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 standardizes and analyzes text at large scale. There is a lot of data collected by companies that comes in the form of text.   * **Spam** filter > automatically scanned using text classification and keyword extraction. And so then, after that, based on the patterns that the NLP finds in those emails, it classifies them as a high chance for spam versus not. * **Search engine** > no longer use just keywords to help you understand, or search your results, or figure out and reach your results. In fact now, they take your search queries, they analyze your intent when you search for information, and then through NLP, they extract and provide the best results * **Language translation** > using NLP to translate languages way more accurately and present grammatically correct results to you. * **Recommendation** systems use NLP for recommending products and services * **Question-and-answering systems**, **chatbots**, and **automated** **customer** **support** can be used to answer questions for millions of customers at the same time * **product categorization**; that is a hierarchical structure of relationships or taxonomies between different products |
| --- |

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

| Feature transformation converts raw text data into numeric data. It separates the raw text and stop-words for meaningful features. With Data preparation we clean and preprocess text to reduce noisy typos, remove unimportant words and phrases, organize documents etc. Then you can remove stop-words. These are words that don’t carry much meaning. With stemming and lemmatization, you can reduce words to their base forms. Word Embeddings capture semantic relationships between words. TF-IDF weights a document based on frequency of a token in the document. |
| --- |

1. Explain lemmatization and provide an example.

| Stemming simply reduces the word to its active singular tense. But Lemmatization takes the word back to its root EX. "The wolfs are running quickly." >> “The wolf be run quick."   * + Playing,player,played => play   + Cats, cats', cats => cats   + am, are, is => be |
| --- |

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

| Term Frequency Inverse Document Frequency (TF-IDF) measures the importance of a word in a specific document.   * If the token appears a lot in the given document >> importance to that document goes up * If the token also appears in a lot of other documents >> importance to that document goes down |
| --- |

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

| Vectorizers and word embeddings both represent text data in a numerical format.  Vectorizers are algorithms that convert text data into numerical data by frequency of words in a given doc. Each vector represents a unique token from the vocabulary. The dimension represents the frequency of that word in the document. Each word is treated independently.  Word embeddings capture semantic relationships. They use neural network-based language models like Word2Vec. Word embeddings use **Cosine similiarty** to compute similarity between two vectors. The same word can have different vector representations depending on its context in the sentence |
| --- |

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 and other models differ in their architecture, complexity, and algorithm approach.  Neural networks consist of interconnected nodes organized in layers. Information flows through these layers, and each node applies an activation function to its input to produce an output.  linear regression, logistic regression, decision trees are simpler and are usually computationally faster. However, they might not be able to handle complex data.  Neural networks use back propagation and gradient descent and forward propagation. This adjusts the weights of the network to minimize the difference between predicted and actual outputs.  You use NN when the data has complex patterns. For Image and Speech Recognition, NLP, a lot of audio or text data. Neural networks often require large amounts of data but they can be computationally intensive. They’re also considered black-box models cause it’s hard to interpret whats happening inside. |
| --- |



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