# 前期准备

## 建议流程

1. 仔细阅读比赛的概述和数据的描述；

2. 查找类似的Kaggle比赛。作为刚加入Kaggle社区不久的新手，我对所有的Kaggle比赛进行了搜索和分析；

3. 研究类似比赛的解决方案；

4. 阅读相关论文，确保我跟进该领域的最新进展；

5. 分析数据，并构建可靠的交叉验证；

6. 进行数据预处理、特征工程、模型训练；

7. 进行结果分析，包括预测分布、错误分析、困难样本等；

8. 根据分析，精心设计模型或设计新模型；

9. 基于数据分析和结果分析，通过设计模型增加多样性，或解决困难样本；

10. 模型集成；

11. 必要时返回前面某个步骤

## 数据情况

Tied comments 不会有相同结果的句子 df[“score”]=df[“score”].rank(method=”first”)

有5302对有2个评估，5302个有1个，当存在相斥情况时引入第三个worker评估

由于计算方式是逐对与专家评估结果进行对比，最大分数不为1.

提供了两两句子的比较结果作为验证集，提交要求在一系列句子上生成一个句子毒性的排名。

Tied comments：df[“score”]=df[“score”].rank(method=”first”)

### 参考注释方法

BWS注释办法：四个组一起比较，进行a、b、c、d的打分。（v，a，d与攻击性也有关联）

## 模型处理参考

### 一

训练代码：怎么训练，简单想法是 拿二分类训练 再换成打分输出

实际 训练一个能判断毒性程度的model 用两个input 进行比较 结合MarginRankingloss 调整模型

MarginRankingLoss: 

当y=1 ,认为x1>x2。若实际也是如此loss=0，否则不为0.y取1 -1（？） 损失函数与目标是否一致（？） 如果模型是训练一个毒性打分模型，两个这样的模型计算损失，是符合的。

## 过去比赛经验

稳健的cv：lb上升，cv对数损失下降才认为合适。

数据增强：反复翻译

第二次比赛：减少身份偏差带来的影响，平衡auc与减少偏差作为metric