xgli的博客

想让博客有个排名，去掉“千里之外”。联系方式：xgli0807@gmail.com

目录视图 摘要视图 RSS 订阅

文本分类（六）：使用fastText对文本进行分类--小插曲

标签： 文本分类-深度学习 fasttext

2016-10-28 21:44 14536人阅读 评论(52) 收藏

分类：

python（26） ML（11） NLP（8）

环境说明：python2.7、linux

自己打自己脸，目前官方的包只能在linux，mac环境下使用。

测试facebook开源的基于深度学习的对文本分类的fastText模型

fasttext python包的安装:

```
1 pip install fasttext
```

第一步获取分类文本，文本直接用的清华大学的新闻分本，可在文本系列的第三篇找到下载地址。

输出数据格式： 样本 + 样本标签

说明：这一步不是必须的，可以直接从第二步开始，第二步提供了处理好的文本格式。写这一步主要是为了记忆当时是怎么处理原始文本的。

```
import jieba
import os

basedir = "/home/li/corpus/news/" #这是我的文件地址，需跟据文件夹位置进行更改
dir_list = ['affairs','constellation','economic','edu','ent','fashion','game','home','house','lottery','
##生成fastext的训练和测试数据集

ftrain = open("news_fasttext_train.txt","w")
ftest = open("news_fasttext_test.txt","w")

num = -1
for e in dir_list:
    num += 1
    indir = basedir + e + '/'
    files = os.listdir(indir)
    count = 0
    for fileName in files:
        count += 1
        filepath = indir + fileName
        with open(filepath,'r') as fr:
            text = fr.read()
        text = text.decode("utf-8").encode("utf-8")
        seg_text = jieba.cut(text.replace("\t"," ").replace("\n"," "))
```

文本分类（五）：使用LDA进行文本的降…
DCX_abc ：666

文本分类（六）：使用fastText对文本进行…
lxg0807 ：@qq_25666147:重新更正了一下代码，最后一个只是一部分代码

文本分类（六）：使用fastText对文本进行…
wzl2015001 ：@Jemila:同问

tensorflow GPU使用问题
wzl2015001 ：博主你好，请问你有没有弄过两块GPU并行运算，怎么弄你知道不？

文本分类（六）：使用fastText对文本进行…
fahaihappy ：@Jemila:请问下 你解决这个问题了吗? zai windows下遇到同样的问题了

文章分类

python (27)
leetcode (0)
ML (12)
NLP (9)
DM (2)
c (2)
java (7)
一周总结 (0)
linux (6)
elsearch (3)
BD (1)
php (1)
tensorflow (9)
总结 (0)
算法 (1)

```
24      outline = " ".join(seg_text)
25      outline = outline.encode("utf-8") + "\t__label__" + e + "\n"
26  #       print outline
27  #       break
28
29      if count < 10000:
30          ftrain.write(outline)
31          ftrain.flush()
32          continue
33      elif count  < 20000:
34          ftest.write(outline)
35          ftest.flush()
36          continue
37      else:
38          break
39
40  ftrain.close()
41  ftest.close()
```

第二步：利用fasttext进行分类。使用的是fasttext的python包。

整理好的数据：百度网盘下载

news_fasttext_train.txt

news_fasttext_test.txt

```
1  # _*_coding:utf-8 _*_
2  import logging
3  logging.basicConfig(format='%(asctime)s : %(levelname)s : %(me
4  import fasttext
5  #训练模型
```

```
6 classifier = fasttext.supervised("news_fasttext_train.txt","news_fasttext.model",label_prefix="__label__")
7
8 #load训练好的模型
9 #classifier = fasttext.load_model('news_fasttext.model.bin', label_prefix='__label__')
```

```
1 #测试模型
2 result = classifier.test("news_fasttext_test.txt")
3 print result.precision
4 print result.recall
```

```
0.92240420242
0.92240420242
```

