UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

Department of Computing

BSc Examination 2016

IS53036A
Introduction to Natural Language Processing

Duration: 2 hours 15 minutes

Date and time: Tuesday 12th January, 14.30 – 16.45

There are FOUR questions in this paper. You must answer Question 1 and no more than TWO of Questions 2 to 4. Question 1 carries 40 marks and all other questions carry 30 marks each. The marks for each part of a question are indicated at the end of the part in [.] brackets.

There are 100 marks available on this paper.

A hand held 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.

**THIS PAPER MUST NOT BE REMOVED
FROM THE EXAMINATION ROOM**

QUESTION 1 (40 marks): you must attempt this question.

a) Explain what is meant by a **corpus** in the context of Natural Language Processing. What is meant by the following terms in the context of corpus linguistics?

i. Comparable corpora
ii. Treebank
iii. Training and test sets
iv. Gold standard **[8 marks]**

b) Using the regular expressions supplied in Appendix A, describe the classes of strings matched by the following regular expressions, giving examples (e.g., (ac)* matches ε, ac, acac ...):

i. a+b*
ii. (a+b)*
iii. a?(b|c)+

Write REEs that will match the following strings:

iv. ‘sing’, ‘bring, ‘belonging’ **but not** ‘singer’, ‘inga’, ‘things’
v. ‘last’, ‘lost’, ‘least’ **but not** ‘ballast’, ‘whilst’, ‘listing’

[10 marks]

c) Using the phrase-structure grammar provided in Appendix B:

i. Write out the shortest sentence generated by the grammar rules.
ii. Write out a grammatical sentence of at least 8 words which is generated by this set of rules.
iii. Draw syntax trees for both of these sentences, according to the grammar rules provided.

[6 marks]

d)

i. Explain the difference between a **grammar** and a **parser**.
ii. Briefly describe one example of **top-down** and one of **bottom-up** parsing regimes, using diagrams as appropriate.

[8 marks]

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

```
[('The', 'DET'), ('Beatles', 'NOUN'), ('were', 'VERB'), ('an', 'DET'), ('English', 'ADJ'),
('rock', 'NOUN'), ('band', 'NOUN'), (',', '.'), ('started', 'VERB'), ('in', 'ADP'), ('Liverpool',
'NOUN'), (',', '.'), ('England', 'NOUN'), ('in', 'ADP'), ('1960', 'NUM'), ('.', '.')]
```

Text to be annotated:

Starting in 1957, John Lennon and several of his friends played in a British band called the Quarrymen. Over the next few years, the members of the band changed, and by 1960, the band was called The Beatles. They first came to the United States in 1964.

(Example and text both from Wikipedia.)

[8 marks]

You should answer no more than two of the remaining questions.

QUESTION 2. Syntax and Parsing

a) Using the probabilistic grammar rules and lexical rules given in Appendix C, draw as many syntax trees as you can for the following sentence. Explain any ambiguities it may have by giving paraphrases for the meanings corresponding to the different syntactic analyses:

“The students copied the diagram on the whiteboard”.

[10 marks]

b) 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. Explain your answer and show your working.

[10 marks]

c) Explain how the phrase structure grammar shown in Appendix B can be modified so that it will generate examples (i-iv) below but not the ungrammatical (v-viii).

i. The artist painted a pretty little red flower.
ii. The student drew a portrait in ink.
iii. The artist holds a brush.
iv. The models have robes.
v. *The artist hold a brush.
vi. *The models has robes.
vii. *The artist slept a brush.
viii. *The artist painted a portrait with brush.

[10 marks]

QUESTION 3. Classification

a)

i. Explain what is meant by *sequence classification* in the context of Natural Language Processing, and describe a suitable application.

ii. Explain the difference between *supervised* and *unsupervised* learning.

iii. Explain what is meant by the *naïve Bayes assumption*.

[12 marks]

b) Suppose a corpus contains 320,000 words, and 80,000 of these are tagged as *NOUN*. The word-form "*can"* occurs 300 times in the corpus, tagged either as *NOUN* or *VERB* (main verb or modal auxiliary). Analysis shows that "*can"* accounts for 0.08% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of "*can"* is tagged as *NOUN*. Explain your answer and show your working.

[10 marks]

c) The Stanford Twitter Sentiment Corpus was constructed by using emoticons to classify tweets as positive or negative. For example, a tweet containing :-) would be classed as *positive*, while tweets that do not contain an emoticon are classed as *neutral*. Discuss some possible advantages and disadvantages of this approach.

[8 marks]

QUESTION 4. Information Extraction

a) Explain the difference between *information extraction* and *information retrieval*. Describe an example of a widely used information retrieval application.

[4 marks]

b) This question requires you to write a regular expression to recognise proper names of *people*, including titles such as *Ms*, *President*. You should assume you are dealing with tokenised text from a news domain (e.g. *Wall Street Journal*, London *Times*, BBC), and use the notation employed by the NLTK `findall` method: each token is indicated with angled brackets, as in this example which might be used to search for phone numbers:

```
r”<\+[0-9][0-9]><[0-9]+><[0-9]+>”
```

You should explain the content of your regular expression, using suitable examples. Explain any assumptions you have made about what kinds of expression can make up a name.

Would you expect this RE to give high or low results for *precision* and *recall*? Justify your answer.

[12 marks]

c) The following text is taken from a Goldsmiths press release published on 27 Oct 2015:

“Music has played a central role in human cultures for thousands of years. For most of us, music is emotional. It moves us. It’s known that stimulating music and low-energy music have different effects on the brain, but how does music affect the connection between our head (ie. the brain) and the heart, the popular seat of emotion? And how does that impact on emotions? To try and find out, researchers from Goldsmiths, University of London have been exploring how the brain responds to our heartbeat while listening to music.”

