Introduction

This document contains all the questions for your first assessment of the Artificial Intelligence 2 module. All questions are compulsory, and you must complete all questions in a Jupyter Notebook file. This assessment is worth 15% of the total marks for the Artificial Intelligence 2 module. You have 3 hours to complete this assessment. You must submit your work through Blackboard in the **NLP CA 1 submission 2** link that can be found under the **Assignments & Tests** link on the left of the Blackboard screen.

I emailed you the name of your assigned text file that you will use to complete each task of this CA assessment. You can find the relevant text file inside the **Sample files** folder on Blackboard. **Important: You must use this text file to answer all questions for your CA. You will lose marks if you have not used the correct text file.**

Q1

Download and open the text file assigned to you from Blackboard. You may need to open the text file with utf-8 encoding. Count the number of pages in the text file, print a sample (first few lines) of the text file contents to the screen.

Q2

Create an NLP document object. Then using a loop, store all the sentences of the document object into an array. Choose one of the sentences from your text document and show the following information using an f-string:

    (a) Token
    (b) Token POS tag
    (c) Token dependency
    (d) Explanation of each token POS tag

Q3

Explain what is meant by a regular expression. Explain this concept using program code to count the words that start with a vowel (a, e, i, o, u) in the entire text document. Display the total number of occurrences of these vowels.

Q4

Explain the concept of POS tags. Then using a sentence of your choice from your assigned text document show the following information using a f-string:

    (a) Token
    (b) POS tag
    (c) Description of the tag

Q5

Explain what a stop word is in Natural Language Processing. Then demonstrate this concept using a loop to find all the stop words within your document object. Next, choose one of these stop words and remove it from the list of recognised stop words for your document object. Finally prove that the stop word has been removed by checking its stop word status using the relevant spacy command.

Q6

Explain what is meant by Named Entity Recognition. Then using a loop, store all the named entities in your document object into an array. Show all the named entities that are in your document object, as well as the general NER label and full description of each label.

Q7

Using the rules-based matching technique, demonstrate this concept to find the number of occurrences of the words **man** or **woman** in the entire text document. Show the number of occurrences of both words. Call the matcher a suitable name. Display the start and end positions of each matching word, as well as each match. Use an f-string to suitably format your display.

And show a total count of all occurrences of these words.

Q8

Using the phrase matching technique, demonstrate the concept of phrase matching using your assigned text document. Demonstrate how the phrase matching technique can find **all occurrences** of the words **laugh**, **cry**, **happy**, **sad** in your document object. For each matched phrase show the following information:

(a) Start position of the matched phrase
(b) End position of the matched phrase
(c) 5 words before and after the matched phrase

Format your output using an f-string.

Q9

Choose one sentence from your document object. Then explain the concept of lemmatization and demonstrate its use by implementing this technique on your sentence of choice.

Q10

Choose one sentence from your document object. Then demonstrate how the dependency visualiser called display can be used to show part-of-speech and dependency tags for this chosen sentence. Configure the display outputs including font type, colour, background colour, and distance of 90.

Important Information

Plagiarism will not be accepted and will result in an automatic mark of zero.

Due Date: Monday 23rd March at 17:00. You must submit your work through Blackboard using the relevant link. Submit your work as a Jupyter notebook file. A cover sheet must also be submitted with your jupyter notebook file.