CIS 700-004 Spring 2019 **Homework 3, part 1: due April 19th, 2019**

For this assignment, we are going to use the Reddit sarcasm dataset. Since sarcasm is difficult to express via text, Redditors frequently end sarcastic comments with "/s". The Reddit dataset uses the presence of this "/s" token to create a labeled dataset of sarcastic and non-sarcastic comments. There will be some slight modifications to the dataset – we've removed metadata and balanced the dataset (only 1% of Reddit comments are actually sarcastic, but 50% of the training and test examples are sarcastic in the modified dataset).[1]

Note that even within the niche of sarcasm-based NLP, there are better, cleaner, and larger datasets than this Reddit dataset. I've selected this one specifically because there are many teachable characteristics of the dataset. Take a look at the dataset and answer the following questions.

1. What are the hardest aspects of this dataset? How noisy do you think the dataset will be? (Hint: consider confusion matrices)

2. Come up with at least 2 data hypotheses: features that distinguish sarcastic comments from non-sarcastic comments that any good model should pick up on.

3. Reddit is based on threads (comments replying to other comments). In the sarcasm dataset, we have potentially sarcastic comments as well as their parents. Consider the following methods of passing the structured data to models.

   (a) Only pass in the candidate comment. Ignore the parent comment.

   (b) Pass in the concatenation of the parent comment and the candidate comment.

   (c) Pass in the parent comment and the candidate comment into separate models.

   Describe the inductive biases imposed by each of the modeling setups. Create a hypothesis for the ranking of the 3 setups.

4. Consider the following 3 models.

   (a) A logistic regression on the average word embedding over the input text.

   (b) A CNN on variable-length word embeddings using 1D convolutions onto a linear layer.

   (c) A bidirectional RNN on variable-length word embeddings.

   For each of the setups, hypothesize which of the models will work the best. (Hint: perform some exploratory data analysis to check whether the RNNs are suitable).

---

[1] Reddit is a free-speech community, so the dataset can be somewhat offensive and vitriolic. These comments do not represent the opinions of the dataset creators (Khodak et. al.) or the CIS700-004 teaching staff.