# EEL6935 Programming HW2

## Natural language processing (NLP)

### Summary

The goal of this assignment is to do sentiment analysis by implementing and training a basic neural network in Python. The result show the accuracy of this basic neural network is not satisfactory (barely above 30%). This shows that sentiment analysis requires more complicated machine learning system.

### Environment Setup

The project is setup in Python 2.x environment, fortunately, the package comes with a requirements.txt file showing all required python modules. After installing python with Anaconda, the setup can be easily done by simply running:

```
> pip install -r requirements.txt
```

In terminal, in my case, is Mac OS terminal running zsh.

### Get Dataset

The code package doesn't comes with dataset, but a shell script to get them instead. Downloading the dataset can be done by running:

```
> cd big_data/datasets
> bash ./get_datasets.sh
```

### Implement and Training Neural Network

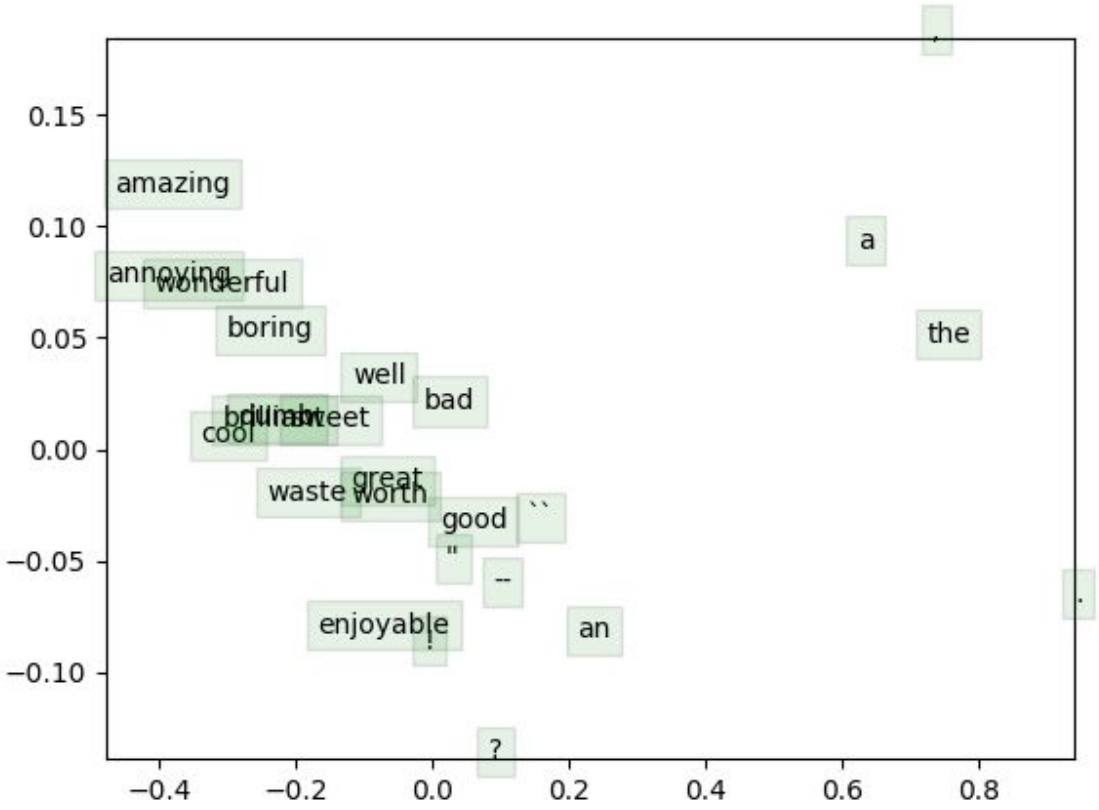
Following the homework instruction, the unfinished python code can be done. The finished codes can be found in the homework code repository hosted on GitHub:

https://github.com/ufjfeng/EEL6935-Assignments/tree/master/NLP

After implementing the neural network, we can train the network with downloaded dataset by running:

```
> python q3_run.py
```

