THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALLS

UNIVERSITY OF LONDON **CO3354 ZB**

BSc Examination

COMPUTING AND INFORMATION SYSTEMS, CREATIVE COMPUTING and COMBINED DEGREE SCHEME

Introduction to Natural Language Processing

Date and Time: Monday 16 May 2016: 14.30 – 16.45

Duration: 2 hours 15 minutes

There are FIVE questions on this paper. Candidates should answer THREE questions. All questions carry equal marks and full marks can be obtained for complete answers to THREE questions. The marks for each part of a question are indicated at the end of the part in [.] brackets.

Only your first THREE answers, in the order that they appear in your answer book, will be marked.

There are 75 marks available on this paper.

A handheld calculator may be used when answering questions on this paper but it must not be pre-programmed or able to display graphics, text or algebraic equations. The make and type of machine must be stated clearly on the front cover of the answer book.

QUESTION 1

a) The following sentences are all ambiguous in some way. Express their different meanings using paraphrases and explain the source of the ambiguity in each case:

i. John punched Bill and he fell over.
ii. Every man loves a woman.
iii. Mary watched the cat with one eye.
iv. Flying planes can be dangerous.

[6]

b) Explain what is meant by a **corpus** in the context of Natural Language Processing. **[2]**

What is meant by the following terms in the context of corpus linguistics? (One mark each for i-viii.)

i. Monitor corpus
ii. Parallel corpora
iii. Comparable corpora
iv. Treebank
v. Sampling frame
vi. Supervised learning
vii. Concordance
viii. Collocation

[8]

c) Annotate the text below with POS (part of speech) tags, using the simplified tagset given in Appendix A as in this example:

An/DET Indian/ADJ teenager/N scored/VD more/ADJ than/P 1,000/NUM runs/N in/P a/DET single/ADJ innings/N to/TO set/V a/DET new/ADJ world/N record/N in/P school/N cricket/N ./.

Where a word has more than one possible POS, explain how you have decided which one to use.

Text (BBC News Website, 5th January 2016):

"I was not thinking of a record," Dhanawade told BBC Hindi."It was not in my mind at all but as soon as I got close to the feat it was clear to me that I could achieve it."

Dhanawade said his father, an auto-rickshaw driver, had pushed him to play and was partly responsible for his success. He said he was ready for international cricket, but intended to first play in the under-19 state team.

[9]

QUESTION 2

Consider the following grammar:

S → NP VP
NP → Det N
NP → NP PP
VP → V
VP → V NP
VP → V NP PP

PP → P NP
Det → the | a | some | ε
N → teacher | student | classroom | books | table | pen | diagram
V → slept | put | saw | called | gave | has | have | wrote | drew | read
P → in | on | at | with | by | from | into

a)
i. Write out two grammatical sentences of at least 8 words which are generated by this set of rules.
ii. Draw syntax trees for both of these sentences, according to the grammar rules above. If your example is structurally ambiguous, you should give all applicable trees.

[9]

b) This grammar will generate sequences that are not grammatical sentences. Explain how it can be modified so that it will generate the grammatical examples (i-ii) below but not the ungrammatical (iii-iv).

i. The student slept in the classroom.
ii. The teacher put some books on the table.
iii. *The student slept the classroom.
iv. *The teacher put on the table.

[8]

c) Explain further how the grammar can be modified to generate examples like i – ii below:

i. The young student quickly read a little red book.
ii. The teacher usually sat at a big old wooden desk.

[8]

NB: you are not required to provide complete grammars in answer to (b) and (c), only to explain how the grammars need to be modified, giving appropriate worked examples. Marks will be awarded for compactness and generality of your solutions, not just whether they match the specific examples provided.

QUESTION 3

a) Briefly explain what is meant by each of the following in the context of Natural Language Processing:

i. N-gram tagging
ii. Conditional frequency distribution
iii. Stopwords
iv. Hidden Markov Model

[8]

b) Using the regular expressions supplied in Appendix B, describe and give examples to illustrate the classes of strings matched by the following regular expressions (for example, (ac)* matches ε, ac, acac ...):

i. a*b+
ii. (a*b)+
iii. [a-z]*|[0-9]*
iv. [a-z0-9]*

[9]

c) Suppose a corpus contains 300,000 word-tokens, and 70,000 of these are tagged as N (common noun). The word-form *house* occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that *house* accounts for 0.3% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *house* is tagged as N. Explain your answer and show your working.

[8]

QUESTION 4

The following paragraph is taken from a Goldsmiths press release dated 8th December 2015:

In May 2014 Goldsmiths held a nationwide competition to design an art gallery for our south east London campus, a venue that will run exhibitions, projects and residencies by leading artists and curators from the UK and abroad. Assemble's innovative design respectfully incorporates the black steel water tanks of the historic Laurie Grove baths, capitalising on the 'raw and robust' construction of the existing structure while building two additional steel frame structures which allow for a varied range of gallery space.

