# bug

师弟，attention写好了打算run一下的，结果发现import出错了；看你修改了\_\_init\_\_.py，我现在对这个导包的方法有点懵了；这个你测试通过后，告诉我导包的方式就行；也不是大问题；需要你debug一下。

这个是一个项目的共同导入初始化。例如，我们的module都需要用到torch。你可以再\_\_init\_\_.py中写入

import torch

\_\_all\_\_=['torch’]

再在需要使用的文件中，写入

## Import sys

Sys.path.append(‘./’) #把init的路径加载到项目中，

## From module import \* #就是把init的所有内容导入到这个文件中

cd /data/wanyao/Dropbox/ghproj-titan/naturalcodev2

记得先根据需要修改yml文件再运行。

Train：

nohup python -u run/baseline/baseline.py > log/baseline.log

import出问题了。。

minor bugs:

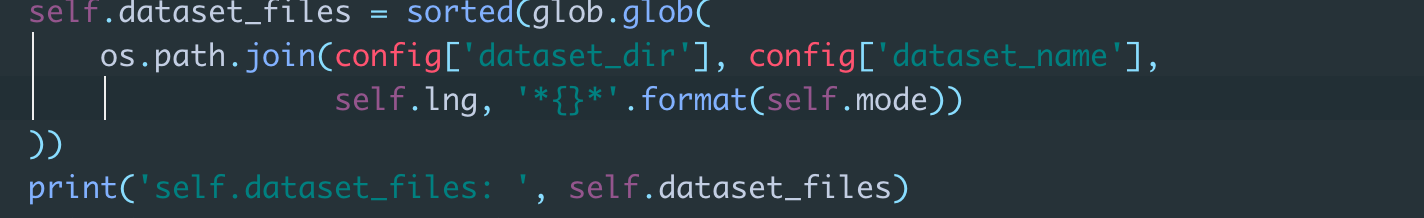
1. Constants: <blank> -> <PAD>, <unk> -> <UNK>, <s> -> <BOS>, </s> -> <EOS>

感觉PAD比blank更常见；<s>这种表示有点太简单，因为代码中也可能出现<s>这个token我觉得

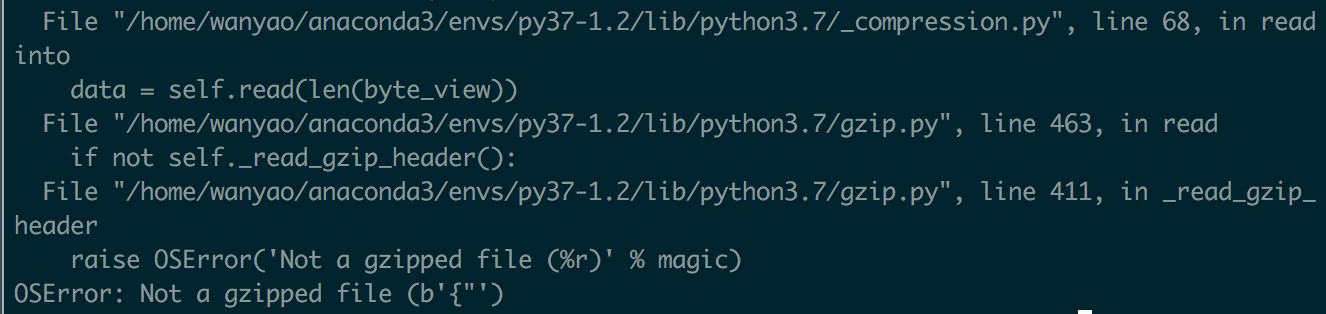
2. uni.py应该是dataloader，所以其类命名为UniDataloader更合适？同样的，tl.py的类命名为TLDataloader更合适

3. dataset中pop这个能去掉么？所有的情况下都保留一样的keys，如果某个key不存在，用{}表示？

5. 在base.py \_construct方法中



这里的file必须是.gz才可以的，一旦我将其解压后就变成.jsonl的文件了，然后就出 bug。



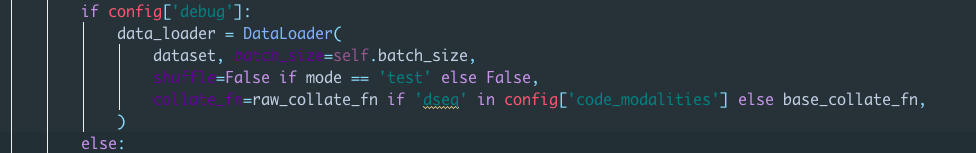
这里应该把文件名的构造变得更加严格一点，或者加一个exception。或者支持一下读jsonl也可以的了。

