This folder contains the files for the project titled -
Question Similarity Predictions using Siamese LSTM Networks by Akila Jeeson Daniel.

The scripts are in IPython Notebooks run on Python 2.7. The modules used in the scripts are - timeit, pandas, re, json, csv, numpy, itertools, nltk, keras (with Tensorflow), __future__ [for print function), sklearn, gensim, matplotlib.

The input file is given in the Input folder - questions.csv which contains the data from the Kaggle competition Quora Pairs dataset.

You need to download word vector embedding files and unzipit to their respective folders in submission folder - Word2Vec - 'GoogleNews-vectors-negative300.bin.gz' [https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing] [3.64 GB uncompressed]
and 'GloVe.6B' [http://nlp.stanford.edu/data/glove.6B.zip] [862.2 MB zipped]

The scripts are in the 'scripts folder'. The data_cleaning.ipynb cleans the questions.csv file and saves the data to data_clean.csv. This script takes significant amount of time to run for the whole file. Please try with a subset.

The Benchmark_TFIDF.ipynb file contains the script for the benchmark model. This file takes about an hour to run for 50000 question pairs.

The data_modelling_rnn.ipynb file contains the script for the final RNN model. The first rnn run in the main() is for the best fit model. 50000 question pairs run takes about 30 mins on a GPU.

The .html files contains a snapshot of a run of each of the scripts.