黑桃：

谈一下自己的感受。

在这之前看过一些机器学习的算法，(大多忘记了)但没有实际应用到项目中，曾经想过自己编程实现算法 ，无奈编程能力不过关。后来无意中看到有这样一个带打比赛的课程，然后就报名了，总的来说收获是非常大的。

整个比赛我们尝试了所有的机器学习模型，最终分数 0.78。

最初我是用的老师给的 svm 代码，跑出来的成绩是 0.74，后经过一些调参，维持在了 0.76，仔细研究老师的代码后，发现老师使用的只是 word 数据，于是想着将 article 和 word 拼接起来，也就是构建新特征，结果是令人惊喜的，一下到了 0.779。尝试了各种模型调参没有显著提升，又放下了一段时间，后来我们尝试了将文章的长度作为特征拼接，结果分数反而降低，故而舍弃。

我们用跑出来的特征，将特征进行两个拼接，三个拼接，然后放模型里跑，见始终不能突破 0.78，于是我们开始尝试模型融合，主要是两方面，一个是概率文件融合，另一个是 result 文件融合，模型融合的核心是模型要好而不同，我们将不同模型跑出来的概率文件和 result 分别融合，最终结果终于突破到 0.782

后面打算再深入学习一些算法，然后提高自己的代码能力。

最后，整理总结很重要！整理总结很重要！整理总结很重要！

Youn：

在职菜鸟。18 年 5 月份接触的 NLP,工作之余断断续续的会看一些 NLP 相关语言模型与算法。这是我第一次接触 NLP 比赛,是从训练营才知道的,就想跟着老师试试水。

一开始使用 tf-idf+LogisticRegression 默认参数分类,随机 9:1 划分训练集验证集,0.76+的查准率,线上 0.76+。后来又试了 LinearSVC，MultinomialNB,RandomForestClassifier 分类算法,发现 LinearSVC 效果略优于 LogisticRegression,

然后就想着去优化提取的特征值,试了 hashing/doc2vec+LinearSVC 默认参数,但是效果都没有 tf-idf 好,后来就 tf-idf+LinearSVC 佛系调参,终于上了 0.77。9.16 号去参加了达观杯线下的前十选手答辩会,收获很多,也开拓了对 NLP 领域的视野。(没找范晶晶大哥照面面基我还是挺后悔的,哈哈)

后面我会跟着 Top10 万里阳光号的算法思路去重新做一下达观杯的分本分类,去尝试分类模型的融合以及深度学习的分类去达到更好的效果,如果有什么好的收获也希望和大家一起分享反馈。(任重道远,希望可以跟着大家好好学习,共同进步)

总的来说,这次达观杯比赛对我来说意义很大,也坚定了以后做 NLP 的决心。

Dejavu 的梦：

某高校本科生。参加比赛前机器学习小白。在官网提供的 baseline 与 jian 的指导下。使用 LR 与 SVM 单模型到了 0.779。提升的方法主要是 jian 的特征工程以及 grid 调参。后结合了深度学习的结果进行模型融合后上到了 0.78 多。

在后半段做的过程中发现先前几个机器学习的方法得到的结果太过于相似【即模型相似度太高】，所以融合的效果并不好。深度学习虽然单模型效果不好【0.75 上下徘徊】，但差异性大。融合后效果后，这便是我上 0.78 的关键。

Louis：

这是我第一次做这种机器学习的实践，由于本科是学数学的，之后学习软件也是一直在做工程上的东西。比赛最开始拿到老师在知识圈的代码按照代码敲完后提交了成绩，然后在群里看大家讨论说特征提取采用 tfidf 会更好，就去学习了这方面的知识然后尝试了修改代码结果确实变好了。之前看过同学的毕设知道 svm 可以做分类，然后又修改了代码，但是电脑跑了 30 个小时左右后崩溃了。没办法只好找了一台内存更大的电脑继续跑最后发现效果还不如 LR，在此之后又尝试了决策树，随机森林，模型融合等模型。最近在做数据降维＋随机森林，但是再一次电脑崩溃了。总的来说，由于自己是个考研党，一天学习的时间不是很多，希望之后虽然是时间少但是不断去探索吧，提升自己的实践经验和算法推导能力。最后想问问大家参数怎样去选择会更好，每次感觉自己在随缘调参心好累。

月光疾飞：

这也是我打过的第一场稍微正式的机器学习的比赛，之前做的都是做的一些例子。

我主要从以几个方面进行总结

1. 开始阶段，由于本身有机器学习和 Python 的部分基础，所以拿到老师的代码后，能看懂并且能去尝试着修改，这样在开始的几天，试过几个独立的模型，比如 LR 的 0.7326，SVM 的 0.7597,LinearSV 的 0.76871，然后做的是初始几个模型的对比，得出在不进行降维等操作，LinearSVC 的效果最好。

2. 第二阶段就到了 9 月 1 了（中间时间太忙，没弄这个比赛）。然后就开始进行降维，模型的融合等操作，队友训练的模型和我的模型进行融合，中间有些结果确实不好，说不清楚这个原因。最好的结果达到 0.779 吧！

3. 第三阶段没有完成，k 折交叉验证做到一半，比赛就结束了，很遗憾的是在比赛中没有跨过 0.78 的坎。然后就是 b 榜成绩也没提交，直接从 100 多降到了 300 多

总之，比赛经验比较缺乏，特征本身也没做得太好，然后就是模型预测效果上不去（可能过拟合、也可能本身效果不够理想）。自对算法的认识不够深入。这个比赛也基本也只是跑了一个大概的流程，加上本身的态度，打这个比赛就是觉得好玩，没太花时间。但每做一部分自己确实在用心，晚上跑数据，早上看结果去调整，有时候也等待着去出结果，看是否会有更好的效果。

