wz323
Data Science HW 4

Project Option 1: Deep Learning and Natural Language Processing

Part 1: Run the example program in TensorFlow and analyze the output

Code for this part: tensorflow.ipynb, data set automatically downloaded.

```
Training Step: 7039  | total loss: 0.08428 | time: 132.231s
| Adam | epoch: 010 | loss: 0.08428 - acc: 0.9846 -- iter: 22496/22500
Training Step: 7040  | total loss: 0.08491 | time: 136.932s
| Adam | epoch: 010 | loss: 0.08491 - acc: 0.9830 | val_loss: 0.77709 - va
l_acc: 0.8008 -- iter: 22500/22500
--
```

By observation, the accuracy of running the example is fluctuating and generally increasing. The final accuracy is around 98%.

Part 2: Pick a traditional non-deep learning machine learning algorithm available in Spark (Naïve Bayes, Random Forests, SVMs, logistic regression, etc.), setup the experiment, and analyze the output.

Code for this part: part2.ipynb. Please use data sets from:
http://ai.stanford.edu/~amaas/data/sentiment/

I experimented using bag of word features and random forest model. The outcome is not bad – 87% accuracy.

Part 3: Using data uncovered from your analysis, explain which segments of the data benefit most from the use of either algorithm from Part 1 and Part 2.

In part 1:
Deep learning benefits the segment of data that are very complex to analysis.

In part 2:
Random forest benefits the segment of data that has a lot of correlations.