* Train a classifier on the similarity score, which can help us to decide what similarity score should be correct.
* Mark high confidence data labels, leave unsure answers to human graders
* Text similarity, question demoting, term weighting and length ratio features
* Use an ensemble of models to generate a score
* Adversarial attacks (cheating/fooling model): <https://arxiv.org/pdf/2201.08318.pdf> maybe get entities then substitute random words in between