希望之后能提升比赛经验和对算法的进一步深入的认识，然后就是一定要用心去打比赛，多花点心思和时间，做好每个阶段。

空晴：

关于这个比赛，首先是在论坛中找到了有人发了个 baseline 代码，运行后得到结果还不错。之后就是参加了咱们这个训练营，学习到一些算法知识，并且尝试改进 baseline 的代码，大概试了 30 几种尝试吧。发现比赛中对机器的内存要求很高，有一些程序竟然超过 32G，造成机器卡死。为避免对内存的消耗，将训练集进行小量的划分，得到训练集和验证集。另外，对参数的设置，是根据网格搜索与交叉验证进行调试，得到符合自己满意的参数。在比赛过程中，发现在自己小量验证集中得到较好的结果，但是在提交的结果中，反倒排名下降了，原因不明。最后得到较好结果是将词集与字集特征简单的融合一起形成一个新属性特征，对这个特征进行 TF-IDF+SVM 算法得到还算不错的排名。总结来说，参赛经验欠缺，对业务理解不熟练，无法找到构建有效的特征属性，缺少对脱敏数据的理解。

金文：

我是第一个接触关于 nlp 算法的比赛，之前不是特别了解这方面的，也是在参加咱们的西瓜书训练营才报名了，然后就是自己一个人跑数据，最开始的时候连 tfidf 都不知道，然后上网查来补充自己不知道的地方，但是毕竟重在参与吗？刚刚下载数据的时候一看 3g 多就觉得这数据真他么多，哈哈哈（工业界的肯定比这多），感觉然后看了看数据标签的比例，看看样本是不是分类均衡，但是几乎样本对应的类别的样本量差距不大，要想办法过滤一下停止词和标点符号。针对脱敏数据，统计了每个词的 idf 值，将那些所有文章都出现的词筛选出来，些词即使不是标点符号，肯定也是一些对文章主题无用的词，可以过滤点。另外，经过分析，发现训练集样本中有文本重复的记录，应该是数据的噪声，有些重复样本的标签居然是不同的。这种噪声数据，感觉应该过滤掉吧，而有很多标签相同的重复样本我就保留一条，其余的也全部给删除了。当时选模型首先就像到了 lr 模型，因为它性能很好，短小而精悍，于是用了 TfidfVectorizer+lr 的训练分数达到了 0.76，感觉还可以，后来就尝试了各种算法 svm、lgb、xgboost 等发现 lgb 的分数就高达到了 0.77239，其他的都是介于之间的，遗憾的是模型融合啥的也没有做，最重要感觉还是自己的特征处理没有做好，也可能是自己第一次参加这样的比赛，没什么经验，还有一个就是自己对脱敏的数据基本无从下手，虽然现在很多比赛的数据都是脱敏的，毕竟数据这个东西谁都不想透露给别人。当然了这之间也跑过老师的代码实例但是电脑配置稍低带不起来

程剑杰：

简单总结

我在打比赛之前也没有接触过 NLP 的一些东西，研究生阶段一直是做的 cv 方面的东西，大多用深度学习的方法去做，基本用不到机器学习算法，为了秋招，想把机器学习算法巩固一下，就参加了这个比赛。

选择模型方面，关于 NLP 在比赛之前还只是知道 RNN 和 LSTM 这两个深度网络，然后老师说用机器学习算法效果不错，也就顺理成章的用了机器学习算法而没有去考虑深度模型。机器学习模型了解的都试了一下，LR，SVM,XGBoost,RF,贝叶斯大概这几种方法，LR 到了 0.75 左右，SVM 到了 0.77 左右，XGboost 但模型也是 0.778 左右，可能没怎么调参吧，因为 tfidf 的特征维度本身就很高，SVM 只需要用 Linear 就够了，速度快效果也不差，贝叶斯的效果不是很好只有 0.56 左右，RF 也是 0.73 左右，效果不是太好，突破 0.78 的瓶颈还是用的模型融合，用的各个单模型的结果多数服从少数的融合方式，还没试过 k 折 stacking 的方式去做一下，时间很紧，当时又是一个人搞的比赛，又遇上了一堆秋招的笔试面试，就没有坚持跟进比赛的进度了，现在秋招有了着落，就想着再重新做一下比赛的内容，多考虑一些内容，也算是个总结。

特罗伊：

我也来说说，请老师同学批评指教。这次共提交了 44 次结果，最终 B 榜公布是 0.789680. 针对这个任务，我尝试了 SVM, LR, XGB, LGBM,NBSVM 这些 ML 的算法，也尝试了 CNN, RCNN, LSTM+ATT, InceptionCNN, BiLSTM 这些 nn 的模型。svm, lr 的特征尝试了字和词拼接后卡方选择（2.8m 维），svm 效果较好，A 榜 0.782；由于 xgb 和 lgb 太吃机器，仅使用 200k 的 tfidf 和 jia 老师提供的 lda,lsa 和 doc2vec 特征拼接，lgb 较高，A 榜 0.77+；nn 中只用了 word_seg，最好的是 RCNN,A 榜 0.77+。

策略与集成：pseudo label（在 svm 上几乎没有提升）；stacking（在基分类器仅用了 svc，

lr，lgb，第二层用的 lr，B 榜 0.7863）；voting(西瓜书 P182，多数投票法，A 榜 0.7897)，最终 B 榜公布是 voting 最高；数据增强，主要是用在 nn 模型中，方法包括词的随机乱序与随机丢弃、文本的截断（前中后各 600 个 tokens ）。总结：由于语料脱敏了，因此词嵌入能够带来的提升有限，猜测这也是这个任务 nn 表现没有那么突出的原因；尽管数据决定了机器学习的上限,算法模型只能逼近这一上限,但是参数决定了模型和算法的上限,要调好参数，就要吃透算法原理，这也是琛享这个班最大的特色吧。希望有机会跟大家组队打其他相关比赛，共同进步！