由于fasttext貌似只提供全部结果的p值和r值，想要统计不同分类的结果，就需要自己 [illegible] 现了。

```
1  # -*- coding: utf-8 -*-
2  """
3  Created on Wed Oct 18 14:17:27 2017
4
5  @author: xiaoguangli
6  """
7  import logging
8  logging.basicConfig(format='%(asctime)s : %(levelname)s : %(me
9  import fasttext
10
11
12 classifier = fasttext.load_model('news_fasttext.model.bin', label_p
13 labels_right = []
```

```
texts = []
with open("news_fasttext_test.txt") as fr:
    for line in fr:
        line = line.decode("utf-8").rstrip()
        labels_right.append(line.split("\t")[1].replace("__label__",""))
        texts.append(line.split("\t")[0])
    #     print labels
    #     print texts
#     break
labels_predict = [e[0] for e in classifier.predict(texts)] #预测输出结果为二维形式
# print labels_predict

text_labels = list(set(labels_right))
text_predict_labels = list(set(labels_predict))
print text_predict_labels
print text_labels

A = dict.fromkeys(text_labels,0)  #预测正确的各个类的数目
B = dict.fromkeys(text_labels,0)   #测试数据集中各个类的数目
C = dict.fromkeys(text_predict_labels,0) #预测结果中各个类的数目
for i in range(0,len(labels_right)):
    B[labels_right[i]] += 1
    C[labels_predict[i]] += 1
    if labels_right[i] == labels_predict[i]:
        A[labels_right[i]] += 1

print A
print B
print C
#计算准确率，召回率，F值
for key in B:
```

```
45    try:
46        r = float(A[key]) / float(B[key])
47        p = float(A[key]) / float(C[key])
48        f = p * r * 2 / (p + r)
49        print "%s:\t p:%f\t r:%f\t f:%f" % (key,p,r,f)
50    except:
51        print "error:", key, "right:", A.get(key,0), "real:", B.get(key,0), "predict:",C.get(key,0)
52
53
```

实验数据分类

```
[u'affairs', u'fashion', u'lottery', u'house', u'science', u'sports', u'game', u'economic', u'ent', u'edu', u'hom                    on',
['affairs', 'fashion', 'house', 'sports', 'game', 'economic', 'ent', 'edu', 'home', 'stock', 'science']
{'science': 8415, 'affairs': 8257, 'fashion': 3173, 'house': 9491, 'sports': 9739, 'game': 9506, 'economic': 923                    'edu
{'science': 10000, 'affairs': 10000, 'fashion': 3369, 'house': 10000, 'sports': 10000, 'game': 10000, 'econom                    10
{u'affairs': 8562, u'fashion': 3585, u'lottery': 96, u'science': 9088, u'edu': 10068, u'sports': 10099, u'game': 10151, u'economi
```

#实验结果

```
1  science:  p:0.841500  r:0.925946r:  f:0.881706
2  affairs:  p:0.825700  r:0.964377r:  f:0.889667
3  fashion:  p:0.941822  r:0.885077r:  f:0.912568
4  house:  p:0.949100  r:0.949100r:  f:0.949100
5  sports: p:0.973900  r:0.964353r:  f:0.969103
6  game:  p:0.950600  r:0.936459r:  f:0.943477
7  economic:  p:0.923500  r:0.911559r:  f:0.917490
8  ent:  p:0.966500  r:0.895073r:  f:0.929416
9  edu:  p:0.949100  r:0.942690r:  f:0.945884
```

```
10 home:  p:0.931500  r:0.922003r:  f:0.926727
11 stock:  p:0.901500  r:0.878998r:  f:0.890107
```

从结果上，看出fasttext的分类效果还是不错的，没有进行对fasttext的调参，结果都基本在90以上，不过在预测的时候，不知道怎么多出了一个分类constellation。难道。。。。查找原因中。。。。

2016/11/7更正：从集合B中可以看出训练集的标签中是没有lottery和constellation的数据的，说明在数据准备的时候，每类选取10000篇，导致在测试数据集中lottery和constellation不存在数据了。因此在第一步准备数据的时候可以根据lottery和constellation类的数据进行训练集和测试分，或者简单粗暴点，这两类没有达到我们的数量要求，可以直接删除掉

- 上一篇 git命令大全
- 下一篇 python numpy.tile函数

相关文章推荐

- NLP丨高级词向量表达（二）——FastText（简述、..
- MySQL在微信支付下的高可用运营--莫晓东
- fastText原理及应用
- 容器技术在58同城的实践--姚远
- fastText原
- 腾讯云容器
- FastText
- 微博热点事

- fasttext的基本使用 java 、python为例子
- SDCC 2017之容器技术实战线上峰会
- fastText Ngram 的处理过程
- SDCC 2017之数据库技术实战线上峰会
- fasttext
- fasttext使用笔记
- Facebook：FastText 理解和在query意图识别的应用
- [转]Facebook 开源的快速文本分类器 FastText

查看评论

kezaoshao0675 18楼 2017-12-05 17:

请问楼主？windows下搭建fasttest环境可以解决了么？卡了好久了

Jemila 17楼 2017-09-27 16:25发表

好奇怪为什么我预测的是nan?两份数据是来自楼主的，从头到尾都没错就是predict 和recall都是nan

Jemila

回复Jemila：米娜桑，fasttext要在linux系统下才可以运行，在windows下缺少一个编译文件

neoson2015

回复Jemila：我也是耶

wzl2015001 Re: 2017-10-14 11:43发表

回复Jemila：同问

fahaihappy Re: 2017-10-10 15:20发表

回复Jemila：请问下 你解决这个问题了吗? zai windows下遇到同样的问题了

lya1224 16楼 2017-09-18 15:

