Spring Deferrals 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 **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

① Add one smoothing / Laplace smoothing

PTO

add 1 to every word count (of word in the vocab)

→ all the selected feature words are assumed to appear in all different class at least once.

② subword - feature engineering with n-grams

to break string into pieces.

(1A) Text classification is ~~to~~ identify the class / intent of a sentence.

i. sentiment analysis (positive / negative)

ii. Suggestion detection (yes / no)

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

(TF) in curr. doc

(IDF) in all doc

$a_1 = a_2 = a_3 = TF = 1$, $N = 3$, $N_w = 3$ $= \log \frac{3}{3} + 1$

$= 1$

$Constitution_1 = 1 \times \left(\log \frac{3}{1} + 1 \right) = 1.4771$

$country_3 = 1 \times \left(\log \frac{3}{1} + 1 \right) = 1.4771$

$it_2 = 3 \times \left(\log \frac{3}{4} + 1 \right) = 2.625$

$it_3 = 1 \times \left(\log \frac{3}{4} + 1 \right) = 0.8751$

$you_1 = 1 \times \left(\log \frac{3}{2} + 1 \right) = 1.1761$

$you_2 = 1.1761$

2A

$p(w_1\ w_2\ w_3\ w_4) = p(w_4 | w_3) \times p(w_3 | w_2) \times p(w_2 | w_1) \times p(w_1)$

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

$p(\text{bed} | \text{to}) \times p(\text{to} | \text{went}) \times p(\text{went} | \text{he}) \times p(\text{he} | \text{And}) \times p(\text{And})$.

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

$= 1 \times 1 \times \frac{1}{3} \times 1 \times 0.08 = 0.027$.

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

↓

add one smoothing.

↓

~~add~~ assign non-zero probability to $p(\text{in} | \text{bed})$

(2B).

it's	2	(0.08)	snoring	1	(0.04)	his	1	(0.04)
raining	1	(0.04)	he	3	(0.12)	head	1	(0.04)
pouring	1	(0.04)	went	1	(0.04)	couldn't	1	(0.04)
the	2	(0.08)	to	1	(0.04)	get	1	(0.04)
old	1	(0.04)	bed	1	(0.04)	up	1	(0.04)
man	1	(0.04)	and	2	(0.08)	in	1	(0.04)
is	1	(0.04)	bumped	1	(0.04)	morning	1	
	9			10			7	= 26

---

it's raining	1	($\frac{1}{2}$)	is snoring	1		he bumped	1	($\frac{1}{3}$)
raining it's	1		snoring he	1		bumped his	1	
it's pouring	1	($\frac{1}{2}$)	he went	1	($\frac{1}{3}$)	his head	1	
pouring the	1		went to	1		head and	1	
the old	1		to bed	1		he couldn't	1	($\frac{1}{3}$)
old man	1		bed and	1		couldn't get	1	
man is	1		and he	2	↑ (1)	get up	1	
						up in	1	
						in the	1	
						the morning	1	

$$p(\text{raining} \mid \text{it's}) = \frac{p(\text{it's raining})}{p(\text{it's raining}) + p(\text{it's pouring})}$$

Sum = 25

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

mention detection = 🟨

mention clustering

① Joe Biden, the position of US ambassador, him, his

② Claire Cronin, nominee, Cronin, key campaigner, her, she, the Majority Leader of the Mass.... House of Rep ...

3A automatic methods for constructing unambiguous meaning representations for linguistic expression.

↳ i → Lexical semantics. (word sense disambiguation)

↳ identify word meaning

ii → compositional semantics (semantic role labelling)

→ identify semantic role for each word/phrase

iii ↳ discourse semantics (coreference resolution)

→ identify which words/phrases refer to the same 'entity' across sentence.

3B Step 1 → identify all mentions

↳ coreference candidates.

↳ (mention detection)

Step 2 ↳ identify all mentions that refer to the same real world entity

↳ (mention clustering)

Section 4: Social Media Analysis

Question 4A **10** Marks

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

Question 4B **10** Marks

approach ? step ?

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

4A spam detection, offensive content detection, misinformation detection, suggestion mining, sentiment analysis.

4B ① Data collection — e.g. Collect data like tweets from social media via APIs.

—

② Data preprocessing - spelling checker, clean emojis.

- process data to prepare input for model training

③ model selection — select a suitable classification model - supervised / unsupervised / etc

—Binary classification / ordinal classification)

4C suggestion classification

- yes/no

suggestion extraction

- span identification of the suggestion

suggestion aspect detection

- identify aspect of the suggestion

suggestions clustering

- aggregate similar suggestions etc.

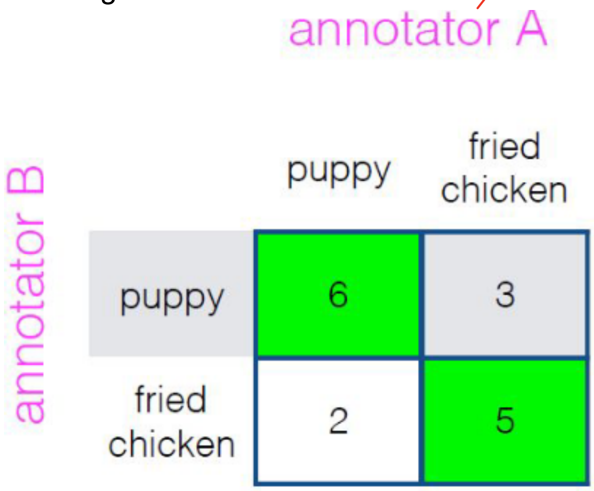
Section 5: Information Extraction and Vector Space Models

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

Tp + TN = 20.

FP + FN = 60.