**HW1**

Top of Form

**Q1**

1 Point

|  |
| --- |
| **Published Date:** Sept. 3, 2020, 4:46 p.m. |
| **Deadline Date:** Sept. 18, 2020, 11:59 p.m. |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
Overview and Assignment Goals:  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
The objective of this assignment are the following:

1. Implement the Nearest Neighbor Classification Algorithm
2. Handle Text Data (Reviews of Movies)
3. Design and Engineer Features from Text Data.
4. Choose the Best Model i.e., Parameters of a Nearest Neighbor Selection, Features and Similarity Functions

**Detailed Description:**

A practical application in e-commerce applications is to infer sentiment (or polarity) from free form review text submitted for a range of products. For the purposes of this assignment you have to **implement a k-Nearest Neighbor Classifier to predict the sentiment for 15000 reviews for movies provided in the test file (test\_file.txt).**

**Positive sentiment is represented by a review rating of 1 and negative sentiment is represented by a review rating of -1.**

In the test file you are only provided the reviews but no ground truth rating which will be used for comparing your predictions.

**Training data consists of 14999 reviews and exists in the file train\_file.txt. Each row begins with the sentiment score followed with the text of the review.**

For evaluation purposes (Leaderboard Ranking) we will use the simple accuracy metric comparing the predictions submitted by you on the test set with the ground truth. Some things to note:

The public leaderboard shows results for 50% of randomly chosen test instances only. This is standard practice in data mining challenges to avoid gaming of the system and prevent overfitting.

The private leaderboard will be released after the deadline evaluates all the entries in the test set. In a 24-hour cycle you are allowed to submit a prediction file 10 times only.

**format.txt shows an example file containing 15000 rows alternating with 1 and -1. Your final submission should be similar to format.dat with same number of rows i.e., 15000 but the sentiment prediction should be generated by your developed model.**

**Rules:**

This is an individual assignment. Discussion of broad level strategies is allowed but any copying of prediction files and source code will result in an honor code violation.

Feel free to use the programming language of your choice for this assignment.

While you can use libraries and templates for dealing with text data you should implement your own nearest neighbor classifier.

**Deliverables:**

Valid submissions to the miner2.vsnet.gmu.edu website

Gradescope submission of report and source code:

Create a folder called HW1\_LastName, and put all the source code there. You should include a readme file in your code folder. The readme file should include information on how to run your code and where your k-NN implementation is.

Submit (on Gradescope) a 3-page, single-spaced report in PDF format describing details regarding the steps you followed for developing the classifier for predicting the product review sentiments. Be sure to include the following in the report:

1. **Please state your miner2 user name, mason userid (eg: gmason76), and your best public score for your submission (at the time of writing the report) at the beginning of your report.**
2. **A detailed, but readable description of your approach, and how and why you chose the parameters, features, distance measure, etc.**
3. **Your methodology of choosing the approach and associated parameters. Any graphs or tables illustrating key experiments you did in the process of choosing your final model.**
4. **Any graphs or tables illustrating key experiments you did in the process of choosing your final model.**
5. **Also submit an archive of your HW1\_LastName folder (.zip or .tar.gz) via Gradescope.**

Make sure you answer all the questions. It is always a good idea to organize your report. Feel free to use subtitles in each section such as Introduction, Approaches, Experimental Results and Conclusion. You can also summarize results in a table and use bold font to highlight key results. If you are using something outside of the class notes, that's fine. Explain briefly to show your understanding to avoid losing points.

**Grading:** Grading for the Assignment will be based on your implementation (40%), report (30%) and ranking results (30%).

Bottom of Form