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?

|  |
| --- |
| NLP is in short Natural Language Processing. It’s defined as a branch of AI that helps machines process and understand human language in speech and text form. In order for machines learning models to process words and blocks of text, the text needs to be transformed into numerical features without it losing its underlying meaning. This subfield of machine learning is important because several data collected to be analyzed is in the form of text not numbers. Some real-world applications of NLP is the automated translated services. |

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

|  |
| --- |
| We have to transform features for NLP tasks because most data collected is in the form of text not numbers. Therefore, what we should do is to first transform the text of the features to number features. Some common techniques to perform this is tokenization and vectorization. Some techniques to extract data from text are sentiment analysis, topic modeling, translation, etc. The strategies to map individual word tokens to a number are vectorization and embedding. |

1. Explain lemmatization and provide an example.

|  |
| --- |
| Lemmatization is when we take each word and convert it to some canonical form. This is used with different verb tenses, noun versions of a verb, possessive and plural forms of a noun. And the motivation behind this is to reduce or memory footprint by reducing the feature set size without hurting the model quality. An example of lemmatization is transforming playing, player, played to play. |

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

|  |
| --- |
| TF-IDF is in short for Term Frequency Inverse Document Frequency. It essentially builds on the count method. We use the term frequency within a doc divided by the document frequency. It’s calculated by starting with computing the frequency in which it appears across all documents. If the token appears a lot in a given document the importance of the document goes up. If the token appears a lot of other documents the importance to the document goes down. The process of TF-IDF is similar to what we see in model building. We first split the data, then import TfidVectorizer then we tokenize and pre-process all specified in the vectorizer, fit the vectorizer to build the word to feature dictionary, then transform the TF-DF logic and convert them from text to numeric data matrix. |

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

|  |
| --- |
| The difference between vectorizers and word embeddings is that vectorizers simply convert words into numbers based on some predetermined rules. Meanwhile word embedding seeks to capture the semantics of words. |

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?

|  |
| --- |
| The difference between a Neural Network and other supervised leaning models is that for something in lets say, linear model like Logistic Regression, it learns one weighted function or line that can be used as a model to make predictions. In neural networks it extends this by using **many** linear functions coupled with linear transformations of those function so that the final combination of them is a non-linear relationship. And overall, neural networks are used to model more complex patterns and relationships compared to other supervised models. Moreover, neural networks I s a composition of simple linear and non-linear transformations of input data. |

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