郑基亮：

本人第一次参加这类比赛，总共提交 3 次

第一次试验一下 beginner 代码

第二次结合了老师给的代码，先生成 tfidi 特征，再基于此生成 lsa，lda 特征，再将原始数据集向量化，然后再基于 lsa,lda,vec 进行特征融合，转换成稀疏矩阵，最后用 linearSVC 训练数据集。

其中 LDA 降维太费时间，两次没有跑完（电脑必须关机了），由于接近提交成绩的时间，将维度降低一半，顺利跑完，最后的成绩是 0.75。

第三次，比赛完成后，又按照原来的维度跑了一遍，提交后显示成绩提升了一点点。

达观杯比赛结束了，这个流程跑下来，仅是 beginner 代码自己尝试了其他几种训练方法，后面提交的成绩，没有再试其他方法。

作为一个年龄较大的上班党，能用于学习的时间不太多。目前每天能保证晚上有半小时到 1 小时的时间来看机器学习相关的内容，白天空闲时间不定，周末的时间稍微能多一些。一般是利用通勤时间或零星时间准备《西瓜书带读》的当天作业，晚上看西瓜书和机器学习实战。

加入这个群，是想进一步弄明白机器学习常用几个算法，推导与应用等。范同学建群的初衷不错，每人准备一方面的内容，写笔记后再大家分享，我个人愿意在这件事情上分配时间和精力。可能需要范同学进一步考虑算法研讨工作的分配以及分享的形式等。

明年的达观杯大赛，希望自己能轻松应对。

杨云亭：

第一次接触达观 nlp，是在报名学习西瓜书这期训练营开始。起初稀里糊涂，然后按部就班，既然学习就好好看一下嘛，nlp 这个方向之前一点都没接触过，身边更多的小伙伴是做图像这块，我的方向更偏理论，西瓜书在大四末拥有，研一也一直涉及，却苦于没有实践的能力。训练营中有一点感动我的是可以带打比赛，出于此，才有缘这个比赛。

比赛经过：

刚开始，下载数据集，了解比赛的任务，通过训练数据集，来预测长文本数据的分类。当初想着就是跟着老师的任务，一步一步来，后来发现要想做得更好，还是需要自己对模型和特征进行一些处理，不能只靠单一的模型和简单的调参。

第一次使用线性模型中的逻辑回归做的，特征使用的 countVectorizer,结果是：0.732...

第二次是群里在讨论说 tfidfVectorizer 的效果更好，实验结果相对提升了一些，达到了 0.76...

之后又尝试过贝叶斯模型，结果只有 0.66.....linearSVC 结果是 0.71....,决策树结果只有 0.52.....，以及将几个模型集成在一起的效果也只有 0.72，集成的模型包括三个：LR,贝叶斯

和决策树，使用的都是 tfidf 特征，最后通过投票的方法决定了最后的分类。期间也曾尝试加入老师在 github 上给出的机器学习中特征处理的代码，机器太差，也没跑出什么，只是试着去明白理解特征处理的过程。

感悟就是对于知识的应用很肤浅，赛制期间所做的准备有点少，只是在自己所知道的几个模型间进行了实现对比，效果也一般。后续还是需要多查资料，多尝试一些更深的方法，而不能只停留在单一模型。

52：

达观杯的比赛梳理

在我看来，这个比赛主要是结合 NLP、机器学习和深度学习等技术，构建文本分类模型，并以此实现长文本的精确分类。

刚接触到这个比赛时，其实我内心深处也是拒绝的。因为 NLP 是我之前完全未接触过的领域，而且脑海里的固有印象中只有深度学习（如 RNN）模型来处理效果会比较好，但我恰好主要的研究方向是机器学习部分。但是队友告诉我，用机器学习也是可以做的，不需要太多的 NLP 相关知识，而我也决定硬着头皮试一下，所以有了今天的比赛梳理。

打开比赛方所给的文本数据，我发现这些是一堆的字，词向量。根据赛事方的描述，二者是独立编码的，因此属于两个类别的特征并且这些特征维度非常大。虽说我当时不懂 NLP 中特征处理方式，但通过查询并与队友交流，我们知道了 tf-idf，countvector 等几种关于文本特征提取的方法。当时我的想法就是先用最基本的特征提取方式，确保可以跑出一个模型，等有一个结果出来时候再继续改进。就这样第一个模型结果诞生，是 tf-idf 提取 word 特征并用 LR 模型训练出来的，大概有 0.7 左右的分数。

其中，但是有些方法提取的特征真的太大，我的小笔记本根本带不动啊啊啊啊啊！（还花了 1600 买了 16g 的内存条。。。）。因为我之前在学校时研究的就是智能算法优化的方向，所以自然的想到用遗传算法等启发式算法进行超参的优化。但是遗传算法的思想其实在我看来是合理的暴力搜索找局部最优的方法，它的缺点就是，太慢了!! 尤其是数据量大的情况下，本身训练一个模型就很费时间。所以我采用了训练数据取原始数据五分之一的方式，使用 LR 进行训练，设置了种群数量为 10，迭代次数为 20。这代码跑了两天，结果不如原始默认参数的 LR。。可能原因是数据量太少，迭代次数太少，种群数量也少。并且，智能算法的超参范围是需要人为给定的，但是我当时对超参的含义还不是很理解，可能导致初始设定的搜索范围就有问题。

之后我尝试了线性 SVM，默认参数，跑了 20 多小时。跑出来结果类别全是 3，这个问题至今未解。。。。。凉凉

最终，我在学习了一些超参的设置后用模型 LogisticRegression(C=10, dual=True) 跑出了分数 0.77+的结果。

Bigjing：

我先分享一下我的比赛经历和感受，达观杯是我参加的第一个比赛，我是小白，基本上边比赛，边实践理论，边完善，第一次跨过 0.778 的坎，关键在模型调参，第二次跨过 0.779 的坎，关键在于特征工程，第三次跨过 0.78 的坎，关键在模型融合……因为小亨邀请我 29 号晚上做个分享，具体再细说。如果你也是小白的话，我有两个建议:【1】不一定学好了所有理论再去打比赛，打比赛和理论研究的区别在于，比赛像做工程，先快速实现再深究，先能做再考虑做好;【2】向你的老师，队友等其他优秀的人学习，这是你进步最快的方法。小

