UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

Department of Computing

BSc Examination 2017

IS53036A
Introduction to Natural Language Processing

Duration: 2 hours 15 minutes

Date and time: 10.00 – 12.15 13 January 2017

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 answer this question.

a) Explain why text generally has to be normalised before any other language processing. **[2 marks]**

Briefly describe the following:

i. Tokenisation
ii. Stemming
iii. Lemmatisation
iv. sentence segmentation **[6 marks]**

b) Using the regular expressions (REs) supplied in Appendix A, describe and give examples of 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+(b|c)?

Write REs that will match the following strings:

iv. ‘sink’, ‘brink’, ‘thinking’ **but not** ‘sinker’, ‘banking’, ‘king’
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 10 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 **parser** and a **recogniser**.

[2 marks]

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

[6 marks]

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

“Bob Dylan is an American songwriter, singer, artist, and writer.” is tagged as

```
[('Bob', 'NOUN'), ('Dylan', 'NOUN'), ('is', 'VERB'), ('an', 'DET'), ('American',
'ADJ'), ('songwriter', 'NOUN'), (',', '.'), ('singer', 'NOUN'), (',', '.'), ('artist',
'NOUN'), (',', '.'), ('and', 'CONJ'), ('writer', 'NOUN'), ('.', '.')]
```

Text to be annotated:

“Dylan's lyrics incorporate a wide range of political, social, philosophical, and literary influences. They defied existing pop music conventions and appealed to the burgeoning counterculture. Initially inspired by the performances of Little Richard and the songwriting of Woody Guthrie, Robert Johnson, and Hank Williams, Dylan has amplified and personalized musical genres.”

(Example and text both from Wikipedia.)

[8 marks]

You should answer <u>no more than two</u> 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.

“Jack shaves and bathes frequently”

[6 marks]

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

[4 marks]

b) Calculate the relative probabilities assigned to different analyses of the sentences by the grammar rules. 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-iii) below but not the ‘ungrammatical’ (iv-vi).

i. The artist slept on the bed.
ii. The artist and the model drank some tea.
iii. The artist drank some tea and the model slept.
iv. *The artist slept the bed.
v. *The artist and drank some tea.
vi. *And the model slept.

[10 marks]

QUESTION 3. Classification and Machine Learning

a)

i. Explain what is meant by a *Hidden Markov Model* in the context of Natural Language Processing. What is it that is “hidden”?

[5 marks]

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

[2 marks]

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

[5 marks]

b) Suppose a corpus contains 350,000 word-tokens, and 85,000 of these are tagged as N (common noun). The word-form *vote* occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that *vote* accounts for 0.5% of all common noun tokens in the corpus. Use Bayes’ formula to calculate the probability that a given occurrence of *vote* is tagged as N. Explain your answer and show your working.

[10 marks]

c) Describe two examples of supervised classification tasks apart from POS tagging. Discuss what kind of **features** of the input might be relevant in each case.

[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 (RE) to recognise *postal addresses*, such as *10 Downing Street, London SW1A 2AA* . 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.

[8 marks]

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

[4 marks]

c) The following text is taken from a Goldsmiths press release published on 27 Oct 2016:

“Researchers from the Department of Computing, in collaboration with Idiap Research Institute in Martigny, Switzerland, are developing a machine with the fluidity to reproduce graffiti tags in pen, paint and neon lights. The Baxter robot has been taught to write its own name in a stylised graphic form known as a tag - which first emerged in late-1960s New York - and is being taught to reproduce other graffiti tags using the same types of movements that a human would use for the skilled task. Drawing machines created to write, copy or generate art have existed for at least three centuries. Designed by Jacquet Droz (1721-90), one of the first drawing mechanical automatons could write sentences in cursive script.”

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

```
['Research', 'from', 'the', 'Depart', 'of', 'Comput', ',', 'in',
'collabor', 'with', 'Idiap', 'Research', 'Institut', 'in',
'Martigni', ',', 'Switzerland', ',', 'are', 'develop', 'a', 'machin',
'with', 'the', 'fluiditi', 'to', 'reproduc', 'graffiti', 'tag', 'in',
'pen', ',', 'paint', 'and', 'neon', 'light', '.', 'The', 'Baxter',
'robot', 'ha', 'been', 'taught', 'to', 'write', 'it', 'own', 'name',
'in', 'a', 'stylis', 'graphic', 'form', 'known', 'as', 'a', 'tag', '-
', 'which', 'first', 'emerg', 'in', 'late-1960', 'New', 'York', '-',
'and', 'is', 'be', 'taught', 'to', 'reproduc', 'other', 'graffiti',
'tag', 'use', 'the', 'same', 'type', 'of', 'movement', 'that', 'a',
'human', 'would', 'use', 'for', 'the', 'skill', 'task', '.', 'Draw',
'machin', 'creat', 'to', 'write', ',', 'copi', 'or', 'gener', 'art',
'have', 'exist', 'for', 'at', 'least', 'three', 'centuri', '.',
'Design', 'by', 'Jacquet', 'Droz', '(', '1721-90', ')', ',', 'one',
'of', 'the', 'first', 'draw', 'mechan', 'automaton', 'could',
'write', 'sentenc', 'in', 'cursiv', 'script', '.']
```