6. .pyc文件不用push上来的，github客户端好像可以设置ignore文件的

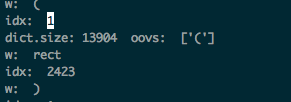


7. module下的Model.py可以删掉的？如果是有用的文件名应该是小写的才对

8. uni.py debug模式，shuffle始终保持False



9. 在调试util.py extend\_dict的时候发现， ‘(’,’)’’:’ 等都不在comment的dict中，很奇怪，难道这么多comment中没有这几个符号？还是标点符号都被处理掉了呢？

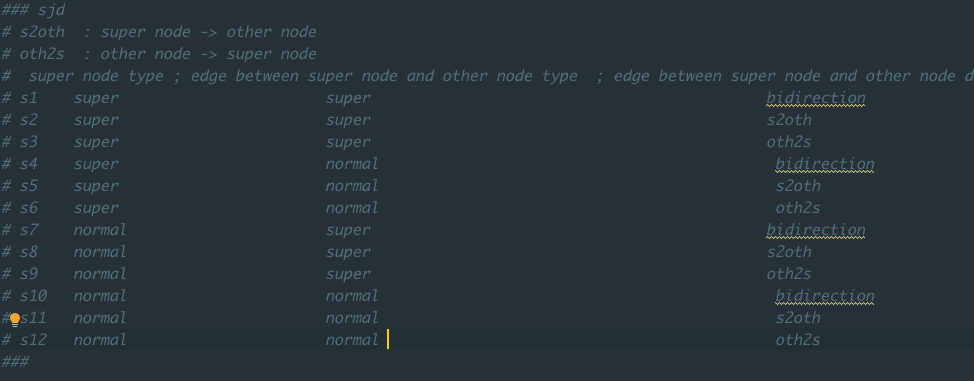


10. 每个yml文件对应一个log文件，log文件的名字应该包括主要的参数，例如batch\_size, learning\_rate, run\_date 等等；这个怎样设置一下比较好了？

11. 尽量少用import \*，调用了什么就import什么比较好，例如经常用到了一个util，我都没法定位到这个util在哪个文件；如果from xxx import xxx的话，这样我就知道这个util函数是在哪个文件里面了。

12. attention=>attn\_type 这个和attention源码中的命名保持一致，也包含更多信息；现在的attention主要实现在GlobalAttention里面，是copy自一个源码

13. 这个可以移动到doc中去



14. code的表示文件夹encoder->code2vec,encoder格局太小，给人一种感觉就是encoder-decoder中的encoder，这个文件夹是要做代码各种表示学习的，包括以后的Bert也是这个下面，是一个fundamental的东西，以后很多上层任务都会用到，不止encoder-decoder框架。

15. import这种方式我觉得用起来还是不是很顺手啊，例如



我想看get\_args在那个文件当中定义的，我必须得通过ctrl+f在工程中搜才可以。现在发现是在\_\_init\_\_.py中，所以这个问题还好。

学长告诉你怎么用pycharm，你按住ctrl，再点击那个你不懂的东西。你就可以进去了

16. task的定义，尽量和我们论文中的名字统一，尽量做到我们的每篇论文都能找到task，即一个task对应一篇论文。Task：MM2Seq （包括Seq2Seq，Tree2Seq，Hybrid2Seq等baselines），DeepCom，Code2Seq，TL，DTRL，KD，SMILE（这个才是我们的论文模型名，基于maml的）

这个按照你说的办

17. Log文件的时间力度没必要搞到分钟，搞到天就行了（同一天运行的程序是一个log文件），因为有时候我们会放一些低级错误，会重新run文件；这样的话没运行一次都生成一个log文件，很多都会是无用的。

原本的log文件就是非常麻烦的，应为这些东西必须保证不一致才行。由于我们之间可能会出现时间冲突的情况， 所以我+了md5码在后面。而且log文件保护的信息太少了，只能想办法隔离。时间精确的可以帮助我们快速回忆训练过程，影响不大的。