白视角，希望对你们有帮助。再次感谢 Jian 老师带我入门。

完美男 Ren：

复盘一波，刚学习 nlp，只会跑老师给的代码，代码有什么用一直是个迷，由于这个比赛数据全是数字，看不出来提取的特征到底有没有用，得到概率大的不一定是有用特征。这个 tfidf，感觉也就只是个提特征的辅助，具体提特征里面的方法还需要继续探究。在实际工程中，切词，预处理等工作做好真的不容易，也要继续研究。综上，预处理怎么搞，特征怎么提，是我目前关注的重点。对于分数，目前不是太在乎，以上工作没搞明白，对分数也没兴趣。群里大佬不少，多多学习，大家互相帮助。正在找工作中，希望找工作的同学也多分享面试经验。

Victor Sun：

作为刚用 python 做机器学习的小白，简单说下参加训练营和达观杯的感受吧。大三时开过数据挖掘课，参加过数模、泰迪杯数据挖掘、天池广东机场流量预测几个比赛，当时都是用 SPSS、MATLAB 做的曲线拟合。大四跟导师学了一年硬件，现在研一又回归到纯软上来。真正起步就是在毕业的暑假，刷视频学习 sklearn 库，用 python 做数据分析，机缘巧合发现训练营，参加达观杯比赛。通过逻辑回归 0.73 的代码入门比赛，去查资料自学经典算法和处理文本数据的方法、交叉验证、网格搜索等，截止时 0.77+。今后多向前辈和老师学习，多通过比赛积累项目经验，进一步学习 ML，为以后学习分布式机器学习或是深度学习打好基础吧。

Beyond：

过去从来没有打过比赛，也对于 NLP 没有什么了解，这是我第一次参加比赛，我最终的成绩是 0.77。

最开始使用 countvector + LR，就是老师给的原始代码跑出来是 0.73。然后，我就在想怎么提高成绩，无意中看见，大家说 tf-idf 效果更好，就改成了 tf-idf＋LR，果然，成绩提高了 0.76。然后我又尝试了用 tf-idf＋LinerSVM 跑，就是用的默认参数，0.77。用 tf-idf＋XGBoost 跑，这个由于机器原因，没有跑了一天多都没跑出来。这时候，我就想，我调一下参数以提高成绩，我调的是 tf-idf＋LinerSVM，调了之后发现，没有明显的成绩提升，就只提升了小数点 2 位以后。在老师对于比赛进行了讲解之后，我就开始尝试先进行特征选择，然后提取 lsa 特征，doc2vec 特征，lda 特征，但是很尴尬的是，由于内存不足，特征融合失败了，所以，我就单单以 lsa + LinerSVM 试了一下，只有 0.5。最后，我融合人 tf-idf + LR、tf-idf + LinerSVM、lsa＋LinerSVM，投票法，结果和我预期的差不多，没有太大的提升，还是 0.77。

整体来说，我觉得更多的这次参加比赛的是给了我一个整体上流程的一个认识。最大的收获是老师讲解的特征的选择，和模型的融合。

Yan：

我也来一波分享！我是从运维转机器学习的，开始只是听说机器学习对智能化运维有帮助，本着对新知识的渴望，渐渐的热爱上了人工智能，尤其是 NLP 方向。

也是从打这次比赛开始，我才知道了什么是特征工程以及各种机器学习的算法。由于几

乎没有任何算法的经验，我先跑了一遍 jian 老师给的代码，居然跑出了 0.73 的成绩，心里还是很喜悦的。经过简单的调参，分数提高到了 0.75，这时候我依然是不知其然的阶段。

接下来，开始猛看了西瓜书，结合《Python 大战机器学习》，终于了解了 tfidf 等特征提取方法，熟悉了逻辑回归、决策树、SVM、贝叶斯分类器、Xgboost 等算法，通过特征工程以及 SVM 的运用，分数提高到了 0.776。只因光荣的晋升为奶爸了，后面没来得及继续训练模型，但这足以激发起了我的学习热情！

接下来我会分篇幅总结我对各种算法的认识与疑问，还有对于 NLP 方面的知识总结。期待各位大牛们的指导与分享！

dongXY：

我是一名在校生，做的医学数据分析，以前没有接触过文本分类，是第一次参加比赛，在这个比赛中只是会用一些分类算法，对于文本类的数据进行特征选择并不熟悉。所以刚开始是用的老师给的代码，只改了 logistic 的一些参数，提交之后得到的结果并不高，只有 0.73277。后来用了随机森林、AdaBoost 的方法，但是结果也并不高，可能是特征选择和调参没有做好。由于前段时间较忙，所以在竞赛截止前也没来得及做进一步的调整。接下来要对文本中的特征选择方法多看下。

机智翔同学：

我写的总结，一个博客：

https://blog.csdn.net/GreatXiang888/article/details/82873435

李知一：

nlp 文本赛，下载数据集，看着数据完全不懂，不过后面老师说加密了，就没纠结了，后面照着手敲代码，本来以为要各种特征处理融合，没想到直接用 CountVectorizer 就搞定了，搜索了它的用法 发现又有另一种 TfidVectorizer 训练文本的方式，说实话特征处理真心有点懵，后面对于模型，用老师给的模型做了调参，特别是 C 正则化，调上调下，发现增加惩罚力度，可提高成绩不过不是很高可以到 0.745，之后特征同样的处理，用朴素贝叶斯发现成绩下降了很多有点懵，发现朴素贝叶斯的模型其实和数学贝叶斯理论还是有差别的，没有数学的严谨性，用可行的计算 st 尽可能的接近贝叶斯理论，大家一起学习吧！

