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

Tuesday 15 May 2018: 14.30 – 16.45

Time allowed: 2 hours and 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**

Patrick Hayes and Kenneth Ford published a paper in 1995 with the title "Turing Test Considered Harmful", in which they argued that attempts to construct systems capable of unconstrained, natural-sounding conversation actually distracted from the proper concerns of AI, and that it would be more productive to shift the goal from "making artificial superhumans which can replace us, to making superhumanly intelligent artifacts which we can use to amplify and support our own cognitive abilities".

Write an essay considering and evaluating arguments for and against this stance, with reference to NLP technologies you have learned about from the course materials, recommended readings and your own independent study. Conclude by stating whether or not you agree with Hayes and Ford's position as summarised above, giving your reasons.

**[25]**

## QUESTION 2

The following paragraph is taken from the first chapter of Bram Stoker's novel Dracula (1897):

*Having had some time at my disposal when in London, I had visited the British Museum, and made search among the books and maps in the library regarding Transylvania; it had struck me that some foreknowledge of the country could hardly fail to have some importance in dealing with a nobleman of that country. I find that the district he named is in the extreme east of the country, just on the borders of three states, Transylvania, Moldavia and Bukovina, in the midst of the Carpathian mountains; one of the wildest and least known portions of Europe. I was not able to light on any map or work giving the exact locality of the Castle Dracula, as there are no maps of this country as yet to compare with our own Ordnance Survey maps; but I found that Bistritz, the post town named by Count Dracula, is a fairly well-known place.*

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

```
['have', 'had', 'some', 'time', 'at', 'my', 'dispos', 'when', 'in',
'london', ',', 'I', 'had', 'visit', 'the', 'british', 'museum', ',',
'and', 'made', 'search', 'among', 'the', 'book', 'and', 'map', 'in',
'the', 'librari', 'regard', 'transylvania', ';', 'it', 'had',
'struck', 'me', 'that', 'some', 'foreknowledg', 'of', 'the',
'countri', 'could', 'hardli', 'fail', 'to', 'have', 'some', 'import',
'in', 'deal', 'with', 'a', 'nobleman', 'of', 'that', 'countri', '.',
'I', 'find', 'that', 'the', 'district', 'he', 'name', 'is', 'in',
'the', 'extrem', 'east', 'of', 'the', 'countri', ',', 'just', 'on',
'the', 'border', 'of', 'three', 'state', ',', 'transylvania', ',',
'moldavia', 'and', 'bukovina', ',', 'in', 'the', 'midst', 'of',
'the', 'carpathian', 'mountain', ';', 'one', 'of', 'the', 'wildest',
'and', 'least', 'known', 'portion', 'of', 'europ', '.', 'I', 'wa',
'not', 'abl', 'to', 'light', 'on', 'ani', 'map', 'or', 'work',
'give', 'the', 'exact', 'local', 'of', 'the', 'castl', 'dracula',
',', 'as', 'there', 'are', 'no', 'map', 'of', 'thi', 'countri', 'as',
'yet', 'to', 'compar', 'with', 'our', 'own', 'ordnanc', 'survey',
'map', ';', 'but', 'I', 'found', 'that', 'bistritz', ',', 'the',
'post', 'town', 'name', 'by', 'count', 'dracula', ',', 'is', 'a',
'fairli', 'well-known', 'place', '.']
```

