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 is the ability of a machine learning model to be able to handle text data. Applications of NLP include automated chatbot, translation services, email filters, search engine results |

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.

|  |
| --- |
| The reason is because there are a lot of filler words that are involved in raw text data that don’t contribute to the semantic meaning of the text itself. We use data cleaning/preprocessing techniques to address this issue such as lemmatization where we find the base word for different words, make n-grams (breaking up sentences into phrases) or removing stop words. After this, we typically vectorize the text data so it can be more easily interpreted by the model |

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

|  |
| --- |
| TF-IDF is a way of vectorizing our raw text data. It essentially scores the importance of a word given its context/document. If the word appears a lot in a specific document, its importance increases, however, if the word appears in many documents, its importance decreases. Then, given these “scores”, it converts the word/token into a numeric form |

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

|  |
| --- |
| Vectorizers mainly just convert words to their numeric forms, whereas word embeddings do something similar, but they have an added feature of maintaining the semantic meaning of the words when they are vectorized |

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.

|  |
| --- |
| Traditional NNs are not able to handle sequences of words as well as a sequence to sequence model. Traditional NNs are not able to very well handle the relationships between words in a sentence, whereas a sequence to sequence model does do a better job because it maintains the order in which each word is analyzed and retains the semantic meaning of the words. The components of a seq2seq model are an encoder and decoder. An encoder takes a sequence of words and converts it to its vectorized form and the decoder takes that vector and produces some output based on a given prompt/task |

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

|  |
| --- |
| Similarities:  Both vectorize word sequences  Differences:  DANs: Takes average of words to convert into vectors. Doesn't retain semantic meaning or ordering of the words. Better for capturing vector similarities and word sentiment since it just looks at the overall picture of the sentence  RNNs: sequentially inputs each word into each layer of the model. The order of the words matter, so it does a better job at capturing the relationships between words. Better suited for translating text for example |



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