Question Answering System

Capstone Project 2 Proposal

Background

Currently, NLP and Deep learning are two very popular topics in the field of Data Science and are of great interest to me. Given the guidelines/freedoms of the second Capstone Project, I decided it best to focus less on a specific 'real-world' client/idea and more so use the project to further develop as a Data Scientist, and more specifically, develop my proficiency in the fields of Deep Learning and NLP both theoretically and practically.

The Idea/Motivation

Sentiment Analysis, Document Summarisation and Question Answering are three topics within NLP that I found very interesting.

-Sentiment Analysis is probably the most 'boring' choice of the three and perhaps the most easy to implement.

-Document Summarisation (of perhaps the GPDR documents) is clearly very important, especially so in the age of short attention spans and copious amount of data of all kinds, however potentially too time consuming for a Capstone Project.

-Question Answering is a field of Machine Learning/NLP where a system can answer questions posed by humans in natural language. This is also very important for the same reasons as Document Summarisation and obviously multiple others that the reader can think of.

I chose the *'Question Answering System'* for my second Capstone Project for the motivating reasons outlined in the 'Background' section as well as the fact that a rich dataset is already available and can be found at https://rajpurkar.github.io/SQuAD-explorer/. Furthermore, the GitHub link contains a leaderboard of the best performing Question Answering models in the world, and it will be a nice benchmark to compare my models to.

The Problem, the Client, the Outcome

In short, there is lots data of all kinds (even this document contains too much textual information for my liking), and so it would be nice to get answers quickly and accurately. Thus, in brief, the goal of this Capstone is to develop a Questions Answering model that will accurately and quickly provide answers to questions asked of it. Clearly, the model can only answer questions of information that was shown to it during training, however having had experience building high performing models for this dataset, a similar methodology could be applied (to some degree) to other datasets and thus (potentially) alot of people (the client) could have a lot of their questions answered quickly (the outcome).

The Approach/How am I going to do this

Currently, in short, it is not clear yet. Given that the major part of this Capstone is the Machine Learning part, there will be a lot of trial and error, tinkering and trying out multiple algorithms/methods. The data acquisition and data wrangling of this project are largely trivial; the dataset is a .json file of Wikipedia articles and questions asked. The details of EDA will be worked out in the process.

Deliverables

Python files/Jupyter Notebooks including detailed documentation and programming choice justification, a video/article describing the project in detail.