李俊：

第三次参加这类比赛，相隔两个月，前两个比赛分别是华录杯的交通路口识别和公交线路预测，路口识别属于图像识别，排名 70+；公交线路预测属于传统的机器学习，排名 10。因此，对于比赛流程比较熟悉，参加 nlp 比赛的目的在于：

1. 大致了解文本分类问题；
2. 了解 CNN 在不同使用场景的表现；
3. 进一步熟悉数据分析过程；

参赛之前对文本处理的了解仅限于：

1. 看过吴军博士的《数学之美》，了解余弦距离在文本分类中的作用。
2. 读过 Denny Britz 的 《Understanding Convolutional Neural Networks for NLP》。

在参加训练营之前，检查了数据是否平衡、是否存在数据缺失。对 word_seg 中的词进行了分析，得到了没累次中出现频率最高的词最为停词，并将只出现过一次的词去掉，进行了预处理。

参加训练营之后，用 logistic 分类进行了训练，并提交了答案，根据源代码得到的结果，提交成绩为 0.73。然后使用自己的方法处理过的 word_seg 提交成绩，比不处理得到的结果也在 0.73，也就是机会没起作用。

然后没有再分配时间处理比赛的内容，而是着重理解 朴素贝叶斯和 支持向量机的算法原理，预计还需要一天的时间，之后会一一测试这两个算法使用的效果。

当然还要结合自己预处理方法理解老师的预处理方法的优势在于哪里。

等到的过去

首先，我接触机器学习时间不太长，是第一次做这样的比赛，明白了参加这样比赛的一个流程。参加比赛前自己学的东西都是按照书本上的，没有一个系统的这样比赛的一个经历，自己在比赛中下载的数据是没想到的，而且由于考试一开始没有提交，后来就按照老师的提交了一次，并且在跑第一次数据的时候跑了几个小时没跑成功，后来用同学的电脑跑了四十分钟。

第二，理解了关于这次比赛的一个基本思路：对数据预处理，训练数据集，预测结果。在做这些工作后，发现自己需要不断的尝试模型和调参，只是尝试了逻辑回归，还有贝叶斯，结果贝叶斯的效果还没有逻辑回归的高。

期间遇到的问题：

我在 pycharm 跑的时候遇到第一次报了一大堆错误，主要自己没有把 pycharm 的配置环境搞好。因为电脑上按了两个关于 python 的，一个是 python IDE

一个是 anaconda，没有设置好。

自己对特征这一块不太熟悉，自己换的时候回经常出现莫名其妙的 bug。，不过 tfidfVectorizer 比 countVetorizer 好一些。

自己也在尝试别的模型，手写代码还没有完成。我先彻底搞懂自己弄得这几个，在尝试一些决策树之类的吧。

书里行中：

第一次参加这类比赛，了解了参加比赛的基本步骤，这次的 nlp 的竞赛，当我看到数据集的大小时，有点懵，数据集太大了。好在后来还是跑出结果，成功提交了。

目前我只跑出来了两次结果，第一次是最基本的逻辑回归，这次是我自己手敲的代码，对于程序的大致意思都了解，毕竟代码不多，但是细致的每个方法具体做什么的，就不是很清楚了，自己百度查看了方法的作用、参数的意义，做了注解，也了解了一些 nlp 方面的知识。

程序基本思路是：导入数据，对数据做预处理，去除一些不需要的特征，在训练集中去掉 article 和 id 两列，在验证集中去掉 article 列，验证集保留 id 是为了对训练好的模型做验证，然后和 id 做对比，如果一致代表预测正确。然后把字符文本转化成数字向量，使用 CountVectorizer 计算词频，进行矩阵化。之后调用逻辑回归，对训练集进行训练，之后在验证集上验证，根据文本的内容预测文本的类别，最终的结果只保留 id 和 class。提交结果文件，得分为 0.73。第二次是使用 lsa 特征、逻辑回归的，实现的流程和上次的类似，只是在

具体特征时使用的不同的方法。得分是 0.75。还有喝多方法没有尝试,之后会继续进行尝试,说说自己遇到的问题,第一次跑程序的时候报错:process finished with exit code -1073740940,网上没有找到具体的解决办法,问了老师,让我更新 pandas,在 anaconda 里面更新,点击之后没反应,就在终端用 pip 更新的,具体是更新还是卸载之后重装记不太清了,之前一直觉得配环境很麻烦,容易出错,不确定因素很多,犹豫之后决定试试,解决了这个问题,不过尽量还是不要动自己已经装好的环境。还有一点因为数据量很大,程序跑起来要很长时间,某些行程序执行要很长时间,debug 也很难,所以我回在每个阶段让程序输出提示,显示程序进行到哪里了,这样会让自己心里有底。程序看下来大致流程相当,主要区别在于特征提取方式和模型的选择,现阶段还是要多研究不同方法的作用以及参数的设置。深度学习老师使用的是 LSTM,这部分代码我还没有跑过,之后跑完在分析吧,同时猜想能不能使用 cnn 或者 attention 模型来试试呢,先把这些基本的都弄明白在去尝试吧。

嗯 哈哈:

复盘 : 简单写了一个博客记录下我的比赛过程 ,有不对的地方还请指教 https://blog.csdn.net/weixin_41246832/article/details/82889063

Where there's desire:

