NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2017/ 2018

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1
Exam(s)	MSc in Computer Science (Data Analytics)
Module Code(s)	CT5101
Module(s)	Natural Language Processing
Paper No.	1
Repeat Paper	No
External Examiner(s)	Professor Pier Luca Lanzi
Internal Examiner(s)	Dr. Michael Madden *Dr. Paul Buitelaar Dr. John McCrae

Instructions: Answer all parts of all questions. There are 4 sections; each section is worth 25 marks (100 marks total). **Use a separate answer book for each section answered**.

Duration	2 hours
No. of Pages	5
Discipline(s)	Engineering and Information Technology
Course Co-ordinator(s)	Dr. Conor Hayes

Requirements:

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

CT5101 Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section 1: Linguistic Structure, Data and Analysis

Question 1A **5** Marks

Consider the following sentence:

the man with the hat and tie came after the man in the dark suit

How many types and tokens are there in the sentence?

Question 1B **15** Marks

Define a constituency (phrase) grammar and lexicon that analyses the following sentence by using the non-terminal symbols 'S, NP, VP, PP' and the pre-terminal symbols 'Det, Noun, Verb, Prep'.

The minister visited the power plant in the south of the country.

Draw a constituency (phrase) structure tree for this sentence, using the grammar and lexicon you defined.

Question 1C **5** Marks

What is the difference between a parallel corpus and a comparable corpus?

PTO

Section 2: Textual Similarity

For this section, consider the following two sentences:

s_1 = *Tusk swipes at May for better border talk.*

s_2 = *Tusk asks May for better border idea.*

Question 2A **5** Marks

Calculate the following similarities for s_1 and s_2:

- Dice similarity using a bag-of-words model
- Jaccard similarity using a bag-of-words model
- The length of the Longest Common Subsequence

Question 2B **15** Marks

Recall that Damerau-Levenshtein Edit Distance is calculated using the following equation:

$$d(i,j) = min \begin{Bmatrix} d(i-1,j)+1 \\ d(i,j-1)+1 \\ d(i-1,j-1)+1 \\ d(i-1,j-1) \ \ if\ s_{1,i} = s_{2,j} \\ d(i-2,j-2)+1 \ \ if\ s_{1,i} = s_{2,j-1} \ \wedge \ s_{1,i-1} = s_{2,j} \end{Bmatrix}$$

What is the Damerau-Levenshtein distance between these sentences? Explain your method.

Question 2C **5** Marks

What is the 'Big O' complexity of the method you used to calculate Edit Distance? Name the methodology you used or could use to reduce this complexity.

PTO

Section 3: Language Modelling

Consider the following poem by A.A Milne as a corpus. Treat each line as a new sentence. Ignoring punctuation, it is 100 words long.

> *The wind on the hill.*
>
> *No one can tell me nobody knows where the wind comes from where the wind goes.*
>
> *It's flying from somewhere as fast as it can I couldn't keep up with it not if I ran.*
>
> *But if I stopped holding the string of my kite, it would blow with the wind for a day and a night.*
>
> *And then when I found it, wherever it blew, I should know that the wind had been going there too.*
>
> *So then I could tell them Where the wind goes.*
>
> *But where the wind comes from nobody knows.*

Question 3A **5** Marks

Calculate the unigram probabilities *ignoring case* for the words: “been”, “had”, “the”, “wind”, “where”.

Question 3B **5** Marks

Calculate the bigram probability ignoring case for the combinations that are not provided in the following table:

$p(w_2\|w_1)$	w_2=been	w_2=had	w_2=the	w_2=where	w_2=wind
w_1=been	0	0	0	0	0
w_1=had		0	0	0	0
w_1=the	0	0	0		
w_1=where	0	0		0	0
w_1=wind	0		0	0	0

Question 3C **10** Marks

State the formula for a **bigram language model** applied to the sentence “the wind had been”. Using this bigram language model calculate the probability of the line “the wind had been”.

Question 3D **5** Marks

p(“The wind had been there”) = 0 given the bigram model. Briefly explain why and suggest a model that produces a non-zero probability for this sentence.

PTO

Section 4: Information Extraction

Consider the following sentences:

> S_1 = *Shares in Smurfit Kappa have risen by over 18pc in early trading on the London Stock Exchange.*
>
> S_2 = *Adidas shares were up 1.6%, marking the biggest increase among the largest shares in Germany.*
>
> S_3 = *United Technologies shares rise 2% as a 'well-known' activist takes a position.*
>
> S_4 = *Shares of Jaypee Infratech climbed over 4 per cent on Tuesday morning.*
>
> S_5 = *Shares of Commonwealth Bank declined 0.92 percent.*

Question 4A **5** Marks

Annotate sentences S_2 and S_4 with Named Entities of type COMPANY, NUMBER, COUNTRY and TIME where appropriate. Use the following annotation format:

> [COMPANY *Adidas] shares were up [NUMBER 1.6] % …*

Question 4B **15** Marks

Assume we want to extract the following relations from the sentences above:

> Shares-Up-Percentage (COMPANY, NUMBER)
> Shares-Down-Percentage (COMPANY, NUMBER)

For instance:

> Shares-Up-Percentage (Smurfit Kappa, 18)

Provide patterns that can be used to extract this information for all companies mentioned in the sentences above.

Question 4C **5** Marks

What is the core difference between ‘open information extraction’ (Open IE) and ‘knowledge base population’ (KBP)?

END