qanda

Question Answering Chatbot created for the purpose of the Lincode Hackathon

Dataset

To construct the dataset

```
 cd data

wget https://rajpurkar.github.io/SQuAD-explorer/dataset/train-v2.0.json -O data/train-v2.0.json

wget https://rajpurkar.github.io/SQuAD-explorer/dataset/dev-v2.0.json -O data/dev-v2.0.json
```

Running the data/DataExploration.ipynb notebook

Warning: The file is very large(1.5gb) and loading the w2vec model consumes a lot of memory

Outside the qanda directory download and extract the W2Vec model using

```
 cd qanda/../

wget -c "https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz"
```

Go back into the qanda repo and start Jupyter Notebook

```
 cd qanda/data/

jupyter-notebook
```

Results of the notebook

Referring to DataExploration.pdf, we can see that a similarity score approach is not very useful as only a ~57% accuracy was achieved in identifying the correct line. Hence we move towards Deep Learning approaches

Identifying the model

On visiting the website https://rajpurkar.github.io/SQuAD-explorer/ we noticed that a majority of the models in the top rankings made use of transformer based architecture. We also briefly looked into BiDAF models (the state of the art circa 2018) and decided that transformer based models were the better choice.

Comparing the Python libraries "transformers" (by HuggingFace) and "allennlp" (by AllenAI) we found the pretrained models from transformers easier to use and integrate into the application.

Of the multitude of pretrained models available in the transformers library we choose to use the "distilbert-base-uncased-distilled-squad" as it was one of the smaller models in the top 5 most downloaded models (keeping the constraint of speed in mind)

Running the Flask App

To start the server

```
 cd quanda
export FLASK_APP=application.py
flask run
```