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

```
['Music', 'ha', 'play', 'a', 'central', 'role', 'in', 'human',
'cultur', 'for', 'thousand', 'of', 'year', '.', 'For', 'most',
'of', 'us', ',', 'music', 'is', 'emot', '.', 'It', 'move',
'us', '.', 'It’', 'known', 'that', 'stimul', 'music', 'and',
'low-energi', 'music', 'have', 'differ', 'effect', 'on',
'the', 'brain', ',', 'but', 'how', 'doe', 'music', 'affect',
'the', 'connect', 'between', 'our', 'head', '(', 'ie', '.',
'the', 'brain', ')', 'and', 'the', 'heart', ',', 'the',
'popular', 'seat', 'of', 'emot', '?', 'And', 'how', 'doe',
'that', 'impact', 'on', 'emot', '?', 'To', 'tri', 'and',
'find', 'out', ',', 'research', 'from', 'Goldsmith', ',',
'Univers', 'of', 'London', 'have', 'been', 'explor', 'how',
'the', 'brain', 'respond', 'to', 'our', 'heartbeat', 'while',
'listen', 'to', 'music', '.']
```

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

```
['mus', 'has', 'play', 'a', 'cent', 'rol', 'in', 'hum',
'cult', 'for', 'thousand', 'of', 'year', '.', 'for', 'most',
'of', 'us', ',', 'mus', 'is', 'emot', '.', 'it', 'mov', 'us',
'.', 'it’s', 'known', 'that', 'stim', 'mus', 'and', 'low-
energy', 'mus', 'hav', 'diff', 'effect', 'on', 'the', 'brain',
',', 'but', 'how', 'doe', 'mus', 'affect', 'the', 'connect',
'between', 'our', 'head', '(', 'ie', '.', 'the', 'brain', ')',
'and', 'the', 'heart', ',', 'the', 'popul', 'seat', 'of',
'emot', '?', 'and', 'how', 'doe', 'that', 'impact', 'on',
'emot', '?', 'to', 'try', 'and', 'find', 'out', ',',
'research', 'from', 'goldsmith', ',', 'univers', 'of',
'london', 'hav', 'been', 'expl', 'how', 'the', 'brain',
'respond', 'to', 'our', 'heartb', 'whil', 'list', 'to', 'mus',
'.']
```

i. List 5 cases where the two stemmers have made the same decision (not counting ‘stems’ which are identical to the original word in the text, such as ‘a’), and 5 where they have made different decisions.

ii. Describe the rules that the stemmers appear to have applied in each of the cases you have identified in (i).

iii. Is there any case where something has been wrongly identified as a ‘stem’? Justify your answer.

[14 marks]

APPENDIX A: REGULAR EXPRESSIONS

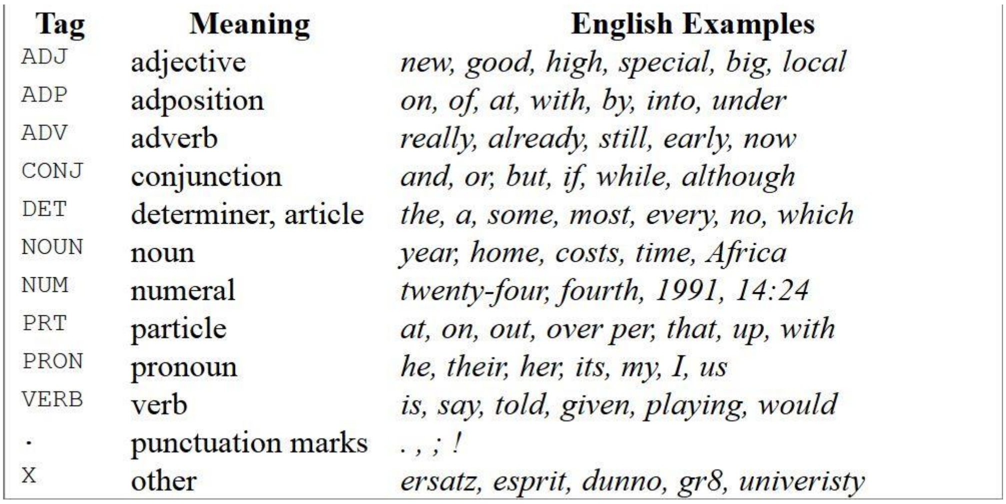
<table>
<tr><th></th><th></th></tr>
<tr><td>.</td><td>Wildcard, matches any character</td></tr>
<tr><td>[abc]</td><td>Matches one of a set of characters</td></tr>
<tr><td>[A-Z0-9]</td><td>Matches one of a range of characters</td></tr>
<tr><td>ed|ing|s</td><td>Matches one of the specified strings (disjunction)</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>
<tr><td>[^abc]</td><td>Caret inside square bracket indicates <i>negation,</i> otherwise:</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>
</table>

APPENDIX B: A phrase-structure grammar

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

Det → the | a | an
N → artist | model | paint | ink | brush | pencil
V → painted | drew | studied | sold | slept
P → in | on | with

APPENDIX C: A universal tagset



Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

From *Natural Language Processing with Python* by Steven Bird, Ewan Klein, and Edward Loper.

APPENDIX D: A probabilistic context-free grammar

Phrasal rules

S → NP VP	[1.0]
NP → Det N	[0.8]
NP → NP PP	[0.2]
VP → V NP	[0.7]
VP → V NP PP	[0.2]
VP → VP PP	[0.1]
PP → P NP	[1.0]

Lexical rules

Det → a [0.5] | the [0.5]
N → diagram [0.2] | formula [0.2] | lecturer [0.2] | whiteboard [0.2] | students [0.2]
V → copied [0.5] | explained [0.5]
P → on [1.0]

END OF EXAMINATION