Capstone Project Proposal Template

Notes:

- This should take no more than one hour to complete – the clearer you are about the business problem you're working to solve with your ML-driven solution, the easier your proposal will be to complete
- This will be uploaded to your repo, which will be a part of your final submission
- Due date for submission is 12/9

Instructions:

1. Download this document as a Word Doc
2. Answer each question using a few sentences, at most
3. Save your completed proposal as a PDF
4. Create a project GitHub repo (if you have yet to do so)
5. Add your instructor as a collaborator (username nickmccarty) to your project repo
6. Add your mentor as a collaborator
7. Push your proposal PDF (created in Step 3) up to your repo
8. Copy the URL corresponding to the location of the PDF in your repo
9. Submit the copied URL using this link

[project name]

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?

  The business problem trying to be solved is the automation of written scenario overviews based on given parameters in the format of tables with words (x rows and y columns) in a Microsoft Word document.

- What industry/realm/domain does this apply to?

  This applies to any industry that needs scenario overviews done for any testing process. (Integration testing, System Acceptance Testing, etc.)

- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)

  The main motivation behind this project is to find a way to apply Artificial Intelligence to my current project to increase efficiency and accuracy.

Data Understanding

- What data will you collect?

  The data that will be collected will be mostly that of any component/module in an official government or licensing authority type of environment.

- Is there a plan for how to get the data (API request, direct download, etc.)?

  The plan as of now is to forge the data as needed to precisely fit the NLP model that will be implemented. However, Kaggle, Google and Amazon all offer datasets that need to be further explored to see if they fit our model's criteria.

- Are the features that will be used described clearly?

  The features that will be used are described very clearly. The features include data sorting, data cross referencing, and data message forwarding.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?

  Removing stop words, changing text case, removing punctuation, and adding additional pretext based on certain trigger words. There may be many more data preparation techniques that might need to be implemented.

- What are some of the cleaning/pre-processing challenges for this data?

  Some of the cleaning/pre-processing challenges for this data would be creating a separate database that would be linked to the data for certain events like pretext formulation. Another cleaning/preprocessing challenge will be finding the correct format for the data for it to fit the modeling techniques that are most common in NLP's.

Modeling

- What modeling techniques are most appropriate for your problem?

  There are two modeling techniques that are most appropriate for this problem. They are both text summarization methods. The first is extractive text summarization. The second is in contrast abstractive text summarization.

  Other potential techniques that could potentially fit for this project are sentiment analysis, named entity recognition, topic modeling, text classification, keyword extraction and lemmatization/stemming.

- What is your target variable? (Remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)

  The target variable in this problem is the summary sequence (which should be the y variable).

- Is this a regression or classification problem?

  This will most likely be a classification problem.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?

  The potential metrics for this project that could be used to determine success are accuracy, precision, recall, F1 Score, AUC, MRR, MAP, RMSE, MAPE, BLEU, METEOR, ROUGE and Perplexity.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?

  The most basic model that could be used for this project is decision tree. Although other supervised machine learning algorithms that could possibly be used for this project are support vector machines, Bayesian networks, maximum entropy, conditional random field, and neural networks/deep learning.

  Other models potentially include Encoder-Decoder used for training and the other could be the Inference model which is used to generate predictions be leveraging some of the layers from the trained model.