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

```
['research', 'from', 'the', 'depart', 'of', 'comput', ',', 'in',
'collab', 'with', 'idiap', 'research', 'institut', 'in', 'martigny',
',', 'switzerland', ',', 'ar', 'develop', 'a', 'machin', 'with',
'the', 'fluid', 'to', 'reproduc', 'graffit', 'tag', 'in', 'pen', ',',
'paint', 'and', 'neon', 'light', '.', 'the', 'baxt', 'robot', 'has',
'been', 'taught', 'to', 'writ', 'it', 'own', 'nam', 'in', 'a',
'styl', 'graph', 'form', 'known', 'as', 'a', 'tag', '-', 'which',
'first', 'emerg', 'in', 'late-1960s', 'new', 'york', '-', 'and',
'is', 'being', 'taught', 'to', 'reproduc', 'oth', 'graffit', 'tag',
'us', 'the', 'sam', 'typ', 'of', 'mov', 'that', 'a', 'hum', 'would',
'us', 'for', 'the', 'skil', 'task', '.', 'draw', 'machin', 'cre',
'to', 'writ', ',', 'cop', 'or', 'gen', 'art', 'hav', 'ex', 'for',
'at', 'least', 'three', 'century', '.', 'design', 'by', 'jacquet',
'droz', '(', '1721-90', ')', ',', 'on', 'of', 'the', 'first', 'draw',
'mech', 'automaton', 'could', 'writ', 'sent', 'in', 'curs', 'script',
'.']
```

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), and 5 where they have made different decisions. The latter may include cases where one stemmer has not removed anything from a word.

[4 marks]

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

[8 marks]

iii. Is there any case where you believe something has been wrongly identified as a 'stem'? Justify your answer.

[2 marks]

APPENDIX A: REGULAR EXPRESSIONS

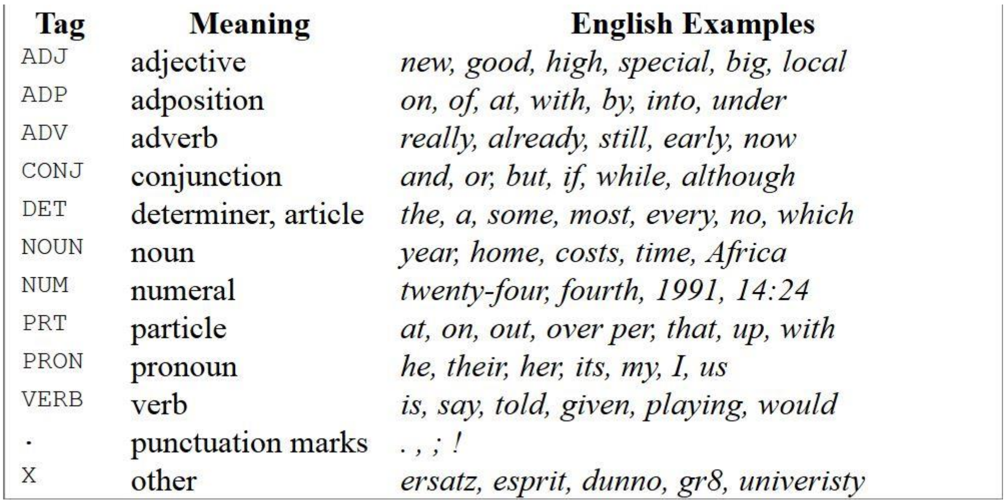
.	Wildcard, matches any character
[abc]	Matches one of a set of characters
[A-Z0-9]	Matches one of a range of characters
ed|ing|s	Matches one of the specified strings (disjunction)
*	Zero or more of previous item, e.g. a*, [a-z]*
+	One or more of previous item, e.g. a+, [a-z]+
?	Zero or one of the previous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

APPENDIX B: A phrase-structure grammar

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

Det → the | a | an
Adj → red | blue | tall | short | young | old | famous
N → artist | model | paint | ink | brush | pencil
V → painted | drew | studied | sold | slept
P → in | on | with

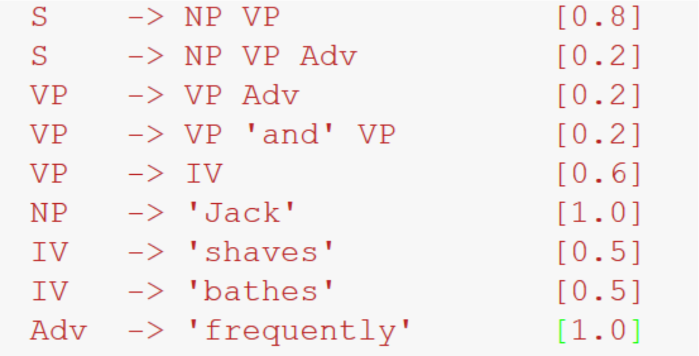
APPENDIX C: A universal tagset



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



```
S     -> NP VP                [0.8]
S     -> NP VP Adv            [0.2]
VP    -> VP Adv               [0.2]
VP    -> VP 'and' VP          [0.2]
VP    -> IV                   [0.6]
NP    -> 'Jack'               [1.0]
IV    -> 'shaves'             [0.5]
IV    -> 'bathes'             [0.5]
Adv   -> 'frequently'         [1.0]
```

END OF EXAMINATION