这个不建议修改

18. nohup python -u /data/wanyao/Dropbox/ghproj-titan/naturalcodev2/run/sl/main.py --yaml sl-ruby.yml --task sl --task\_mode train --multi\_processing True > /data/wanyao/Dropbox/ghproj-titan/naturalcodev2/run/sl/SL\_TRAIN\_2020-Jan-14\_03-27-48.d41d8cd98f00b204e9800998ecf8427e.log 2>&1 &

2>&1 &要不去掉？不然我想停掉nohup文件，还得kill，之前没有这个我直接ctrl+c就可以终止程序的。

nohup可以根据个人喜好选择，我建议个人定义不用的main来区别

19. yml里面可以再按模态网络结构再抽一层出来，例如

lstm: （for Seq）

* rnn\_layer\_num
* rnn\_hidden\_size

tree-lstm: （for AST）

- tree\_leaf\_subtoken

- tree\_lstm\_cell\_type

ggnn: ～（for CFG）

这个可以的根据你的喜好修改

这个暂时改，现在没有时间，而且可以用

20 global\_attention.py不应该放在codesum文件夹下，因为attention是一个通用的模块，所以我单独搞了个attention文件夹，专门实现各种attention的。我现在搞不清哪个attention是我昨天修改了的了。你说那个是自己找了一个，但是里面又有我备份的一个forward\_()函数，待会确认下。

这个global\_attention.py你可以根据自己的要求修改，我之后再阅读你的代码

21. src目录下的log，我觉得可以直接把get\_logger当作utils里面的一个函数；这样就可以把src/log目录删掉了。

暂时不改

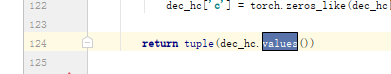
sjd=========

## 22 values

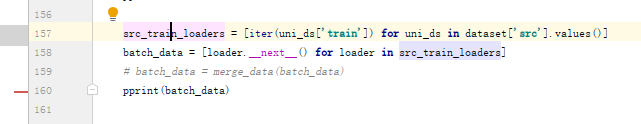
不同python版本，对于相同dict，通过.values（）得到的value的排列顺序可能不一样，所以对于可能被顺序影响的情况，应该避免使用.values, 应该按照设定的顺序，用key取出value



naturalcodev2/src/module/code2vec/multi\_modal/mm\_encoder.py

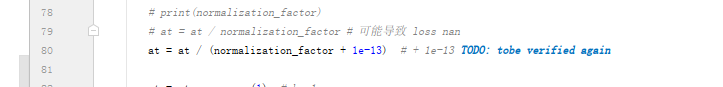


naturalcodev2/src/dataset/xlang\_dataloader.py

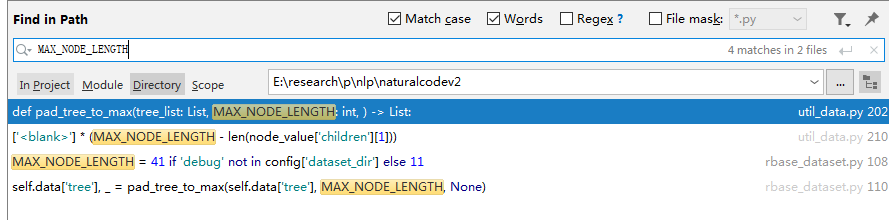


## 23 intra attention 添加的小数字改成clamp？

naturalcodev2/src/module/attention/intra\_attention.py



## 24 不同语言也是同样的MAX\_NODE\_LENGTH的？



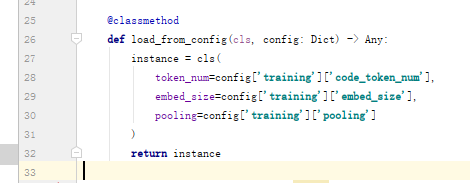
## 25 self.data['tok'] 和 self.data['code']是一样的吗？ 有啥问题吗？

naturalcodev2/src/dataset/base/sbase\_dataset.py



## 26 code\_token\_num和pooling在ruby.yml里没有？

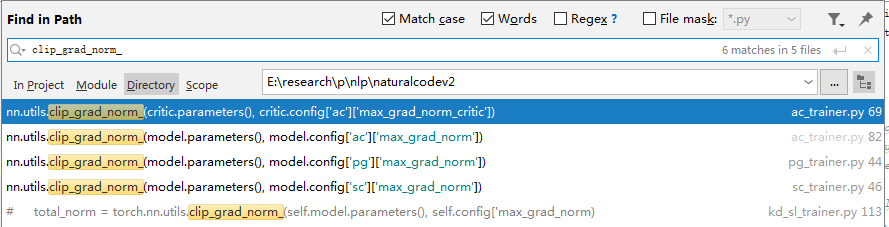
naturalcodev2/src/module/code2vec/base/emb.py



