NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2015/ 2016

Course Instance Code(s)	1CSD1
Exam(s)	Computer Science – Data Analytics
Module Code(s)	CT5101
Module(s)	Natural Language Processing
Paper No.	1
Repeat Paper	No
External Examiner(s)	Professor Liam Maguire
Internal Examiner(s)	Professor Gerard Lyons *Dr. Paul Buitelaar Dr. Georgeta Bordea Dr. John McCrae Dr. Ian Wood

Instructions: There are 6 sections, choose 4 sections and answer all questions from those sections. Each section is worth 20 marks (80 marks total). **Use a separate answer book for each section answered**.

Duration	2 hours
No. of Pages	8
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

CT5101 Natural Language Processing

Exam Duration: 2 Hours

You must complete 4 of the 6 Sections

Section 1: Linguistic Structure, Data and Analysis

Instructions: Provide answers for questions 1A and 1B

Question 1A 10 Marks

Define a formal grammar (grammar rules + lexicon) that can be used to analyse and/or generate the following sentence. The grammar should use phrase symbols such as NP, VP, AP, PP and terminal symbols such as Noun, Verb, Adjective, Preposition, etc.

He bought a very old and expensive book at the secondhand bookshop with the big windows in the square beside the church.

Question 1B 10 Marks

Provide binary vectors for "*book*" and "*novel*" with vector length equal to the number of word types in the following text. The context window for constructing the vector is the sentence in which the words "*novel*" or "*book*" occur. You should use morphological normalization, e.g. inflection, as well as other normalization such as lower/upper case.

He bought an expensive book at the bookshop.
The novel he bought turned out to be a very expensive mistake.
The books at this bookshop are mostly expensive.
Novels are often cheap to buy, although not at this bookshop.

Section 2: Probability and Classification

Instructions: Provide answers for questions 2A and 2B

Consider the task of predicting verb tense using three features: word length, word ends in 'd' and word ends in 'n'. We have the following data points:

Word	**Ends(d)**	**Length**	**Ends(n)**	**PastTense**
played	1	6	0	True
need	1	4	0	False
worked	1	6	0	True
had	1	3	0	True
smitten	0	7	1	True
kitten	0	6	1	False

Question 2A 5 Marks

What is the feature vector for “smiled” for predicting whether it is past tense?

Question 2B 15 Marks

Under a Naive Bayes model, what is the predicted value for *PastTense* for the word “smiled”? Show calculations to justify your conclusion.

Hint: Treat the features as discrete values.

Section 3: Model Evaluation and Unsupervised Learning

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

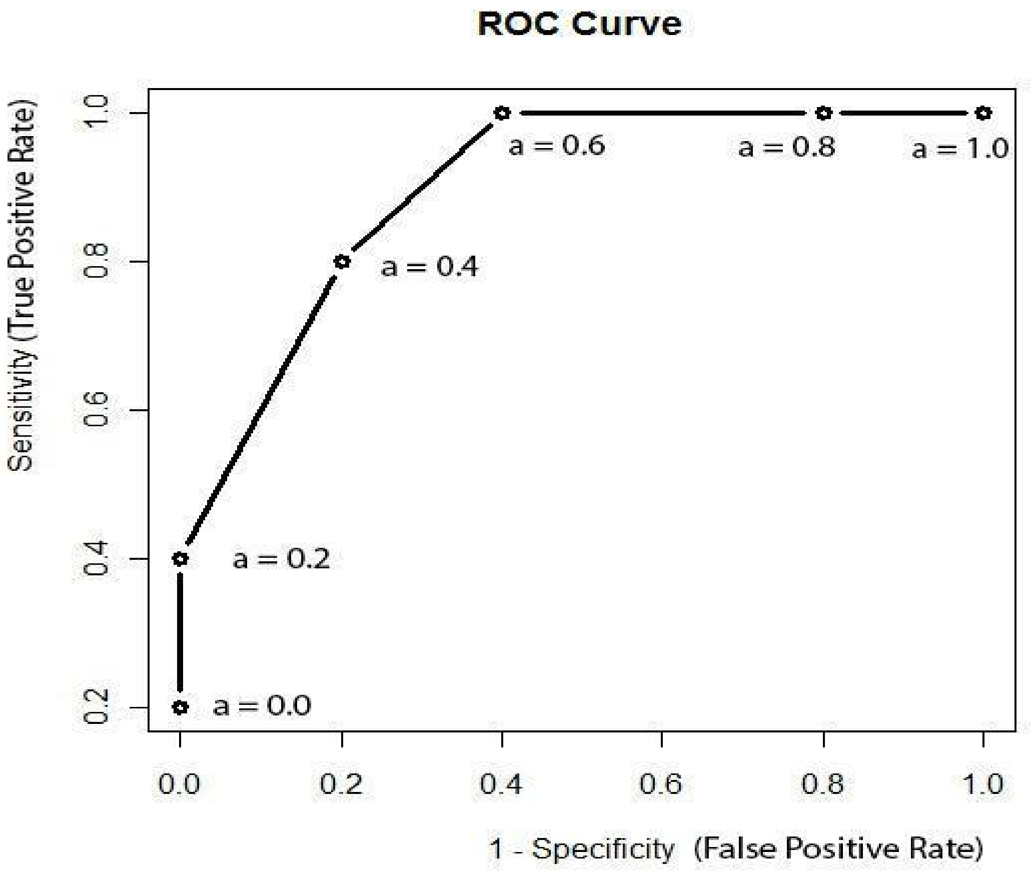
Question 3A 5 Marks

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

Question 3B 5 Marks

We are tasked with developing a classifier to predict sentiment towards a product. In this task, it is important to detect all expressions of positive sentiment whilst minimising the number of false detections.

