



## **Spring Deferrals 2021/2022**

<b>Course Code(s)</b>	<b>Instance</b>	1CSD1, 1CSD2, 1SPE1, 1MAO2, 1MAI1
<b>Exam(s)</b>		MSc in Computer Science (Data Analytics), MSc in Computer Science (Artificial Intelligence), MSc in Computer Science (Artificial Intelligence) - Online
<b>Module Code(s)</b>		CT5120, CT5146
<b>Module(s)</b>		Introduction to Natural Language Processing, Introduction to Natural Language Processing - Online
<b>Paper No.</b>		1
<b>External Examiner(s)</b>		Dr John Woodward
<b>Internal Examiner(s)</b>		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.**

<b>Duration</b>	2 hours
<b>No. of Pages</b>	6
<b>Discipline(s)</b>	Computer Science
<b>Course Co-ordinator(s)</b>	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

**10 Marks**

Explain what is meant by text classification and give **two** examples of tasks that may be solved by means of text classification

### Question 1B

**10 Marks**

State the formula for TF-IDF.

Consider the following corpus, treating each sentence as a separate document

- You are called upon to deliberate on a new Constitution for the United States of America
- Yes, my countrymen, I own to you that, after having given it a thorough consideration, I am clearly of opinion it is in your interest to adopt it
- It is not a new observation that the people of any country seldom adopt and steadily persevere for many years in an erroneous opinion respecting their interests

Calculate TF-IDF vectors for each document containing the following words: a, constitution, country, it, you

### Question 1C

**5 Marks**

Suggest a solution to the problem of out-of-vocabulary words in text classification

**PTO**

## Section 2: Sequence Models

### Question 2A

5 Marks

State the formula for a bigram language model

### Question 2B

10 Marks

*it's raining it's pouring  
the old man is snoring  
he went to bed  
and he bumped his head  
and he 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.

### Question 2C

5 Marks

Using the probabilities calculated above, what is the probability of the sentence “And he went to bed”

### Question 2D

5 Marks

The probability for the sentence “He went to bed in the morning” is zero. Suggest a modification to the bigram language model to produce a non-zero probability for this sentence.

PTO

### Section 3: Semantic Analysis

#### Question 3A

10 Marks

**Define** semantic analysis. **List** and **explain** at least **three** tasks that it entails.

#### Question 3B

5 Marks

Explain the steps involved in coreference resolution.

#### Question 3C

10 Marks

Consider the following text:

*Joe Biden has announced Claire Cronin as his nominee for the position of US ambassador to Ireland. Cronin was a key campaigner for him in her home state of Massachusetts, where she was serving as the Majority Leader of the Massachusetts House of Representatives.*

Apply the steps explained in 3B to resolve all coreferences in the given text.

**PTO**

## Section 4: Social Media Analysis

### Question 4A

10 Marks

**List** and **explain five** different applications of NLP in social media.

### Question 4B

10 Marks

**List** and **explain three** different task formulations of sentiment analysis. **Give one** example for each task.

### Question 4C

5 Marks

What does suggestion mining involve? (**List** and briefly **explain** at least **three** sub-tasks)

**PTO**

## Section 5: Information Extraction and Vector Space Models

### Question 5A

10 Marks

Consider the following texts:

Doc 1: Government published NPHET advice

Doc 2: NPHET concerned about outbreaks in workplaces

Doc 3: NPHET advice says it is impossible to predict the trajectory of Covid-19

Create a Term-Document matrix (alphabetically sorted)

### Question 5B

10 Marks

Explain what Inter-annotator agreement is and calculate Cohen's kappa for following following annotation matrix

		annotator A	
		puppy	fried chicken
annotator B	puppy	6	3
	fried chicken	2	5

### Question 5C

5 Marks

Calculate precision, recall and F1 for the following

# Gold Standard items (GS) 40

# Extracted items (EX) 60

# Correctly extracted items (CEX) 20

END