UNIT 7 ASSIGNMENT

Use ML for Text Analysis

## 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 7 Written Portion

# Choosing Your Model

Answer the questions below about using text as data and word embedding.

## Questions:

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

|  |
| --- |
| Natural Language Processing focuses on the interaction between computers and human language. Its goal is to allow computers to generate, understand, and interpret human language. Examples of this are speech recognition and voice assistants such as Siri, Alexa, and Google Assistant. Another example is when it can be used for machine translation such as Google Translate. It can interpret typed text, images of written language, as well as spoken word. |

1. Why and how do we have to transform features for NLP tasks? Provide some examples of commonly used techniques.

|  |
| --- |
| Transforming features for NLP tasks allows you to process textual data into a numerical format for models to interpret. Some commonly used techniques for feature transformation are Tokenization, which splits text into individual words (tokens) that can be then analyzed. Another technique is Bag-of-Words, which represents text as a collection of unique words and uses their frequencies. Each document has a numerical vector and each dimension represents a specific word with the value indicating the frequency of the word in the document. Another technique is TF-IDF, which was used in the Unit 7 assignment to assign a weight of importance to each word in the document relative to its frequency. It captures the significance of each word in different documents in this way. |

1. Explain lemmatization and provide an example.

|  |
| --- |
| Lemmatization is used to reduce words to their base form, called lemma. This represents the meaning of the word without worrying about its grammatical form. An example of lemmatization is turning “The squirrels are climbing over the fences” to “The squirrel are climb over the fence”, where “squirrels” and “fences” are turned into the base forms of “squirrel” and “fence”. |

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

|  |
| --- |
| TF-IDF stands for Term Frequency Inverse Document Frequency. It is a numerical representation that measures the importance of a word in a document. Term Frequency measures the frequency of a term in a document, the higher the frequency of the word, the larger of a value is assigned to it. Inverse Document Frequency measures how rare a word is in the document, the IDF value quantifies how much information a term provides. TF is a raw score value while IDF is the logarithm of the total number of documents divided by the number of documents containing the term (word). Then, once these two values are found, the TF-IDF calculation is the multiplication of the two values. TF-IDF score then shows the importance of a term in the document and represents each document as a vector with each dimension corresponding to a term and the value in each dimension is the score of the term in the document. |

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

|  |
| --- |
| Vectorizers and word embeddings are two different approaches to representing textual data in a numerical format. Vectorizers convert individual words or documents into numerical vectors where each dimension represents a word or feature and their size depends on the number of unique words. An example of a vectorizer is TF-IDF and they treat words as independent units. Word embedding, on the other hand, generates dense vectors of fixed decisions and captures relationships between words (instead of treating them independently). They are learned through unsupervised training rather than predefined algorithms and can learn representations of similar words to those they have worked with before. Therefore, they can capture more nuanced information. |

1. What is the difference between a Neural Network and the other Supervised Learning models that you have implemented? When should neural networks be used?

|  |
| --- |
| Some supervised Learning models used in this program thus far are Linear Regression, Logistic Regression, and Decision Trees. Linear Regression differs from Neural Networks as it has a limited capacity in capturing complex nonlinear patterns, while Neural Networks can handle them. Logistic Regression is used for binary classification tasks and models the relationship between input features and the probability of them belonging to a specific class. Neural Networks can once again handle more complex classification problems. Lastly, Decision Trees can handle both classification and regression problems and are simple to interpret but cannot handle complete relationships. Neural Networks excel compared to all three of these Supervise Learning Models. Because they are more suited and can handle more complex relationships. They are well suited for large datasets and complex problems—tasks that require complex pattern recognition. |

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