THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALLS

UNIVERSITY OF LONDON **CO3354 ZA**

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. Mary told Susan to finish her dinner.
ii. Flying drones can be dangerous.
iii. Every picture tells a story.
iv. The artist painted a picture on his back.

[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. Isolated corpus structure
ii. Temporal corpus structure
iii. Standoff annotation
iv. Treebank
v. Training and test sets
vi. Gold standard
vii. Inter-annotator agreement.
viii. Stopwords

[8]

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

A/DET wet/ADJ ,/, but/CNJ mild/ADJ ,/, December/NP was/VD a/DET record-breaking/ADJ month/N ,/, the/DET UK/NP Met/NP Office/NP final/ADJ figures/N show/V

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

Text (BBC News Website, 5 th January 2016):
It was the warmest December since records began in 1910 – and the wettest of any calendar month on record. Mean temperatures were about 4C (7.2F) above the long-term average. The Met Office says there is a direct link between the warmth and the record rains that brought widespread floods across Scotland, Northern Ireland and northern England.

[9]

QUESTION 2

Consider the following grammar:

S → NP VP
S → S Conj S
NP → Det N
NP → PN
NP → NP Conj NP
VP → V
VP → V NP
VP → V NP PP
PP → P NP

Det → the | a | some | ε
N → police | constable | constables | handcuffs | burglar | burglars | bed | telephone | gun | table | garden | knife | knives
PN → Holmes | Watson | Moriarty | Lestrade
V → slept | put | saw | called | gave | has | have | arrest | arrests | is | are
P → in | on | under | by
Conj → and | or

a)

i. Write out two grammatical sentences of at least 10 words which are generated by this set of rules.
ii. Draw syntax trees for both of these sentences, according to the grammar rules above.

[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 constable arrests a burglar.
ii. The knife is on the table by the bed.
iii. *The constable arrests a burglars.
iv. *The knives is on the table by the beds.

[8]

c) What problem could this grammar cause for a top-down parser? Explain how the rules could be modified to get round this problem.

[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.

QUESTION 3

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

i. Precision and recall
ii. Frequency distribution
iii. Sentiment analysis
iv. Normalising text

[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-z]|[0-9])+

[9]

c) Suppose a corpus contains 400,000 word-tokens, and 80,000 of these are tagged as N (common noun). The word-form *cook* occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that *cook* accounts for 0.4% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *cook* is tagged as N. Explain your answer and show your working. Show your final and intermediate results to no more than two decimal places.

[8]

QUESTION 4

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

“Assemble, the London based architecture collective chosen to design a new art gallery at Goldsmiths, University of London has been named winner of the Turner Prize 2015. Assemble were chosen to design our new art gallery. Building work begins next year. They are one of three Turner Prize 2015 nominees to have Goldsmiths connections – Bonnie Camplin is a Lecturer in Fine Art in the Department of Art, and Janice Kerbel both a Reader and a Fine Art graduate, having completed her MA with us in 1996.

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

```
['assembl', ',', 'the', 'london', 'bas', 'architect', 'collect',
'chos', 'to', 'design', 'a', 'new', 'art', 'gallery', 'at',
'goldsmith', ',', 'univers', 'of', 'london', 'has', 'been', 'nam',
'win', 'of', 'the', 'turn', 'priz', '2015', '.', 'assembl', 'wer',
'chos', 'to', 'design', 'our', 'new', 'art', 'gallery', '.', 'build',
'work', 'begin', 'next', 'year', '.', 'they', 'ar', 'on', 'of',
'three', 'turn', 'priz', '2015', 'nomin', 'to', 'hav', 'goldsmith',
'connect', '-', 'bonny', 'camplin', 'is', 'a', 'lect', 'in', 'fin',
'art', 'in', 'the', 'depart', 'of', 'art', ',', 'and', 'jan',
'kerbel', 'both', 'a', 'read', 'and', 'a', 'fin', 'art', 'gradu',
',', 'hav', 'complet', 'her', 'ma', 'with', 'us', 'in', '1996', '.']
```

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

```
['assembl', ',', 'the', 'london', 'base', 'architectur', 'collect',
'chosen', 'to', 'design', 'a', 'new', 'art', 'galleri', 'at',
'goldsmith', ',', 'univers', 'of', 'london', 'has', 'been', 'name',
'winner', 'of', 'the', 'turner', 'prize', '2015', '.', 'assembl',
'were', 'chosen', 'to', 'design', 'our', 'new', 'art', 'galleri',
'.', 'build', 'work', 'begin', 'next', 'year', '.', 'they', 'are',
'one', 'of', 'three', 'turner', 'prize', '2015', 'nomine', 'to',
'have', 'goldsmith', 'connect', '-', 'bonni', 'camplin', 'is', 'a',
'lectur', 'in', 'fine', 'art', 'in', 'the', 'depart', 'of', 'art',
',', 'and', 'janic', 'kerbel', 'both', 'a', 'reader', 'and', 'a',
'fine', 'art', 'graduat', ',', 'have', 'complet', 'her', 'ma',
'with', 'us', 'in', '1996', '.']
```

a)

i. Explain what is meant by **word stems**, with references to examples from the above text.

ii. Give an example of how stemming can be useful in real-world NLP applications.

iii. Explain the difference between a **stemmer** and a **lemmatizer**.

[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	[0.7]
NP → NP PP	[0.3]
VP → V NP	[0.8]
VP → V NP PP	[0.2]
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

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

END OF PAPER