复盘:之前本科毕设做的是 NLP 相关的新闻分类,用的是 TF-IDF 模型+Word2vec 模型来对新闻标题进行特征建模,之后利用 SVM 模型进行分类。在分类模型上也只了解了 SVM 模型,本次比赛中,利用老师给的 Logistic 回归模型代码跑了结果大概是 0.73 左右,在详细了解了各参数的意义之后,添加了正则项后,发现得分有一点提高。在之后打算利用之前做过的 TF-IDF 模型+word2vec 来表示文本,但是遇上开学以及出差各种事情没来得及提交最终结果。最近时间比较多,希望可以和大家一起学习讨论。

爦:

复盘: 这个比赛目前现在还是沿用老师给的代码,跑了几次分数在 0.77+,但由于机器较旧,时间开销太大,训练一次接近一个小时,后面就没有继续跑了。这次是第一次接触 nlp 的比赛,对特征处理还不太熟,之前稍有接触过信用欺诈类的比赛,同样有一个很明显的感觉是特征才是最关键一个模块。感觉对于不同的场景,选用模型,调参,融合的方法其实大同小异。特征工程却是五花八门需要因地制宜的。所以感觉接下来在学习算法与实践的同时也要多训练提取特征的能力。本人还是个小白,有什么说得不对大家多多指正哈。

Poplar:

复盘:1、特征处理方式(使用 SVM 评估特征):

<1>hash:评估分数为 0.40319343

<2>tfidf:评估分数为 0.78026469

<3>tf:评估分数为 0.69909475 2、对步骤 1 使用嵌入式方式提取特征(LSVC_l2 选择权重大的特征)

<1>hash_select_96:评估分数为 0.31754680

<2>tfidf_select_901290:评估分数为 0.78670366

<3>tf_select_55821：评估分数为 0.71236186　　3、对步骤 2 使用 lsa 提取特征 hash_select_96_lsa_200:只有 96 个特征，lsa 设置降维后的特征个数为 200，200>96，不能执行

tfidf_select_901290_lsa_200：评估分数为 0.75933537

tf_select_55821_lsa_200：评估分数为 0.61060835　第二步和第三步反复降维，导致特征性能越来越差，效果：tfidf>tf>hash。所以在第二阶段只选择 tfidf 并用嵌入式方式提取特征。　　第二阶段特征处理与模型融合

1、特征处理方式(使用 SVM 评估特征)：

<1>word_seg_tfidf：评估分数为 0.78026469

<2>article_tfidf：评估分数为 0.77620433

<3>word_seg_tfidf_select：评估分数为 0.78670366

<4>article_tfidf_select：评估分数为 0.78121792

这 4 个分数不错，开始使用模型训练　LR 评估 word_seg_tfidf.pkl 模型花费 17.902118917306264min,验证集分数：0.7611031189320066

LR 评估 word_seg_tfidf_select_LSVC_l2_901290.pkl 模型花费 9.791357700030009min,验证集分数：0.7608957407801605

LR 评估 article_tfidf.pkl 模型花费 17.423505647977194min, 验证集分数：0.7569512913524576

LR 评估 article_tfidf_select_LSVC_l2_394883.pkl 模型花费 12.556433924039204min,验证集分数：0.7580162458017851　SVM 评估 word_seg_tfidf.pkl 模型花费 4.21601463953654min,验证集分数：0.7802646979637434

SVM 评估 word_seg_tfidf_select_LSVC_l2_901290.pkl 模型花费 2.7235752662022907min, 验证集分数：0.7867036649078799

SVM 评估 article_tfidf.pkl 模型花费 4.143068528175354min, 验证集分数：0.7762043317616406

SVM 评估 article_tfidf_select_LSVC_l2_394883.pkl 模型花费 2.915249772866567min,验证集分数：0.7812179249833233　最后出现一个问题：对特征 article_tfidf_select 过采样操作。验证集分数：0.9056957488006369，训练结束，耗时:4.6130322058995565min。提交结果为 0.77228.出现过拟合现象，修改参数但是没有太大的办法。最终使用 SVM 模型提交结果为 0.773

it's me：

复盘：刚接触这个比赛比较早，当时只是简单跑了一下大佬给的 baseline，0.77+，跑完提交结果,完了就不知道该怎么做了,就搁置了,后来无意中在公众号中看到了这个训练营，然后就报名参赛了，和队友一起，左后结果提到 0.78+，初赛 68，不过复赛惨不忍睹。。希望和大家一起学习，一起 upupup

小地瓜：

复盘：之前学习了一些理论，第一次实践是跟着打达观杯比赛，用了逻辑回归算法和朴素贝叶斯算法，开始的时候手动调参，结果不是很理想；然后改用交叉验证+网格搜索的方法进行调参，但是发现太耗时；打比赛的感觉是调参完全不知道怎么调，全凭试。个人觉得理论是基础，实践很重要，从实践中总结，进行理论升华。

向着光亮那方：

复盘：第一次参加这类比赛，跑了老师提供的代码，成绩 0.73。之后一点一点的看老师写的代码。现在大二，大数据专业，自己在学习人工智能，数据挖掘，刚刚起步，希望在群里学习，进步。

崔自鑫：

复盘：第一次参加这类比赛，没有什么经验，跑了跑 jian 老师提供的代码，成绩 0.778。之后开学了，没什么时间做了。期间主要学习了 lr 模型，svm，word2vec 等。学习了 scikit 等库的使用。走了一遍比赛流程，对 nlp，文本分类有了初步了解。以后应该多学习别人的经验，不断进步。

低碳南：

复盘:第一次参加 nlp 的比赛，平时接触 cv 的多一些，用了 LR，NB，SVM 几个模型，单个的分数都不高，0.76 来，后面因为比较忙也就没有再研究了。平时虽然也看了些机器学习理论的知识，但实践用的时候还是不能很好的应用，代码能力也还需要努力提高，所以参加比赛还是很有帮助的！

CJ-:

