First deliverable

# Goal of the project

Sentiment analysis on a Chinese restaurant review data.

# What I have done

1. Get code samples ready
   1. A sample using “snownlp” and naïve bayes to do sentiment analysis on a small restaurant review dataset (which I will refer to as the **toy** dataset from now on) <https://github.com/panluoluo/crawler-analysis>;
   2. The logistic regression sample on wine dataset <https://github.com/python-engineer/pytorchTutorial>. It uses batch gradient descent (BGD).
2. Combine the two samples above and compare the performance of three methods: “snownlp”, naïve bayes, and logistic regression.
3. Run the code on toy dataset. The baseline is like

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| snownlp | .763 |
| naïve bayes | **.899** |
| Log reg (BGD, epoch = 100) | .780 |
| Log reg (BGD, epoch = 1000) | .810 |
| Log reg (BGD, epoch = 10000) | .813 |

1. Prepare the dataset
   1. Download it from <https://www.kesci.com/mw/dataset/5e946de7e7ec38002d02d533>
   2. Drop rows with empty “rating” or “comment” in Excel.

# Problem

The dataset is just too large. The csv file is about 500 M, but feed it to the program (cuda enabled) will result in (after running over 10 minutes) an error:

MemoryError: Unable to allocate 1.35 TiB for an array with shape (1048575, 177020) and data type int64

Currently I am limiting the number of features to control the data size.

max\_features = 10000,

I will appreciate it if someone can help me out here!

# What I am going to do

Improve the code quality by learning from the sample in the textbook.

Use mini-batch and torch.dataloader (will this make the program more efficient?).

Compare different optimizers.

Try different segmentation: 1-gram, 2-gram, PkuSeg, THULAC, etc.

Try out multinomial logistic regression.

There are columns of user ID, restaurant ID, rating of environment, flavor, and service besides the overall rating in the dataset, maybe I can do some more analysis?