NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2016/ 2017

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1
Exam(s)	Computer Science – Data Analytics
Module Code(s)	CT5101
Module(s)	Natural Language Processing
Paper No.	1
Repeat Paper	Yes
External Examiner(s)	Professor Liam Maguire
Internal Examiner(s)	Dr. Michael Schukat *Dr. Paul Buitelaar Dr. John McCrae Dr. Ian Wood Dr. Mihael Arcan

Instructions: Answer all questions in all sections. 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	✔	No	☐
MCQ	Yes	☐	No	✔
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

Section 1: Tagging and Parsing

Instructions: Provide answers for questions 1A and 1B

Question 1A 10 Marks

Give an example of an ambiguous English sentence and explain how a part-of-speech tagger **or** a parser can resolve this ambiguity.

Question 1B 15 Marks

Consider the probabilistic context-free grammar as follows

$P(NP \rightarrow N) = 0.4$
$P(NP \rightarrow A\ NP) = 0.3$
$P(NP \rightarrow N\ NP) = 0.1$
$P(NP \rightarrow DET\ NP) = 0.2$
$P(N \rightarrow good) = 0.2$
$P(N \rightarrow work) = 0.4$
$P(N \rightarrow purchase) = 0.4$
$P(A \rightarrow good) = 0.8$
$P(A \rightarrow kind) = 0.2$
$P(DET \rightarrow the) = 0.5$
$P(DET \rightarrow a) = 0.5$

What is the probability of the noun phrase ‘a good work’ (use NP as the start symbol). Provide **all** valid parse trees.

Section 2: Machine Learning

Instructions: Provide answers for questions 2A, 2B, and 2C

Question 2A 5 Marks

Briefly explain why it is important to have separate data for training and testing supervised classification models.

Question 2B 10 Marks

The F_β measure is frequently used for text classification models. Write down the formula for the F_β measure. Briefly discuss the impact of varying the β parameter. In what circumstances would you choose a low β value such as 0.2?

Question 2C 10 Marks

Real world data for binary classification of emotion classes such as *disgust* are usually highly unbalanced, with much fewer examples expressing the target emotion class than those that do not express it.

Briefly discuss why the F_1 measure is usually a better choice than overall prediction accuracy for evaluation of a binary classification model with unbalanced test data.

Calculate the F_1 measure for *disgust* given the evaluation results below.

Emotion	**True Positive**	**False Positive**	**True Negative**	**False Negative**
joy	50	10	35	5
sadness	12	9	66	13
disgust	5	15	75	5
anger	20	5	65	10

Section 3: Machine Translation

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

Question 3A 5 Marks

Name at least three different approaches to Machine Translation.

Question 3B 10 Marks

Given the translation table below, provide four English sentences that can be generated for the German sentence "*das Haus ist klein"* and compute the translation probabilities for these sentences.

f = "das"		f= "Haus"		f = "ist"		f = "klein"	
e	p(e|f)	e	p(e|f)	e	p(e|f)	e	p(e|f)
the	0.7	house	0.8	is	0.8	small	0.4
that	0.15	building	0.16	's	0.16	little	0.4
which	0.075	home	0.02	exists	0.02	short	0.1
who	0.05	household	0.015	has	0.015	minor	0.06
this	0.025	shell	0.005	are	0.005	petty	0.04

Question 3C 5 Marks

Why is count smoothing important in language model generation?

Question 3D 5 Marks

Briefly explain how the BLEU score for automatic translation evaluation is calculated.

Section 4: Sentiment Analysis

Instructions: Provide answers for questions 4A, 4B, and 4C.

Sentiment lexicon:

Word	**Part of Speech**	**Sentiment value**
small	Adjective	-1
tiny	Adjective	-2
comfortable	Adjective	+1
delicious	Adjective	+2
fresh	Adjective	+1
variety	Noun	+1
delicious	Adjective	+2
noisy	Adjective	-2

Sentences:

1. *Our room was tiny and the bath was small too.*
2. *The bed was very comfortable, but the walls were thin and we were unfortunate to get noisy neighbours.*
3. *The breakfast did not have much variety, however everything was fresh and delicious.*

Question 4A **10** Marks

Using the following two features:
Feature 1: Count of positive sentiment adjectives in the sentence.
Feature 2: The sum of sentiment values for all the word of the sentence (Use 0 for words not in the table).

Fill in the table with the feature values for the provided sentences. Use the given lexicon to calculate the values. Use 0 if a word is not present in the lexicon.

Sentence	Feature 1	Feature 2	Aspect, Sentiment Class
#1	0	-3	Room, Negative
#2	1	-1	Bed, Positive
#3	2	4	Breakfast, Conflict

Question 4B **5** Marks

Do you think Features 1 and 2 are useful to predict the sentiment class of a given aspect? Explain.

Question 4C **10** Marks

Suggest a third feature, which you think would be useful for predicting the aspect based sentiment class of a given sentence. Explain how you would calculate the value of this feature.

END