**CS498 AMO Homework 7**

Team :

Minyuan Gu ([minyuan3@illinois.edu](mailto:minyuan3@illinois.edu), netid minyuan3)

Yanislav Shterev ([shterev2@illinois.edu](mailto:shterev2@illinois.edu), netid shterev2)

1. (**5 points**) Page 1: First-glance analysis. Plot the distribution of comment length.
2. (**20 points**) Page 2: tokenization preprocessing. Put the snippet of your code here and give an example result of one comment.
3. (**20 points**) Page 3: vectorization preprocessing. Put the snippet of your code here and give an example result of one comment.
4. (**20 points**) Page 4: classifier one. Train classifier on the training dataset and report the accuracy on the test dataset.
5. (**20 points**) Page 5: classifier two. Use **the same** train-test split in the previous step. Train this classifier on the training dataset and report the accuracy on the test dataset.
6. (**10 points**) Page 6: result analysis. Which classifier is better? Why?
7. (**5 points**) Page 7: code.

**Libraries used & Reference:**

**David Forsyth’s book** - Applied Machine Learning

**Trevor Walker’s lecture and sample code** – CS-498 Lecture videos

**csv** – for reading data from csv format: <https://docs.python.org/3/library/csv.html>

**Yelp’s reviews dataset** - <http://courses.engr.illinois.edu/cs498aml/sp2019/homeworks/yelp_2k.csv>

**Numpy** - <http://www.numpy.org/>

**matplotlib** - to plot the sum-squared error, mean images and distances plots: <https://matplotlib.org/>