NUI Galway
OÉ Gaillimh

Autumn Examinations 2018/ 2019

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1
Exam(s)	MSc in Computer Science (Data Analytics)
Module Code(s)	CT5120
Module(s)	Introduction to Natural Language Processing
Paper No.	1
Repeat Paper	Yes
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. Enda Howley

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

CT5120 Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section 1: Semantics

Instructions: Provide answers for questions 1A, 1B and 1C

Question 1A **10** Marks

Consider the following frequency vectors:

black	4	4	2	6	0

white	4	8	0	2	10

Using cosine similarity, compute the distributional semantic distance between ‘black’ and ‘white’.

Question 1B **10** Marks

Consider the following sense definitions for ‘bank’:

bank#1 - the slope beside a body of water

bank#2 - a financial institution that accepts deposits and channels the money into lending activities

Now consider the following occurrence of ‘bank’ in this sentence:

The bank was left free to offer interest on demand deposits.

How would you apply the Lesk algorithm to disambiguate ‘bank’ between the two senses given above?

Question 1C **5** Marks

Define homonymy, synonymy, and antonymy.

PTO

Section 2: Part-of-speech tagging

Instructions: Provide answers for question 2A and 2B

Question 2A **10** Marks

Consider a Hidden Markov Model with the following probabilities (S designates the start state):

$p(w_i\|t_i)$	w_i = the	w_i = University	w_i = of	w_i = Ireland
t_i = B	0.2	0.7	0.2	0.4
t_i = I	0.2	0.2	0.5	0.4
t_i = O	0.6	0.1	0.3	0.2

$p(t_i\|t_{i-1})$	t_{i-1} = B	t_{i-1} = I	t_{i-1} = O	t_{i-1} = S
t_i = B	0.1	0.3	0.3	0.3
t_i = I	0.6	0.4	0.0	0.0
t_i = O	0.1	0.2	0.3	0.4

What is the probability of the sequence “the University of Ireland” being tagged as “O B I I”?

Question 2B **15** Marks

Given an *annotated* text corpus, describe how you would find probabilities such as given in the table above. Write any algorithms you would use in pseudo-code.

PTO

Section 3: Sentiment Analysis

Instructions: Provide answers for question 3A, 3B, 3C and 3D

Question 3A **10** Marks

Explain **two** challenges for automatic approaches to sentiment analysis

Question 3B **5** Marks

What is a sentiment lexicon and how may it be used as a feature in a sentiment analysis classifier?

Question 3C **5** Marks

Provide a suggestion of **one** way in which negation may be handled in a sentiment analysis.

Question 3D **5** Marks

What is meant with *aspect-based* sentiment analysis? Give an example of an aspect.

PTO

Section 4: Information Extraction

Instructions: Provide answers for questions 4A and 4B

Consider the following corpus of sentences about company acquisitions, with named entity annotation (COM:company) and gold standard labeling if the sentence does or does not express a company acquisition:

#	company acquisition Y/N	Sentence
1	Y	*[COM Salesforce] to acquire data analytics firm [COM Tableau] in $15.7 billion deal.*
2	Y	*[COM Bird] confirms acquisition of [COM Scoot].*
3	Y	*[COM Shutterfly] to be merged with [COM Snapfish] after $2.7B acquisition.*
4	N	*A $3.3 billion [COM Walmart] acquisition of [COM Jet.com] is under discussion.*
5	Y	*[COM Mediahuis] acquisition of [COM INM] approved by regulator.*

Question 4A **15** Marks

What is the Precision, Recall and F-score of an information extraction system that has just one pattern, applied to the 5 sentences given above:

*[COM X] * acquisition of * [COM Y]*

Explain how you derived your answers.

Question 4B **10** Marks

Give the formula for Cohen's kappa coefficient. What is it used for in information extraction?

END