Kaggle比赛链接：https://www.kaggle.com/c/quora-question-pairs/data

比赛介绍：这是一个文本二分类的比赛，检测问题对是否是同一个意思。kaggle网页上有详细的描述。

要求：
1.team name：学号+姓名；
2.一人一组，代码不要抄袭，不要分享；
3.Deadline：6月23日 11：59pm；
4.实验分数包括排名+实验报告的质量（详细程度以及逻辑是否清晰合理）

温馨提示：这次的数据集数量较上次是大许多，需要些许的运算时间，请大家最好不要赶ddl。

实验报告包括内容：

1.代码+实验报告pdf 打包成zip（注意是zip 不是rar）。

2.pdf内写明学号+姓名；打包文件夹 和 实验报告 命名都是学号+姓名。如果提交第二个版本，在后面加上v2

3.内容至少包括：

3.0使用的系统+编码语言+环境说明（使用的包的版本）。

3.1 简单的流程图：数据处理方法-选择算法-调参方法-如何防止过拟合等。

3.2 详细介绍一下流程图各块的内容，要在对应的地方贴一下核心代码。

3.3用自己的语言介绍一下你选择的模型的原理，不要百度的介绍，不要水字数。

3.4需要截图你在kaggle上的排名位置。

3.5其他部分 随意发挥 期待看到你们优秀的方法

4.提交方法：ftp:// 172.18.50.234 用户名：student 密码：2019

比赛内容：一个文本二分类的比赛，检测问题对是否是同一个意思

1.做一些文本分析，问题包含的词个数，字母大小写，可以过滤一些标点符号，tf-idf观察低频词对分类影响
2.通过统计得到文本特征（比如问题1问题2有多少个一样/不一样的词，问题1问题2tf-idf低频词个数等等），再用分类模型分类

3.word embeding：**word2vec**，（Glove，fastText，ELMO，GPT，Bert）
4.CNN，RNN, GRU, LSTM，（+attention）

深度学习的框架：keras，pytorch，tensorflow