UNIT 8 ASSIGNMENT

Natural Language Processing

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Unit 8 Written Portion

# Implementing NLP Tasks

Answer the questions below about natural language processing.

## Questions:

1. What is NLP? What are real-world applications of NLP?

|  |
| --- |
| Natural Language Processing (NLP) is a field of Artificial Intelligence concerned with how computers understand and interact with human language. NLP applications are sentiment analysis (determine the tone of the text), topic modeling, machine translation, text summarization, text extraction, text classification, text generation, etc. |

1. Why and how do we have to transform raw text data for NLP tasks? Provide some examples of commonly used techniques in the NLP pipeline.

|  |
| --- |
| NLP use some techniques to convert raw text data. Common techniques are removing punctuation, converting to lowercase, removing stop words such as “and”, “is”, and “the”, lemmatization (identify root meaning), tokenization(break text into smaller units), etc. |

1. What is TF-IDF? And how is it calculated?

|  |
| --- |
| Term Frequency Inverse Document Frequency(TF-IDF) use the term frequency within a doc divided by the document frequency to assess a word's importance. If the token appears a lot in a given document, importance to that document goes up, and vice versa. It helps identify keywords that distinguish a document from others. |

1. What is the difference between vectorizers and word embeddings?

|  |
| --- |
| Both are used to represent text data numerically. However, word embeddings seek to capture the meaning of word within a body of text, while vectorizers simply convert words into numbers based on some predetermined rules. For a small vocabulary full of high-frequency words, TF-IDF is a good choice, and for larger vocabularies full of low-frequency words, word embeddings are a good choice. |

1. What is the difference between a traditional neural network and a sequence-to sequence model? Why should sequence-to-sequence models be used in NLP? Explain the components of a sequence-to-sequence model.

|  |
| --- |
| Sequence-to-sequence models are special types of neural networks have been developed to deal with text data. Sequence-to-sequence models should be used in NLP because texts are highly dependent on the context. The components are encoder and decoder. Encoder processes the input sequence and captures its meaning, and decoders generates a new sequence based on the encoder’s output. |

1. Compare and contrast a deep averaging network to a recurrent neural network.

|  |
| --- |
| Deep Averaging Networks (DAN) and Recurrent Neural Networks (RNNs) both handle sequential data in NLP. DAN treat a sequence as a bag of words, averaging word embeddings to capture overall sentiment. It is simpler and faster than RNN but it ignores the ordering of the input text. On the other hand, RNNs process the sequence word-by-word, allowing them to capture the order and dependencies between words. |



*To submit this assignment, please refer to the instructions in the course*.