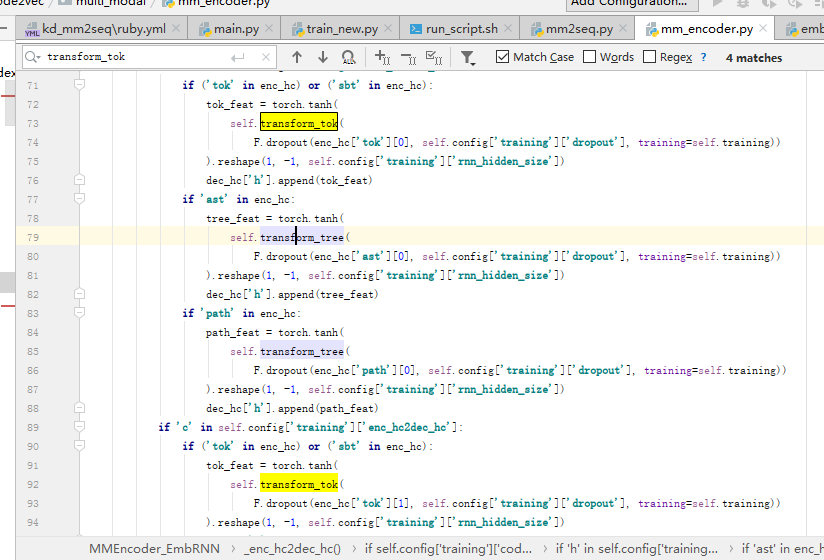
## 27 梯度裁剪

一般用不着裁剪，

建议代码里加个判断，max\_grad\_norm<0或者为None时，不裁剪梯度



## 28 transform\_tok在c和h之间共享？transform\_tree在ast和path之间共享，c和h之间共享？

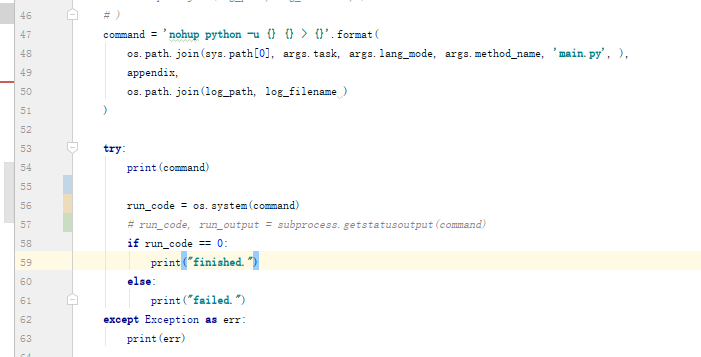


==============

# 已解决

## logging 输出定向到文件

需要使用nohup，且 getstatusoutput 的方式不行，要用os.system

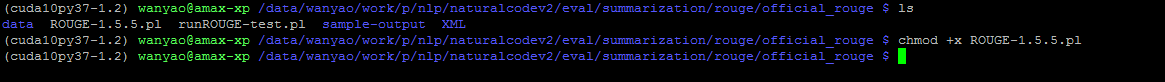


## 安装rouge

sudo apt-get install libxml-parser-perl

有依赖问题就

sudo aptitude install libxml-parser-perl

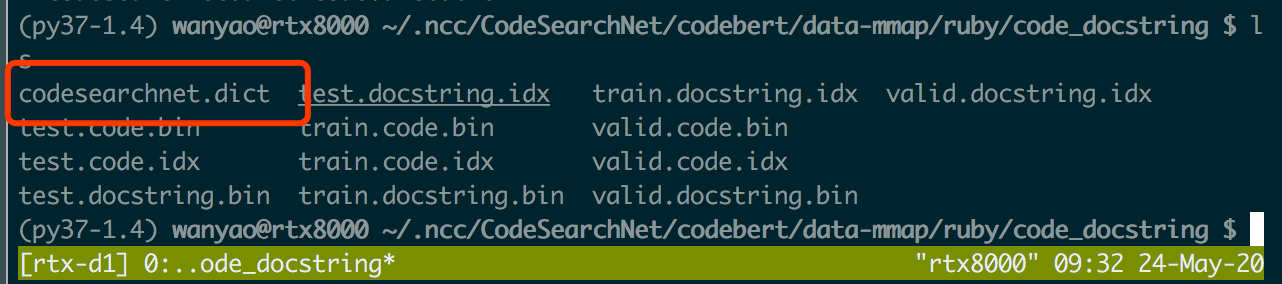


sudo perl -MCPAN -e 'install XML::Parser'