This will take a while since the training has 40,000 iterations and the neural network is implemented in python. After training, a plot indicating the training result of each word vector is shown as:



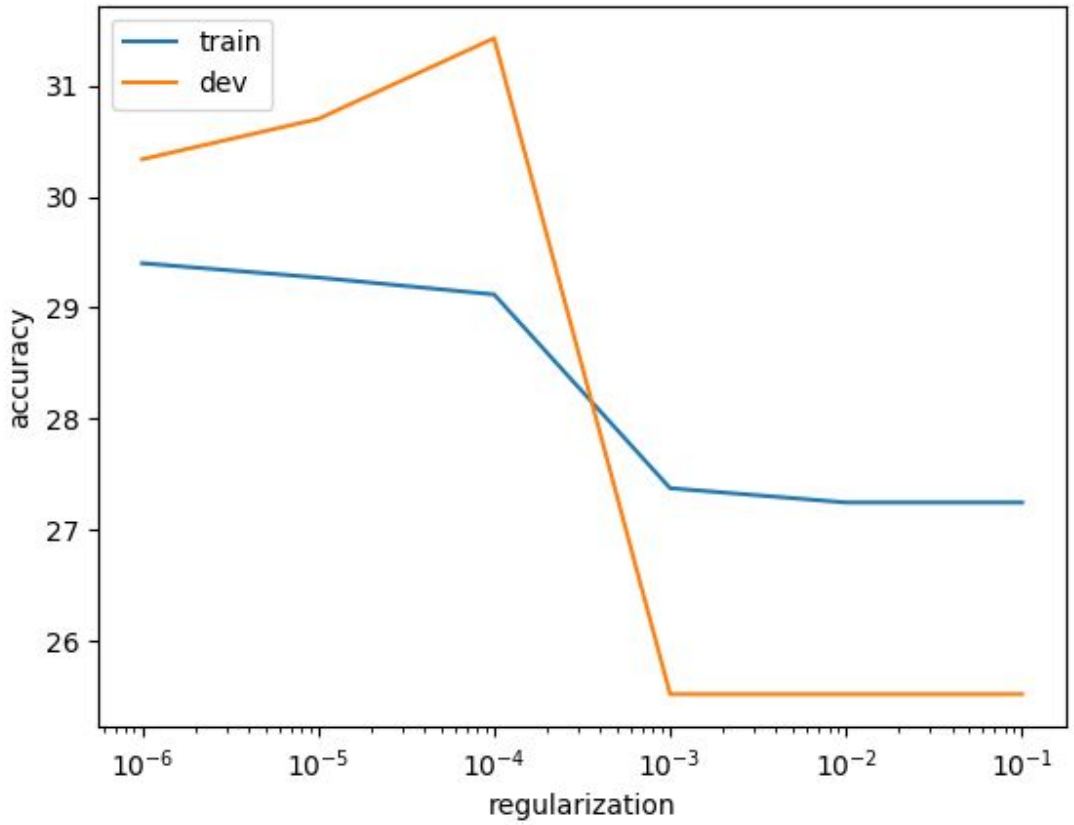

## Regularization

In order to select the best regularization parameter, we ran a experiment in Q4 by testing the network performance with regularization coefficient vary from 1e-1 to 1e-6. This can be achieved by add a line:

REGULARIZATION = [10 ** -i for i in reversed(range(1, 7))]

Into q4_sentiment.py.

After another simulation of, the result of different regularization result is shown.




=== Recap ===

| Reg | Train | Dev |
|---|---|---|
| 1.000000E-06 | 29.400749 | 30.336058 |
| 1.000000E-05 | 29.272004 | 30.699364 |
| 1.000000E-04 | 29.119850 | 31.425976 |
| 1.000000E-03 | 27.375936 | 25.522252 |
| 1.000000E-02 | 27.247191 | 25.522252 |
| 1.000000E-01 | 27.247191 | 25.522252 |

Best regularization value: 1.000000E-04
Test accuracy (%): 27.556561

The plot and recap shows that 1e-4 seems to be a relative good choice for regularization since it produces the best result for sentiment analysis.