This is the result of running the above text through the Lancaster Stemmer:

```
['in', 'may', '2014', 'goldsmith', 'held', 'a', 'nationwid',
'competit', 'to', 'design', 'an', 'art', 'gallery', 'for', 'our',
'sou', 'east', 'london', 'camp', ',', 'a', 'venu', 'that', 'wil',
'run', 'exhibit', ',', 'project', 'and', 'resid', 'by', 'lead',
'art', 'and', 'cur', 'from', 'the', 'uk', 'and', 'abroad', '.',
'assemble's', 'innov', 'design', 'respect', 'incorp', 'the', 'black',
'steel', 'wat', 'tank', 'of', 'the', 'hist', 'laury', 'grov', 'bath',
',', 'capit', 'on', 'the', "'raw", 'and', 'robust', "'", 'construct',
'of', 'the', 'ex', 'structure', 'whil', 'build', 'two', 'addit',
'steel', 'fram', 'structures', 'which', 'allow', 'for', 'a', 'vary',
'rang', 'of', 'gallery', 'spac', '.']
```

And this is the result of running it through the Snowball Stemmer:

```
['in', 'may', '2014', 'goldsmith', 'held', 'a', 'nationwid',
'competit', 'to', 'design', 'an', 'art', 'galleri', 'for', 'our',
'south', 'east', 'london', 'campus', ',', 'a', 'venu', 'that',
'will', 'run', 'exhibit', ',', 'project', 'and', 'resid', 'by',
'lead', 'artist', 'and', 'curat', 'from', 'the', 'uk', 'and',
'abroad', '.', 'assembl', 'innov', 'design', 'respect', 'incorpor',
'the', 'black', 'steel', 'water', 'tank', 'of', 'the', 'histor',
'lauri', 'grove', 'bath', ',', 'capitalis', 'on', 'the', 'raw',
'and', 'robust', "'", 'construct', 'of', 'the', 'exist', 'structur',
'while', 'build', 'two', 'addit', 'steel', 'frame', 'structur',
'which', 'allow', 'for', 'a', 'vari', 'rang', 'of', 'galleri',
'space', '.']
```

a)

i. Explain what is meant by **text normalisation** in the context of natural language processing. What benefits derive from normalising text?

ii. Why is it recommended to use an "off-the-shelf" stemmer rather than coding your own using regular expressions?

[9]

b)

i. Make a list of rules which the Lancaster stemmer seems to have applied in this example and discuss the motivations for the rules.
ii. Make a similar list for the Snowball stemmer and note any cases where the two stemmers have different results.

[12]

c) Are there any cases where you think the rules were applied incorrectly? If so, give up to four examples. Justify your answer.

[4]

QUESTION 5

A probabilistic phrase structure grammar

Phrasal rules

S → NP VP	[1.0]
NP → Det N	[0.8]
NP → NP PP	[0.2]
VP → V NP	[0.9]
VP → VP PP	[0.1]
PP → P NP	[1.0]

Lexical rules

Det → a [0.5] | the [0.5]
N → equation [0.2] | formula [0.2] | teacher [0.2] | blackboard [0.2] | students [0.2]
V → copied [0.5] | explained [0.5]
P → on [1.0]

a) The rules shown above make up an example of a probabilistic or weighted grammar. What advantages can such grammars have over conventional phrase-structure grammars?

[6]

b) Using the probabilistic grammar rules and lexical rules given above, draw as many syntax trees as you can for the sentence:

"The students copied the equation on the blackboard".

Explain any ambiguities it may have by giving paraphrases for the meanings corresponding to the different syntactic analyses.

[9]

c)

i. Calculate the relative probabilities assigned to different analyses of the sentences by the grammar rules. You may omit the lexical probabilities as these make no difference to the outcome.

ii. Discuss whether the results agree with your intuitive understanding of the sentence.

[10]

APPENDIX A: NLTK SIMPLIFIED PART-OF-SPEECH TAGSET

ADJ	adjective	*new, good, high, special*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
EX	existential	*there, there's*
FW	foreign word	*dolce, ersatz, esprit, quo*
MOD	modal verb	*will, can, may, must*
N	noun	*year, home, costs, time*
NP	proper noun	*Adam, Paris*
NUM	number	*twenty-four, fourth, 1991, 14:24*
PRO	pronoun	*he, their, her, its, my, I, us*
P	preposition	*on, of, at, with, by, into, under*
TO	the word to	*to be or not to be*
UH	interjection	*ah, huh, oops,*
V	verb	*is, has, get, do, make, see, run*
VD	past tense	*said, took, told,*
VG	present participle	*making, going, playing*
VN	past participle	*given, taken, begun, sung*
WH	*wh*-determiner	*who, which, when, what, where, how*

APPENDIX B: REGULAR EXPRESSIONS

.	Wildcard, matches any character
^abc	Matches some pattern abc at the start of a string
abc$	Matches some pattern abc at the end of a string
[abc]	Matches one of a set of characters
[A-Z0-9]	Matches one of a range of characters
ed|ing|s	Matches one of the specified strings (disjunction)
*	Zero or more of previous item, e.g. a*, [a-z]*
+	One or more of previous item, e.g. a+, [a-z]+
?	Zero or one of the previous item (i.e. optional), e.g. a?, [a-z]?
a(b|c)+	Parentheses that indicate the scope of the operators

END OF PAPER