唔。。。

第一次参加DataWhale 组织的NLP 学习营，简单记录一下学习过程，也是第一次写学习笔记或者Blog之类的，完全不知道应该怎么写。

第一个任务是关于赛题理解

**赛题名称：零基础入门NLP之新闻文本分类**

**赛题任务：对新闻文本分类，典型的字符识别问题。**

**运行环境**

我计划本次用比较熟悉的CoLab环境，应该很多大神都熟悉的环境，但是第一次写文章应该还是要写细一点，也检查一下我对细节内容的理解。

*CoLab:*

是谷歌开发的用来帮助研究人员或者学生研究和联系机器学习或深度学习 数据分析处理的云端实验环境，是Jupyter notebook 格式，方便扩散分享。用户会有免费的CPU和GPU计算资源，还有专用于Tensorflow的TPU, 付费版比普通版多出10GB的RAM，并且更有可能分配到优质的GPU和TPU资源，可以和google drive 结合方便数据文件的读取，但是付费与否GPU内存都是一定的。

**数据分析:**

数据分为两个部分“

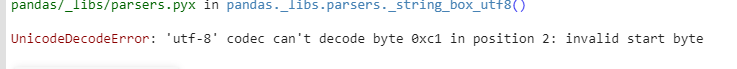
1. Train\_set
2. Test\_set

*第一步，下载数据：*

首先从天池上面下载到的是csv是一数据文件的汇总，简单的上传到Google drive里面，然后利用pydrive和google库文件获取访问权限并且下载数据到CoLab的环境里。

问题一，加载的时候报错:

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xc1 in position 2: invalid start byte



因为一直都是处理的英文数据所以对中文编码不熟悉，查过资料之后发现 中文编码有中国国家标准总局发布的 GBK18030, GBK, GB2313, 以此向右兼容。

于是在 pandas.read\_csv 函数里面设置encoding 参数为 ‘gb18030’ 问题解决

编码相关内容：引至<https://juejin.im/post/5de4b692e51d457a2e2a9d92>

参数设置详见pandas官网文档：

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html>

本来计划直接在colab里面下载数据包并且解压，时间原因 过后在研究怎么操作。目前和第一步一样手动下载数据解压后上传到google drive然后加载到colab环境里

等待过程中研究一下本次任务----理解数据？

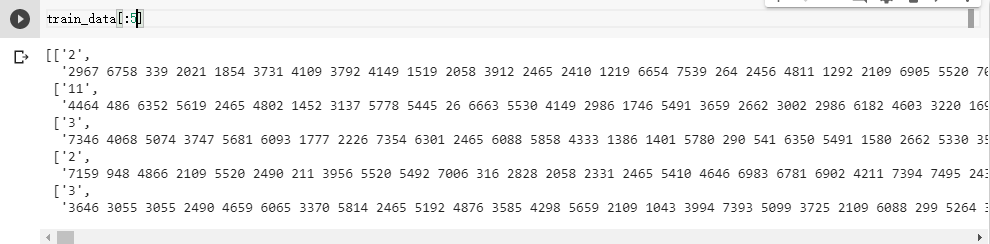
*第二步，研究数据：*

数据是十四个类别的文本分类问题，14个标签分别对应0-13，为预防选手人工标注所以数据进行了字符级别的匿名，因为之前接触到的都是英文的NLP 任务，没有了解过 解题思路中提到的中文分词技术，所以本身没有觉得字符级别编码带来了什么难点，可能后续学习过程中会慢慢体会到，四条思路提示分别是TF-IDF, fasttext，WordVec+ Bert 词向量，前三个都了解过并且应用给英文单词的 representation，抛开字符数字化编码的问题， 我所理解的本赛题，如果后续可以用到RNN本质上就是一个N – M 的seq2seq 模型，可以用到attention分析不同的单词对预测结果的影响度，当然本身IF-IDF就已经能够包含了这一层信息。

思路二 fasttext应该是facebook开发的word2vec的方法，英文里面是把一个单词拆分成不同的部分，每个部分有n个字母，叫做 n-gram,中文里面的话我觉得应该是会把一整个text分成不同的部分，每个部分有n个字符，实现文档级别的编码（有待查证）

Bert词向量 （有待学习）

Train\_set.csv 上传完毕，和加载NLP\_data\_list\_0715一样，把它加载到Colab里面，因为label和text是在一个str里面 用分隔符 ‘\t’ 隔开， 所以需要进一步处理将label和text分开方便后续处理，处理完的结果如图：



因为已经对字符进行了编码所以能够省去对字符的手动数字化操作，还是蛮方便的。

接下来查找学习关于以上提到的未接触到的知识点：

1. 中文分词技术
2. Bert 词向量
3. 如何利用fasttext对中文进行编码

中文分词：

百度百科对于中文分词的定义是讲连续的字符按照一定的规则重新组合成为词序列的过程，由于中文字符每一个都是独立的所以在字符或者语言的数字化表达的过程中会比英文复杂很多，例如：英文单词father 父亲/爸爸 在中文中需要对字符 ‘父‘，’亲‘之间的出现规则进行学习或者用规则规定从而让他组合成一个有意义的中文单词，由于本文档是字符集别加密过的所以没有办法用rule-based方法进行人工中文分词，百度百科里面提供的分类方法有三大类：a>字符匹配；b>理解法； c>统计法

第一个方法类似于 rule-based method，由于字符加密，本赛题不可用。

解释：例如 字典里面有单词 ‘你好‘，’萌宠‘， 如果在文档里面匹配到相对应的字符串，进行切割分类。

第二个方法在百度百科里面没有详细陈述，但是用到理解法应该也是基于rule-based 的方法进行的，后续会查找一些文献仔细研究。

第三个方法虽然在字符级加密的的情况下我认为同样适用，因为不论字符改变成什么样子， p(a|b)是始终不变的，p(a|b)代表字符b出现的情况下字符a接下来出现的概率，因为中文字符会产生不同长度的单词，我认为在利用概率法的时候可以思考一下是否能用到马可夫模型。

总结，基于以往的经验，我的思路大多都是基于神经网络的，但是发现很多大佬直接用机器学习算法进行数据操作也取得了不错的成绩，计划下一步去好好学扎实一些机器学习算法，实验对比ml或者dl方法的区别。