数据集

Jigsaw Unintended Bias in Toxicity Classification（Detect toxicity across a diverse range of conversations）

2019年一个比赛的数据集

Background: When the Conversation AI team first built toxicity models, they found that the models incorrectly learned to associate the names of frequently attacked identities with toxicity. Models predicted a high likelihood of toxicity for comments containing those identities (e.g. "gay"), even when those comments were not actually toxic (such as "I am a gay woman"). This happens because training data was pulled from available sources where unfortunately, certain identities are overwhelmingly referred to in offensive ways. Training a model from data with these imbalances risks simply mirroring those biases back to users.

模型目标：建立在各种对话中公平运作的toxicity模型

数据情况：

total size： 2.38 GB

数据类型：约 200 万条公众评论数据

The text of the individual comment is found in the comment_text column. Each comment in Train has a toxicity label (target), and models should predict the target toxicity for the Test data. This attribute (and all others) are fractional values which represent the fraction of human raters who believed the attribute applied to the given comment. For evaluation, test set examples with target >= 0.5 will be considered to be in the positive class (toxic).

进展：目前在参考一个案例，写Benchmark Kernel