【复盘】完全是小白，按照 jian 老师提供的代码敲了一遍，就是 logicstic 回归和 svm 的两种，提交结果在 0.73 左右，首次接触 npl，现在研一应用统计专业，导师研究深度学习算法方向，希望能尽快入门机器学习，有一定的数学基础，但是编程能力严重不足，希望能留在本群做一些力所能及的事情，一般不会打扰到大家，蟹蟹

刘易斯:

[复盘]因为之前出差，所以关注的不是特别多。就是按照 jian 老师的流程走了一遍，模型没有太多的改动，LR 的结果也就 0.712，然后特征方面本来对我很重要。。但是没太多时间弄了。我也是刚开始做机器学习，算法西瓜书刷了，实践非常不够。加之有时差(在巴黎)，所以 follow 的不是很紧，不过我也在努力。主要领域是研究 3d 建模和脑机接口。相对接触 info 的东西少一点。

No. 0758:

【复盘】本次参加训练营主要是为了夯实理论知识，所以这次达观杯的比赛不是很关心排名，因为也没有时间在做，就提交了两次结果。一开始的话按部就班跑了下 0.73 多的样子，后面是做了次特征工程和模型调参，0.769 吧，stacking 和其他的文本特征没有时间再尝试。

【个人背景】研究生刚毕业没多久，毕业论文做的是微博情感分析，用到了文本分类，lda，word2vec 和图网络的构建，目前在互联网金融做数据分析，手上的业务主要还是风控建模，比赛经验的话有在天池上参加过工业数据的比赛，

【期望】由于研究生是自学的机器学习和编程方面的东西，所以感觉理论方面掌握的有

些粗糙，所以想通过训练营一是学习交流下同行的经验，二是积累项目经验，把一些平常工作用不到的算法多应用

林东涛:

我之前接触比较多的是图像处理方向尤其是目标检测,nlp 方面就只知道 lstm 和 GRU,word embedding、word2vec 也只知道个大概,根本没有特征工程的概念。比赛时,首先用群里大佬发的 tf-idf 加 svm 的代码跑了一下大概有 0.769 左右。后来用了 jian 老师的机器学习代码按流程做了特征提取、选择、融合，用 ligthbgm 跑了一遍大概也在 0.77 左右。后来用回 svm 分类,调了一下正则项,能达到 0.776。后来把各种特征融了一遍,效果都差不多。最后尝试跑了一次深度学习方法代码,由于比较慢,跑了一天才十个左右 epoch,就放弃了。通过比赛,我学习了 nlp 比赛的一般流程,也手撕了相关代码,挺有意思的。

繁小星:

现在是研一，在自学西瓜书，以及一些机器学习的书。自学 pytorch ，TensorFlow 以及 caffe 等深度学习的框架。然后 9 月初看到这个带打课程,就想通过这个平台学到更多东西。刚刚接触这些，实践也比较少，所以在公众号看到这个训练营有实战比赛就想都没想就参加了。比赛中，用逻辑回归的方法，最好的分数达到了 0.773，大家都在这个阶段，所以排名比较靠后,不过我会尽力争取。后面打算用 cnn 或者 rnn 的方法。通过这次比赛收获了很多，也很感谢 jian 老师。以后希望能多多参加比赛，吸取经验。

so...:

之前看了一些 kagle 上的比赛代码，书本看的不多，偶尔会看一些博客。达观的这次比赛前期因为电脑跑不动卡了比较久。后面出次提交只有 0.73。成绩比较差。后面增加了特征文本序列特征，大致调了一下 0.75。第一次参赛主要是体验了整个参赛流程，体会到特征对于预测的重要性。特征工程做得好，对模型的依赖会相对减少。小白一枚，希望多多交流指点

Rubick:

机器学习方面小白一枚，之前的学习和课题都与此无关。初来新学校发现实验室的主要研究方向在计算机视觉方面，所以报名参加了这次训练营。开始由于在假期，时间充裕，完成每天任务的同时，还有时间看些其他的东西，感觉节奏并不快。在初期，按照老师的要求学习书上内容，并按照最早的 baseline 参加了比赛，并对机器学习的编程环境有了一定的熟悉。随后自己主要对 baseline 上面的一些参数，按照作业是上面的介绍进行调参。连续提交了十几次，但是结果比最初好的只有一两次。接下来的新的 baseline 由于内存问题，在几种方法分别跑出结果之后，并不能完成融合，所以没有能提交这一版本的结果。虽然还是进行了一些尝试，但最终还是没有解决 8g 内存笔记本跑不动的问题。总的来说，在这几次的改代码、调参数的过程中还是获得了很多收获，远比单纯的看书和自己摸索要好。再之后就开学了，每天留给训练营的时间在减少，但我依然在坚持，希望能在这里有更多的收获。

ShannonGo:

各位好，我之前过了一遍 coursera 上的 andrew ng 的机器学习，然后又过了一遍深度学习的专项课程，算是有了一点基础，但很惭愧感觉效果并不是很好，实践不够。这次就是用 lr,svm 跑了一下，成绩为 0.768，其他的方法还不怎么会....特征工程也不是很了解，以后有问题还望大佬们指教

Mr Li:

复盘：一开始参加比赛的时候在讨论区看到大佬发的 baseline，然后自己运行了一下，成绩大概在 0.777，之后对里面的参数进行了调整，有一些提升，达到了 0.778，也试过其他的一些方法，比如深度学习网络，成绩很差，只有 0.73 左右，也试过 svm，但是数据量太大了，根本跑不动，一直卡着，之后 jian 老师发了他的一些代码，lr,决策树啥的，还有一些降维 pca，模型融合 lgb 等等，也都尝试过，但是也没有达到之前的分数，经过这次比赛，我对这些模型实现有了一定的了解，但是对其中的原理还不是很清楚，以后会多看看公式，了解模型本质

布朗:

