Computer Science Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of these areas so that you will have a complete and realistic overview of your project. Your course instructor cannot sign off on your project topic without this information.

Note: You must fill out and submit this form. Space beneath each number will expand as needed.

Any cost associated with developing the application will be the responsibility of the student.

INFORM INSTRUCTOR:

Potential use of proprietary company information: (Y/N)

ANALYSIS:

1. Project topic AND description:
   My project topic is going to be using Natural Language Processing (NLP) to perform analysis on emails received from various sources. I intend to analyze the words in the email and determine whether the email is spam or not. The expected outcomes will be *not spam* or *spam*. This will be done by using neural networks to create a program that can teach itself to identify word patterns that are indicative of spam.

   Client:
   The client for this product will be *Dungeons Inc.* The company is responsible for designing, developing, and maintaining dungeons for many wizards, witches, dragons, and liches. Because of the nature of their business, their records are always kept secure. A weak point has been identified in their email system. This program seeks to cover that vulnerability and filter out any spam or malicious emails being sent to the employees.

2. Project purpose/goals:
   The purpose of this project will be to showcase my ability to create an email spam filter. Email spam filters are used throughout most fields of work for the purpose of securing the employee accounts and company data. The goal of this project is to develop a working program that can receive data, process it, and return information to the user that is beneficial to business operations.

3. Descriptive method:
   This project will use the *K-means clustering* algorithm as a method for learning from the input data. This will allow the program to train and recognize similar patterns and sentences that are more common in tandem with another.

4. Predictive/Prescriptive method:
   For the predictive method I have chosen to use the *Naïve Bayes Classification* algorithm to calculate the likelihood that a given email is spam. This will create a model that for every word found in the message it will calculate the probability of that word being found in spam. This is determined by the frequency of each word in emails pre-marked as spam in our data set.

DESIGN and DEVELOPMENT:

1. Computer science application type (select one):
   - Mobile (indicate Apple or Android)
   - Web
   - Stand-Alone

2. Programming/development language(s) you will use:
   Development will be done with Python using data analytics tools and libraries (matplotlib, pandas, numexpr, etc.)
   The core algorithm will be written in Jupyter Notebook.

3. Operating System(s)/Platform(s) you will use:
   Windows10

4. Database Management System you will use:
   N/A

5. Estimated number of hours for the following:
   i. Planning and Design: 10
   ii. Development: 20
   iii. Documentation: 5
   iv. Total: 35

6. Projected completion date:
   November 20, 2022

IMPLEMENTATION and EVALUATION:

1. Describe how you will approach the execution of your project:
   1. Acquire dataset from Kaggle.com
   2. Familiarize myself with and sanitize the data.
   3. Create a model with the data
   4. Train the model
   5. Evaluate
   6. Determine the efficacy of the program. If I am unhappy with the results, I will go back to step 3 and start again.
   7. Create documentation and graphics for data.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

STUDENT SIGNATURE

By signing and submitting this form, you acknowledge any cost associated with development and execution of the application will be your (the student) responsibility.

COURSE INSTRUCTOR'S NAME:

_____

COURSE INSTRUCTOR APPROVAL DATE: 11/15/2022