SECTION A
Candidates should answer ALL of Question 1 in Section A.

Question 1

(a) Which of the following statements are true of Zipf's law?
Select ALL statements that apply.

i. Word rank and word frequency are inversely related
ii. Word rank and word frequency are positively correlated
iii. It applies to many naturalistic phenomena
iv. It describes a power law relationship between word rank and word frequency

(b) Which of the following similarity measures are commonly used in document retrieval?
Choose ONE option.

i. Cosine similarity
ii. Euclidean distance
iii. Manhattan distance
iv. String edit distance

(c) Which of the following are training algorithms used by word2vec?
Select ALL statements that apply.

i. Continuous bag of words
ii. Skipgram
iii. Negative sampling
iv. Hierarchical softmax

(d) What is the perplexity of a string of random digits?
Choose ONE option.

i. 10
ii. It depends on the precise digits
iii. Less than 10
iv. Greater than 10

(e) Which of the following techniques is developed specifically for visualization of high dimensional data?
Choose ONE option.

i. Singular value decomposition
ii. Non-negative matrix factorization
iii. Principal components analysis
iv. T-distributed stochastic neighbor embedding

(f) How does lemmatization differ from stemming?
Select ALL statements that apply.

i. Stemming only works for regular verbs
ii. Lemmatization is informed by the linguistic context
iii. Stemming is a more crude, heuristic process
iv. Stemming requires access to a lexical database

(g) What are some of the shortcomings of traditional 'one-hot' vector encodings'?

Select ALL statements that apply.

i. They tend to be very short
ii. They tend to be relatively sparse
iii. They tend to be very dense
iv. They tend to contain many zero elements

(h) Which of the following would you use to show the correlation between true positive labels and predicted positive labels?
Choose ONE option.

i. Confusion matrix
ii. Probability density function
iii. Network diagram
iv. Bar chart

(i) Which of the following techniques can be used to identify groups of topically-related documents within a corpus?
Choose ONE option.

i. k-Means clustering
ii. k-nearest neighbour
iii. Hierarchical softmax
iv. Principal components analysis

(j) What would you use to measure the proportion of spam messages that are correctly predicted as spam by your classifier?
Choose ONE option.

i. Negative predictive value
ii. Precision
iii. Recall
iv. Accuracy

SECTION B

Candidates should answer any **2** questions from Section B.

Question 2

(a) What are disjunction, grouping and precedence for pattern matching in regular expressions? Give an example to illustrate each. [6]

(b) Why is pattern matching by regular expressions sometimes described as 'greedy'? How might you avoid this behaviour? [4].

(c) What is the motivation of using n-gram probabilities in Natural Language Processing? Outline how n-grams might be used for calculating probabilities in detecting spelling errors. Start by outlining how you would use n-grams for word prediction. [8]

(d) What is the role of the stemming process in NLP? With the help of two examples briefly discuss advantages and disadvantages of such a process. [8]

(e) Briefly discuss the problems a tokenizer might encounter when processing texts which contain a single quote character (‘). [4]

Question 3

(a) Imagine you are planning to write a natural language interface to communicate with your computer. [12]
Develop a simple context-free grammar (CFG) that accepts grammatically correct commands from a user such as:

“open a new browser window”

“go to the homepage”

“search for past NLP exams”

“exit”

“print the exam”

Ensure that ungrammatical commands such as the following are rejected by your grammar:

* *“go to”*

* *“open”*

* *“print exam”*

(b) Give an example of a sentence that has at least two plausible, semantically different syntactic analyses. Draw the corresponding dependency trees and explain the difference in meaning. [6]

(c) What is coordination? Why is it problematic in dependency parsing? How would you capture coordination in a dependency structure? [6]

(d) What is ellipsis? Why is it challenging to parse accurately? Give examples of different kinds of ellipsis. [6]

Question 4

(a) Suppose you have a text classifier that is overfitting to the training data. Describe three strategies to address this situation. [6]

(b) Briefly explain how you might use pointwise mutual information to measure the likelihood of two independent events co-occurring. How might you apply this to sentiment analysis? [6]

(c) Compare and contrast the main approaches to measuring lexical similarity. What are their relative strengths and weaknesses? [6]

(d) Suppose you are doing bag-of-words text classification on a document. The raw input is a single string containing the text of the entire document. Describe the process to convert the raw input to a feature vector. [6]

(e) Explain the difference between intrinsic and extrinsic evaluation. Give examples of each in the context of NLP applications. [6]