NUI Galway
OÉ Gaillimh

Autumn 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	Yes
External Examiner(s)	Professor Liam Maguire
Internal Examiner(s)	*Dr. Paul Buitelaar Dr. Georgeta Bordea Dr. John McCrae Dr. Ian Wood
Instructions:	Answer all questions. Use a separate answer book for each section.
Duration	2 hours
No. of Pages	4
Discipline(s)	Engineering and Information Technology
Course Co-ordinator(s)	Dr. Conor Hayes

Requirements:

Release in Exam Venue	Yes	✔	No	
MCQ	Yes		No	✔
Hand out	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 all the 4 Sections

Section 1: Linguistic Structure, Data and Analysis

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.

The church in the square was built in the twelfth century at the height of the Middle Ages.

Question 1B 10 Marks

Provide binary vectors for "*church"* and "*mosque"* 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 "*church"* or "*mosque"* occur. You should use morphological normalization such as inflection analysis.

The Paris church is very old.
The mosque is older than the city itself.
The city is very old, but the mosque is new.
The Notre Dame is a famous church in Paris.

Section 2: Probability and Classification

Question 2A 5 Marks

Consider the following table of probabilities, $p(X_{i+1} \mid X_i)$ represents the probability of X X_{i+1} appearing immediately after X_i and $p(X_i \mid \text{word})$ is the probability of a word belong to class X.

$p(X_{i+1} \mid X_i)$	X_{i+1} = Noun	X_{i+1} = Verb	X_{i+1} = Adj
X_i =Noun	0.3	0.6	0.1
X_i =Verb	0.6	0.2	0.2
X_i =Adj	0.8	0.1	0.1
X_i =Start	0.5	0.3	0.2

$p(X_i \mid word)$	X_i = Noun	X_i = Verb	X_i = Adj
I	0.8	0.1	0.1
love	0.5	0.4	0.1
Chinese	0.4	0.1	0.5
food	0.8	0.1	0.1

If you assume that this is a Hidden Markov Model, what is the probability of the sentence "I love Chinese food" given that the hidden states are (X_0=Start, X_1=Noun, X_2=Verb, X_3=Adj, X_4=Noun)

Question 2B 10 Marks

The diagram below gives the tableau of the Viterbi Algorithm for calculating the most likely sequence of hidden states for the sentence "I love food" for the Hidden Markov Model given above. The cells give the previous state that was most likely to transition into this state and the probability of this state. Complete this tableau for the last word "food".

	I	**love**	**food**
Noun	Start → 0.40	Noun → 0.06	
Verb	Start → 0.03	Noun → 0.096	
Adjective	Start → 0.02	Noun → 0.004	

Question 2C 5 Marks

Based on the tableau above, what is the most likely sequence of hidden states for the sentence "I love food".

Question 2D 5 Marks

Write down a bag of words representation for each of the following two sentences. Be sure to label each dimension with the corresponding word and to use the same representation scheme for both sentences.

The quick green dog
Green frogs on a dog

Question 2E 5 Marks

Calculate the cosine similarity of the two sentences above, given the bag of words representations from the previous answer.

Section 3: Model Evaluation

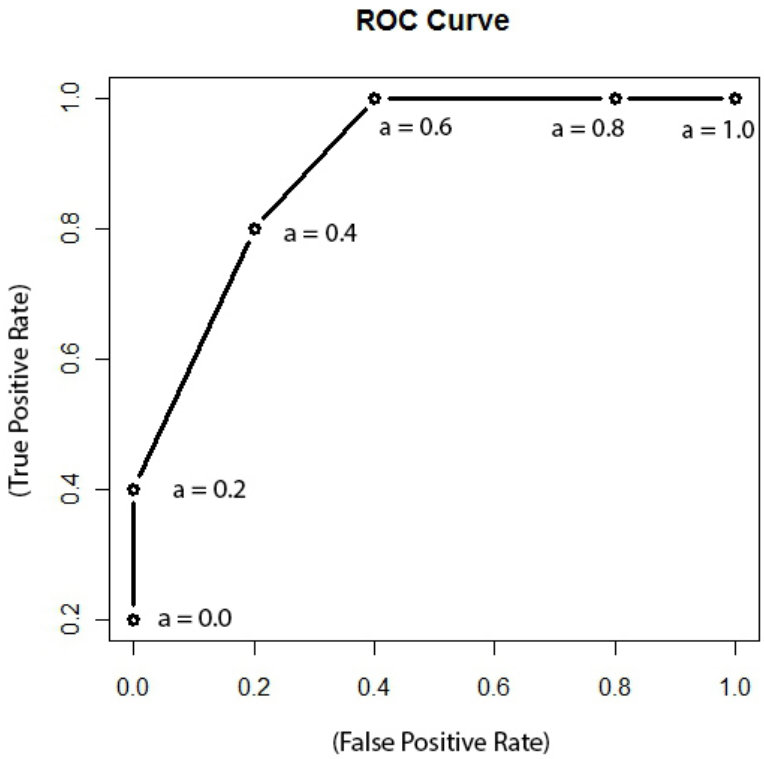
Question 3A 5 Marks

What is “overfitting” and how does it relate to the use of separate training and test data for supervised classification models?

Question 3B 5 Marks

We are tasked with developing a classifier to predict original homework. It is important in this task to detect all cases of plagiarism (negative cases) whilst maximising the number of accurate predictions of original work (positive cases).

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



Section 4: Information Extraction

Consider the following extract from an Irish Times article about the Rio Olympic Games:

Rory McIlroy will not compete in the upcoming Rio Olympics - with the risk of the Zika virus one he is “unwilling to take”.
The World number four golfer joins a growing list of golfers including Vijay Singh, Marc Leishman, Adam Scott, Louis Oosthuizen and Charl Schwartzel who have already said they will not feature in Rio.

On Wednesday morning, McIlroy released the following statement:

“After much thought and deliberation, I have decided to withdraw my name from consideration for this summer’s Olympic Games in Rio de Janeiro.’

Assuming that you want to evaluate a system for Named Entity Recognition that extracts the following named entities:

Rory McIlroy
Zika
The World
Vijay Singh
Marc Leishman
Adam Scott
Louis Oosthuizen
Charl Schwartzel
Olympic Games
Rio de Janeiro

And given that based on the gold standard annotated text the following entities should be extracted:

Rory McIlroy
Vijay Singh
Marc Leishman
Adam Scott
Louis Oosthuizen
Charl Schwartzel
McIlroy
Rio de Janeiro

Question 4A 5 Marks

What is the Precision, Recall and F-score of the Named Entity Recognition system?

Question 4B 5 Marks

Give an example of an extraction pattern that could be used to extract hyponyms for “golfers” from this news article.

Question 4C 5 Marks

Give one example of pronominal coreference from the news article above.

END