Given the following ROC curve, what would be the best parameter choice, ‘a’, for this task:



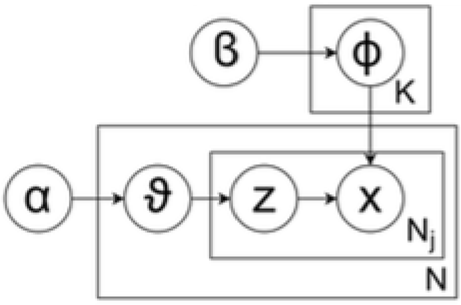
Section 3 is continued on the next page

Question 3C 10 Marks

Latent Dirichlet Allocation (LDA) is a Bayesian generative model. A process is provided that randomly generates latent variables and observed words, starting with Dirichlet priors for topic-word distributions and document-topic distributions. Training the model consists of finding estimates of the most probable values for unobserved (latent) variables.

Given the information below, briefly describe the generative process for LDA. For example, the first step could be described as "draw each θ_j from a dirichlet distribution with weight parameter α".

N	Number of documents in corpus
N_j	Length of j^{th} document
K	Number of topics
W	Number of unique tokens in corpus
α	K-dimensional weighting of topics
β	W-dimensional weighting of words
$\theta_j = (\theta_{1j} \dots \theta_{Kj})^T$	$\theta_j \sim Dir(\alpha)$
θ_{kj}	Likelihood of topic k in document j
$\phi_k = (\phi_{1k} \dots \phi_{Wk})^T$	$\phi_k \sim Dir(\beta)$
ϕ_{wk}	Likelihood of word w given topic k
z_{ij}	Topic for i^{th} position in j^{th} document
x_{ij}	Word for i^{th} position in j^{th} document



Section 4: Syntax - Part of Speech Tagging and Parsing

Instructions: Provide answers for questions 4A and 4B

Consider the following PCFG:

Rule	Probability
S → NP VP	1.0
NP → EX	0.8
NP → DT NN	0.2
VP → VB NP	0.4
VP → VB to VP	0.6
EX → there	1.0
VB → seems	0.3
VB → is	0.4
VB → be	0.3
DT → a	1.0
NN → problem	1.0

Question 4A 10 Marks

What is the probability of the sentence “there seems to be a problem”?

Question 4B 10 Marks

The grammar accepts the ungrammatical sentence *“there seems a problem”. How would you change the grammar to fix this? (The grammar should still accept “there is a problem”)

Section 5: Sentiment Analysis

Instructions: Provide answers for questions 5A, 5B and 5C

Given the following sentiment lexicon:

Word	**Part of Speech**	**Sentiment value**
small	Adjective	-1
tiny	Adjective	-2
correct	Verb	+1
more	Adjective	+1
good	Adjective	+2

And the following opinionated sentences:

- *Our room was tiny and the bath was small too.*
- *There was a lot of confusion around what my reservation entailed, I had to correct them about not having breakfast, not paying with reward points, and being the only one in the room.*
- *The breakfast room was so small that there was no room for more than 20 people having breakfast at the same time.*
- The *room service food was good.*

Consider the task of aspect-based sentiment analysis with the aspect 'bedroom'.

Question 5A 5 Marks

Given that Feature 1 is 'True' if the sentence refers to the given aspect, else 'False', provide the missing values for Feature 1 in the table below.

Sentence	**Feature 1**	**Feature 2**	**Feature 3**	**Aspect based sentiment**
#1	True			Negative
#2	?			Neutral
#3	?			Neutral
#4	?			Neutral

Section 5 is continued on the next page

Section 5 continued

Question 5B 10 Marks

Propose Feature 2, and Feature 3 , and fill their values in the table from Question 5A. These features can be based on the given sentiment lexicon.

Question 5C 5 Marks

Based on the data points, which feature do you think is most useful in detecting the sentiment towards the aspect *Bedroom*?

Section 6: Information Extraction

Instructions: Provide answers for questions 6A, 6B, 6C and 6D

Consider the following sentences about company mergers and acquisitions:

1. *LinkedIn announced the $1.5 billion acquisition of online education company Lynda.com.*
2. *We, however, are more interested in the announcement that LinkedIn is buying Connectifier, and to be more specific, why they are buying it.*
3. *Word on the streets of Israel is that Facebook is close to the acquisition of Pebbles Interfaces, a gesture control technology startup.*
4. *Facebook is buying a product search engine called TheFind to bring in new technology and talent to improve Facebook ads.*
5. *Apple revealed Thursday its acquisition of ed tech startup LearnSprout for an undisclosed amount.*
6. *Apple's acquisition of Emotient gives it access to technology that can measure a customer's emotions through facial recognition.*
7. *German industrial giant Siemens has hired Deutsche Bank to help it assess a possible acquisition of Spanish wind turbine maker Gamesa.*

Question 6A 5 Marks

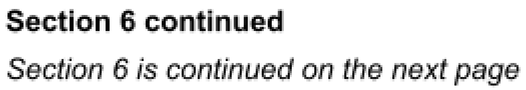
What are the 5 basic text processing steps of an information extraction system for extracting information about company mergers and acquisitions?

Question 6B 5 Marks

Give one example of the output for each of these steps when applied to one of the sentences provided above.

Question 6C 5 Marks

Give an example of pronominal coreference from the sentences above.



Section 6 continued

Section 6 is continued on the next page

Question 6D 5 Marks

Consider that only sentences 1-6 describe actually completed acquisitions.

What is the Precision, Recall and F-score of an information extraction system that makes use of the "*X * acquisition of * Y*" pattern to detect completed acquisitions, where X and Y are company names.

END