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

```
['hav', 'had', 'som', 'tim', 'at', 'my', 'dispos', 'when', 'in',
'london', ',', 'i', 'had', 'visit', 'the', 'brit', 'muse', ',',
'and', 'mad', 'search', 'among', 'the', 'book', 'and', 'map', 'in',
'the', 'libr', 'regard', 'transylvan', ';', 'it', 'had', 'struck',
'me', 'that', 'som', 'foreknowledg', 'of', 'the', 'country', 'could',
'hard', 'fail', 'to', 'hav', 'som', 'import', 'in', 'deal', 'with',
'a', 'noblem', 'of', 'that', 'country', '.', 'i', 'find', 'that',
'the', 'district', 'he', 'nam', 'is', 'in', 'the', 'extrem', 'east',
'of', 'the', 'country', ',', 'just', 'on', 'the', 'bord', 'of',
'three', 'stat', ',', 'transylvan', ',', 'moldav', 'and', 'bukovin',
',', 'in', 'the', 'midst', 'of', 'the', 'carpath', 'mountain', ';',
'on', 'of', 'the', 'wildest', 'and', 'least', 'known', 'port', 'of',
'europ', '.', 'i', 'was', 'not', 'abl', 'to', 'light', 'on', 'any',
'map', 'or', 'work', 'giv', 'the', 'exact', 'loc', 'of', 'the',
'castl', 'dracul', ',', 'as', 'ther', 'ar', 'no', 'map', 'of', 'thi',
'country', 'as', 'yet', 'to', 'comp', 'with', 'our', 'own', 'ordn',
'survey', 'map', ';', 'but', 'i', 'found', 'that', 'bistritz', ',',
'the', 'post', 'town', 'nam', 'by', 'count', 'dracul', ',', 'is',
'a', 'fair', 'well-known', 'plac', '.']
```

a)

i. Explain what is meant by **word stems**, with references to examples from the above text. **[2]**

ii. Give an example of how stemming can be useful in real-world NLP applications. **[3]**

iii. Explain the difference between a **stemmer** and a **lemmatizer**. **[2]**

b)

i. Make a list of rules which the Porter stemmer seems to have applied in the example, and discuss the motivations for the rules.

ii. Make a similar list for the Lancaster stemmer and note any cases where the two stemmers have different results.

**[7x2]**

c) Are there any cases where you think the results are not genuine stems? If so, give up to **FOUR** examples. Justify your answer.

**[4]**

## QUESTION 3

Consider the following grammar:

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

Det → the | a | his | her
N → bed | bedroom | trees | end | bird | garden
PN → Margaret
V → sat | saw
P → in | on | under | by | of | at

a)

i. Write out **TWO** sentences of at least 10 words but with different syntactic structures, which are grammatical and meaningful according to your knowledge of English and are generated by this set of rules.

**[2]**

ii. Draw syntax trees for both of these sentences, according to the grammar rules above. You should draw all applicable trees if your sentences are structurally ambiguous.

**[6]**

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

i. Margaret put her book on the table by the window.
ii. In her attic bedroom, Margaret sat on her bed and watched the trees.
iii. *Margaret put her book.
iv. *Margaret sat her bed.

**[12]**

c) What problem could this grammar (prior to your modification) cause for a recursive-descent parser? Explain how the rules could be modified to get around this problem.

**[5]**

**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, with appropriate worked examples.

## QUESTION 4

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

i. Constituent structure
ii. Tokenisation
iii. Sentiment analysis
iv. Chunking

**[4x2]**

b) Using the regular expressions supplied in the Appendix, 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. ([^aeiou]*|[0-9])

**[4x2]**

c) Suppose a novel contains 189,000 word-tokens, and 45,750 of these are tagged as N (common noun). The word-form *love* occurs 41 times in the novel, tagged either as N or V. Analysis shows that *love* accounts for 0.05% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *love* is tagged as N. Explain your answer and show your working. Show your final and intermediate results to no more than two significant figures.

**[9]**

## QUESTION 5

### A probabilistic phrase structure grammar

| | |
|---|---|
| S -> NP VP | [1.0] |
| VP -> VP Adv | [0.2] |
| VP -> VP 'and' VP | [0.2] |
| VP -> IV | [0.6] |
| NP -> 'Alicia' | [1.0] |
| IV -> 'sings' | [0.5] |
| IV -> 'dances' | [0.5] |
| Adv -> 'beautifully' | [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? Explain the purpose of the numbers in square brackets [.].

**[9]**

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

"Alicia sings and dances beautifully."

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

**[6]**

c)

i. Calculate the relative probabilities assigned to different analyses of the sentences by the grammar rules. Which analysis has the highest probability?

**[6]**

ii. Discuss whether the results agree with your intuitive understanding of the sentence. **[4]**

## APPENDIX: 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, <i>e.g.</i> a*, [a-z]*</td></tr>
<tr><td>+</td><td>One or more of previous item, <i>e.g.</i> a+, [a-z]+</td></tr>
<tr><td>?</td><td>Zero or one of the previous item (<i>i.e.</i> optional), <i>e.g.</i> 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**