Capstone Proposal

The problem I want to solve or work on is a question and answer algorithm of the Stanford University Question and Answer Dataset (SQuAD).  Question and answering is a hot topic in Natural Language Processing (NLP) and will require research on the latest state of the art deep learning methods for NLP.

The SQuAD data set has over 100,000 questions from approximately 19,000 paragraphs of text. The data can be downloaded from the following link [The Stanford Question Answering Dataset (rajpurkar.github.io)](https://rajpurkar.github.io/SQuAD-explorer/).

This problem is a supervised learning problem where the model will be trained on questions and paragraphs as inputs and the answer locations as the labels.  The text will have to be preprocessed into tokens and word embeddings.  Spark will be used to do as much of the preprocessing as possible and falling back to Python libraries when needed.  Preprocessed data can be stored into a database in a cloud platform or locally.  The necessary post and get methods can be developed to do this like it was an API.  Tensorflow and/or Pytorch will be used to build out the different layers to train and test the model.  Layers will probably include a word embedding layer, transformer layer, and traditional layers. 

The final deliverable will consist of the preprocessing script(s) or notebooks, any code used to work with the database/API, scripts with the Tensorflow or Pytorch training, testing, and predicting code, and other final deployment requirements my mentor advises me on. How the project will be wrapped up for deployment is still in planning.  All of the code will be stored on the project github repository.

Given that question and answering algorithms use deep learning methods it will take some significant computing resources on a cloud environment to train the model.  I am not sure how much but it will be good experience to work with the cloud resources.