接触 python,sklearn 的一些库 是从今年 5 月份自学，也是个小白，当时基本的机器学习算法学过一遍， 但实践不多。刷过 kaggle 上面的数据分析， 最近比较忙，达观杯刷得不多，主要实践过的算法有 logistic,svm, nb； 特征用的 word 的 tf-idf，也实验过 lsa，LDA 跑不起来 ，但是用起来效果不好。实践过的算法参数有研究过；现在线上分数 0.776+，正在实践模型融合和大杀器 lightgbm 和 xgboost,adboost；

Illusions:

防被踢我也回一个吧~达观杯我就是按照老师的要求，调了 lr 和 svm 的包，算出结果就提交了，主要是我自认为文本分类这块有许多我还不了解的地方，比如 lda 原理倒是学过，但这个特征处理方面我也感觉有点无从下手，gensim 也不太会用。至于调参和模型融合，我之前用著名的 iris 数据集试过一些，基本都是用的 sklearn 的 gridsearchcv，所以感觉差不多还是调包。

所以最近我在试着做 kaggle 上面已经完结的最适合入门的房价预测的比赛，主要也是因为这个几乎不需要业务知识，房子属性的好坏大家都了解。首先就很无脑的把所有类别型属性 onehot 编码，然后数值型缺失值就填充上 mean，training set 里房价用 log(x＋1)的函数平滑化了一下，用 ridge 回归就做了一下，然后把结果做了图发现 alpha 取 15 时，均方误差最小(这个 alpha 就是 lr 和 svc 包里的 C 的倒数，l2-norm 权重)。

然后现在在做的就是特征工程——比如对于类别型变量，把 nan 少的特征直接填充成其“众数”，因为 nan 少的我看了一下，的确是特征值缺失;然后 nan 较多的特征，有许多这个 nan 也是个特征取值，比如“房子是否有地下室”这个特征，nan 表示的是无地下室，不是缺失，把这样的列的 nan 用 pandas.get_dummies()方法 one-hot 编码的时候把 dummy_na 设成 true，意即把 nan 项也作为一个特征取值进行 one-hot 编码。

此外，有些类别型特征取值是明显有优劣之分的，比如“房屋外墙用料评级”，分了 5 档，我觉得如果这个进行无脑 one-hot 的话，有可能会丢掉人家给排好的优劣之分的信息，所以想试试把它转为数值型的[1,2,3,4,5].

这些还正在做，晚上打算处理好数据用不同模型再跑一次，顺便做一下模型融合，看看到底有木有提升。

Younha:

简单分享下我的感受。本科跟导师在做深度学习的项目，目前计算机研一，毕业暑假来研究生学校跟着导师项目组学习(项目组主要做 AI+医疗)，开始学习 NLP，文本分类，表示学习等等。报这个训练营，主要是自己想再深入机器学习的基础部分。达观杯比赛最好 0.768+，分别用 LR,RF,NB 结合 TF-IDF 做的，stacking 有考虑不过没时间做(开学事情比较多)。目前待审稿 paper 一篇(RF,LR,Xgboost 的模型组合)，最近在研究 RNN 与聚类，从底层数学公式开始，理解原理，代码实现算法。也希望多参加比赛，让自己加快学习步伐。

刘泽鸿:

研一的时候，自己也看过西瓜书前十一章，虽然大部分还看的懂，也了解一些深度学习的知识，主要是 cnn，对于 nlp 之前没有了解过，这次算是第一次接触。因为实践比较少，所以在公众号看到这个训练营有实战比赛就想都没想就参加了。比赛中，最好的分数达到了 0.778+，最终排名 150 左右，使用的是 SVC 结合网格搜索，最后得到的最优结果参数是 penalty=l1,C=1。后面也尝试过使用 linearRegression，但结果都没有之前的好。通过这次比赛收获了很多，也很感谢 jian 老师。以后希望能多多参加比赛，吸取经验。

陈 sir:

我也复一下我的盘，第一次没有什么技术，有一些感受吧。我是一名研二的学生，就是人工智能方向的，研一学了很多专业课大概有十七八门吧（no NLP），机器学习，神经网络，图像处理，数据挖掘等等，自己也自学了李航统计学、西瓜书之类的，反正就是有一点理论基础吧，但是像打比赛或者做机器学习相关项目的实战几乎没有，也与课业压力有关，反正感觉对学的东西特别虚，不实在，总是处于了解的阶段。

研二定的方向是 NLP 社区问答系统方面的，自己没接触过，也是从头开始，在复现论文的时候，尤其感觉实战应用技能真的是很欠缺，算法理解也不深刻，所以就参加了西瓜书学习，希望在这两个方面做个提升，也希望在研二这一年完成自己的毕业设计，同时参加大型比赛，拿个好名次，增加自己的就业能力和机会吧。

达观杯也是我第一次打比赛，以前总是说说没开始过，这次终于实战了一把，因为毕设的原因这段时间很忙，所以是按照老师给的学习计划在走，没有多做什么工作，但是老师布置的任务都是认真完成的，因为这也是毕设和比赛都用到的知识，达观杯只提交了两次，第一次是逻辑回归算法，第二次是 SVC 算法，现在成绩是 0.777，排名四百基本。这其中收获主要有三点：1.突破了一下自我，以前总觉得比赛很神秘，会特别难，但是当你参与其中的时候，并且认真对待，你会发现你也会成功，并且会变得优秀 2.对算法的理解更加深刻，理解算法我觉得有两点，一个是多看，记住，一个是应用，通过实战对算法在项目中的应用和参数的调整来加深自己对算法的理解。3.实战能力的提高，这也是我现在最亟需的技能了…，不管是在毕设还是比赛，这都是必修…，虽然比赛成绩我只提交了两次，但是我每次都会把每行代码用到的函数，参数等都吃透，包括在复现论文的时候，反正，自我感觉，这对我的实战能力的提高帮助很大。

建议的话跟范晶晶同学的一样，工科这种专业，实战是最好的学习。