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 project 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 project 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.

*Complete each project part as you progress through the course. Wait to submit the project until all parts are complete. Begin your course project by completing Part One below. A submit button can be found on the final Course Project assignment page. Information about the grading rubric is available on any of the course project assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the project.*

Week 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 stands for Natural Language Problem. It is the class of ML and AI methods that deal with text as data. Some real world applications of NLP are automated language translation, chatbots and topical classification of news articles. |
| --- |

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

| Text needs to be converted to a structured numeric representation to be compatible with machine learning algorithms. Two approaches to do this are Vectorization, which maps a string token to a feature id and uses one of a few numeric representations. The other is Embedding, which maps a string token to a fixed length vector of numbers. |
| --- |

1. Explain lemmatization and provide an example.

| Lemmatization is the process of converting a piece of text to a more canonical form, usually as a means for dimensionality reduction. One example is the conversion of words like ‘reading’ and ‘reader’ to ‘read.’ |
| --- |

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

| TF-IDF stands for term-frequency, inverse-document-frequency. This is one method for converting text to a numeric representation. The TF component counts the frequency by which a word appears in a given example. The IDF is the inverse of the number of examples in which the word appears. The simplest variation is to multiple the above two components. There are variations however where the log of one or both components are used. |
| --- |

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

| Vectorization maps a string token to a feature id and uses one of a few numeric representations. Embedding maps a string token to a fixed length vector of numbers. |
| --- |

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?

| Neural Networks can be used like any other standard supervised learning tasks for either classification or regression problems. Generally one should use NN’s when one expects that the modeling problem is highly non-linear and requires more powerful methods. One should also consider how much data is available, as NN’s require a lot of data to be trained effectively. |
| --- |



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