Cairo University
Faculty of Computers and Information

اسم الطالب

رقم الطالب

Final Exam

Department: Computer Sciences

معالجة اللغات الطبيعية

Course Title: Natural Language Processing

Course Code: CS462

Semester: Spring 2017/2018

Instructors: Dr. Hanaa Bayomi

Date: 2 Jun. 2018

Exam Duration: 2 Hours

Question	Mark	Signature
One		
Two		
Three		
Four		
Five		
Six		
Seven		
Eight		
Nine		
Total Marks		

60

Total Marks in Writing: _____

Question 5 : Mark each statement with T or F in the right side: (10 Ma

1. Stemming can be used for the purpose of keyword normalization, the process of converting a keyword into its base form.	(
2. Word2Vec model is a machine learning model used to create vector notations of text objects. Word2vec contains multiple deep neural networks	(
3. IR (information Retrieval) and IE (Information Extraction) are the two same thing	(
4. Given a stream of text, Named Entity Recognition determines which pronoun maps to which noun.	(
5. Given a sound clip of a person or people speaking, determine the textual representation of the speech. "Text to speech"	(
6. Parts-of-Speech tagging determines all part-of-speech for a specific word given as input	(
7. Parsing determines Parse Trees (Grammatical Analysis) for a given sentence.	(
8. Co reference Resolution is Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities")	(
9. Machine Translation Converts human language to machjne language	(
10. Natural Language generation is the main task of Natural language processing.	(

Question 6: the following is a probabilistic context-free grammar with start (

S → NP V NP (1.0)
NP → N (0.5) | A N (0.3) | NP N (0.2)
V → forecasts (0.2) | calm (0.2) | describe (0.2)
| warns (0.2) | predicts (0.2)
N → forecasts (0.4) | seas (0.4) | office (0.2)
A → calm (0.4) | stormy (0.5) | meteorological (0.1)

(a) Draw all possible parse trees for the sentence
"Meteorological office forecasts calm seas"
And calculate their probabilities, showing how your answers are derived.

c) What is the main problem of RNN and how to overcome it?

d) Define the main proposed architectures for generating Word2vec?

Question 3 Consider using a parser with the following: **(11 Marks**

(partial) grammar		(Partial) lexicon		
S -> NP VP	VP -> V	2012	N	from
NP -> Det N	VP -> VP PP	Switzerland	N	in
NP -> N	X -> VBP VBG	USA	N	increasing
VP -> X PP	NP -> NP PP	are	VBP	the
PP -> P N		exports	N	to
		exports	V	

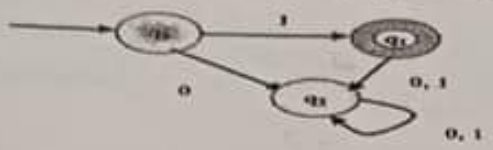
1) Using the CYK algorithm, parse the following sentence with the above lexicon/gra

"The exports from the USA to Switzerland are increasing in 2012 "

Provide both the complete, fully filled, data structure used by the algorithm, as wel as the result of the parsing in the form of a/the parse tree(s).

a) Convert a regular expression 1(0|1)*0 to the FSA

b) Write a regular expression for the language accepted by the following FSA



c) How many bi-grams can be generated from given sentence:

"Analytics Vidhya is a great source to learn data science"

d) In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times. What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "data" appears in approximately one-third of the total documents?

e) Write regular expressions for the following languages.

1) the set of all alphabetic strings

2) the set of all lower case alphabetic strings starting with b

3) the set of all strings with three consecutive repeated words (e.g., "Human Human Human" a "the the the")

4) Any string that starts with one lowercase character, and either ends with two digits or with th vowels

c) In a standard tf-idf information retrieval system, can a query retrieve a document that contain any of the query words? Justify your answer.

d) An information retrieval system with high precision and low recall can be useful for:
(a) Retrieving all relevant documents from a database of legal cases.
(b) Retrieving some interesting documents for a given a topic from the web.
(c) Retrieving a large set of interesting documents for a given a topic from the web.
(d) Checking the existence of a document in a very large document collection.

- **Choose all possible useful situations in the above list (maybe several).**

- **Define the notions of precision and recall.**

Question 2

(10 Marks)

a) What are the differences between RNN and Neural Network?

b) What are the main modules of LSTM?

2) Now suppose that we transform these two rules $X \rightarrow VBP\ VBG$ and $VP \rightarrow X\ PP$

To $VP \rightarrow VBP\ VBG\ PP$. Using the CYK algorithm, the modified grammar and lexicon to analyze the same sentence.