Program 4

Submit Assignment

Due May 10 by 11:59pm **Points** 100 **Submitting** a file upload **Available** until May 10 at 11:59pm

For this assignment, you'll be using some real-world data to predict product quality. The data will need some initial preparation before processing.

Your data file is based on reviews from The Ramen Rater, a site for serious ramen fans. The data consists of 2580 product reviews. For each, you have:

- Identifier (unique integer)
- Brand (355 values)
- Variety (2413 unique values)
- Style (8 values)
- Country of origin (38 values)
- Stars (numeric rating)
- Top Ten (text, missing on most entries, ignore)

You will need to do some initial data configuration and recoding. This can (and should) be done separately, before using TensorFlow to build your network. Specifically:

- Some companies have multiple products, others only appear once. Count the number of times each company name appears. Any company appearing only once should be replaced with "Other". (With only 1 product, there's no way to generalize about the company.)
- The "variety" field has the richest information, but also needs some distillation before we can use it conveniently as input into a neural network. We'd like to identify specific features--flavors, descriptions, etc--that might tell us something. So we have to get it into that form. Again, this should be done separately: Take the text from the Variety column, and break it into words. Count the occurrences of each word and find the 100 most common words. Keep ONLY those words for classification. (Note that some reviews might have several of those popular words, others might not have any.) Code for the presence or absence of each of the most common 100 words. (Note that this is not one-hot coding; if 5 of the words are present, then all 5 should be accounted for in input, not just one.)

Once the data is ready for input, build a neural network in TF to predict the overall rating. Your rating should be by categories, 0-5 stars. Use 1-hot coding for the output (maximum value is taken as the classification).

Write up a short report discussing:

- The configuration of your network, along with a brief overview of your TF code. It is not necessary to discuss the (python, perl, whatever) code you used to clean up the data.
- Your cross-validation strategy
- Summarize your results
- Any further questions you'd like to discuss, ideas for extending this further, etc.

Submit your report and your source code by the deadline.

This data came from: **https://www.kaggle.com/residentmario/ramen-ratings#ramen-ratings.csv (https://www.kaggle.com/residentmario/ramen-ratings#ramen-ratings.csv)**

This dataset is republished as-is from the original **BIG LIST** (https://www.theramenrater.com/resources-2/the-list/) on **https://www.theramenrater.com/** (https://www.theramenrater.com/).

ramen-ratings.zip