information Extraction

Meat, cereals, dairy, fruit and vegetables are likely to be the worst affected as the war in Ukraine combines with production lockdowns in China and export bans on key food stuffs such as palm oil from Indonesia and wheat from India, the grocery trade body IGD warns.

Products that rely on wheat, such as chicken, pork and bakery items, are likely to face the most rapid price rises as problems with exports and production from Ukraine, a big producer of grain, combine with sanctions on Russia, another key producer.

Source: The Guardian

Question 3A: **10 Marks**

From the text above extract at least two hypernym relations by means of Hearst patterns

Question 3B: **5 Marks**

Indicate two named entities within the text above and provide the appropriate entity class.

Question 3C: **10 Marks**

What is a tagging scheme that would allow a tagging model such as a hidden Markov model in order to tag named entities within a text? Give an example of this tagging on one sentence from the text above.

PTO

Section 5: Semantic Analysis

Question 5A **10** Marks

List and **explain** the three levels of semantic analysis and **give** an example of a task for each level.

Question 5B **10** Marks

Define Word Sense Disambiguation (WSD) and **list** three NLP applications that can benefit from it.

Question 5C **5** Marks

Consider the following sentence:

Bob hit Scott with the bottle.

Identify the "semantic roles" for each entity of the event expressed by the given sentence.

END

5A.

1. lexical semantics - identify meaning of word
    A. with word sense disambiguation
        a. knowledge-based with LESK (unsupervised)
2. compositional semantics - identify the semantic role for each word / phrase
    A. semantic role labelling
        a. use FrameNet - supervised to identify roles
        b.
3. discourse semantics - identify which words / phrases refer to the same entity across sentences
    A. coreference resolution
        a. mention detection / clustering

5B.

As words can have different meaning due to homonymy, hence WSD is to find out the correct word sense based on the context.

1. suggestion detection
2. sentiment analysis
3. offensive content detection.

5C.

bob - agent
scott - experiencer
the bottle - instrument