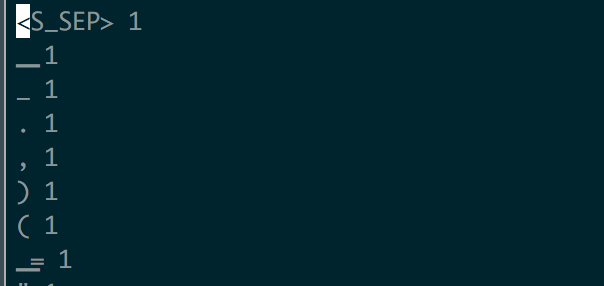
## 解除cuda异步，利于调试

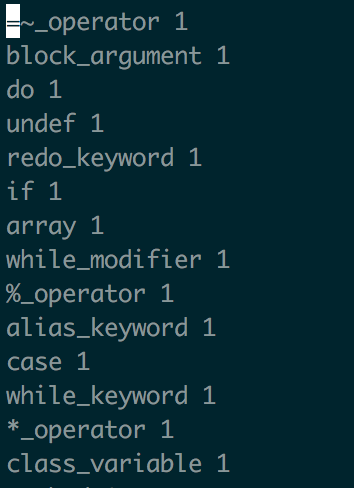
CUDA\_LAUNCH\_BLOCKING = 1

1. 文件命名现在是docstring\_path，改成path\_docstring更好一点，和code\_docstring保持一致
2. codesearchnet.dict 改名成code.dict.txt (code为source\_lang的名字，这个应该有个joined\_dictionary的判断，如果 joined\_dicstionary为1，则生成一个code.dict.txt即可；如果joined\_dicstionary为0，则生成两个个dict文件，docstring.dict.txt即可)。.txt后缀是为了说明这个是txt文件，不是字典文件。这个也是尽量和fairseq保持一致。

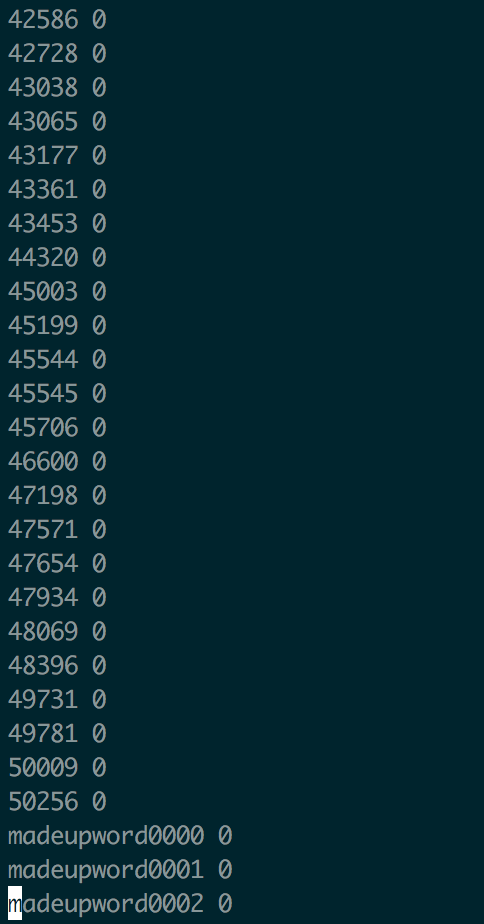


1. code\_docstring下面的codesearchnet.dict和 docstring\_path下面的codesearchnet.dict和不同。一个考虑了special token，一个没有。最好统一成都不考虑special token。Special token在setup­ task的时候手动add\_symbol.





1. src不要缩写了，和source\_lang保持一致，用source。tgt也一样。
2. 进入wiki\_roberta看下fairseq自带的数据处理生成的数据长啥样，他们的dict.txt有padding的，主要是为了加速用的。



1. xxx
2. xxx