楼主，我用你的语料和代码测试了一下，怎么p和r不稳定，好的时候0.91，不好的时候是0.86，这是什么原因呢

fancy1010 15楼 2017-09-05 15:16发表

楼主，我打算做新闻推荐，大概先将新闻分为10个类左右，再对每个分类下新闻进行聚类，话题生成。请问下这个适用吗，并且数据量需要多大呢

俞驰

load_model这句是不可运行的，会出现如下报错：
Traceback (most recent call last):
File "/home/appleyuchi/下载/THUC/THUCNews/csdn_example
odel_test.py", line 11, in <module>
classifier = fasttext.load_model('model', label_prefix=&#3
_label__')
File "fasttext/fasttext.pyx", line 154, in fasttext.fasttext.loa
model
Exception: fastText: Cannot load model due to C++ extension failed
allocate the memory
然后，github上的解决方案，在linux下面也是不行的：
https://github.com/salestock/fastText.py/issues/125

huangrs12 Re: 2017-08-30 16:34发表

回复俞驰：请问fasttext有没有跑出来呀，我跑出来的准确率特别低。同样的训练数据和测试数据svm的准确率都有0.78fasttext是准确率只有0.126

lxg0807 Re: 2017-08-31 20:01发表

回复huangrs12：这个svm需要提取特征，你说的fasttext准确率这么低，应该是有问题的，你是几分类呀

huangrs12 Re: 2017-09-01 10:

回复lxg0807：总共有13个类别，7000条训练数据。不知道问题出现在哪了

lxg0807 Re: 2017-09-02 22:0/发表

回复huangrs12：可能是你的训练数据集太少了，fasttext算是浅层的神经网络，如果你的数据集不够大，可能svm确实会比较好

lxg0807

回复俞驰：有没有可能是你的gcc版本问题呀？

俞驰

我看了下数据集接近两G啊

俞驰

news_fasttext.model.bin文件在哪里有啊？谢谢

---

qq_36958979 11楼 2017-06-14 17:08发表

用Windows搭建环境，已下载清华大学新闻文本语料以及训练文本和测试文本，程序运行后报错如下，请教下博主如何解决，多谢多谢。
Traceback (most recent call last):
File "C:/Users/Administrator/PycharmProjects/untitled/satest.py", line 23, in <module>
files = os.listdir(indir)
FileNotFoundError: [WinError 3] 系统找不到指定的路径。: 'C:/Users/Administrator/Desktop/THUCNewsaffairs/'

lxg0807 Re: 2017-06-14 17:47发表

回复qq_36958979：您好，需要更改路径的，可以直接从第二步开始。

lxg0807 Re: 2017-06-15 09:57发表

回复lxg0807：你好，对不起误导您了，这个pip安装这个包只能是在linux和mac下面。windows需要特殊安装。

---

qq_36958979

回复lxg0807：我根据Windows的指引下载了visual++后成功pip install了fasttext，并且修改了清华大学标注语料的目录地址，跑程序遇到很多问题。

---

qq_36958979

楼主最后的问题解决了吗，请问？
我跑出的结果也是NAN

Jemila Re: 2017-09-27 16:36发表

回复qq_36958979：我的结果也是nan，请问解决了么

qq_36958979 9楼 2017-06-14 15:

好的，谢谢大神，btw请教下已标注的文本语料有大小的要求吗？

huangrs12 Re: 2017-08-30 16:

回复qq_36958979：请问下你的训练数据多大啊，我的训练数据才800K，训练的效果特别的不好。

shuaya001 8楼 2017-06-12 17:38发表

在windows下python3使用fasttext有bug，在github上看官方没有修复这个bug，测试在mac下无问题

qq_36958979

博主用fasttext 做过短文本分类吗？

huangrs12

回复qq_36958979：请问你做的短文本分类效果怎么样，我也是要做短文本分类。一句话大概10个字以内

lxg0807 Re: 2017-06-12 16:43发表

回复qq_36958979：用过fasttext做过句子的二分类。效果一般吧。 可以根据你的业务试试。

程序猿进化之旅 6楼 2017-05-18 17:48发表

楼主是在windows 上搭建的环境吗？

lxg0807 Re: 2017-06-12 16:

回复程序猿进化之旅：是linux平台，但是这个应该与平台无关吧 处理好文本的编码就好

程序猿进化之旅 Re: 2017-06-14 21:

回复lxg0807：windows 上也需要gcc 4.7以上吧，我没试过，楼主没有考虑在windows 上测试下？给我们带带路啊！

lxg0807 Re: 2017-06-15 09:57发表

回复程序猿进化之旅：你好，对不起误导您了，这个ip安装这个包只能是在linux和mac下面。

程序猿进化之旅

回复lxg0807： 楼主,linux 平台上运行到labels dict = [e[0] for e in classifier.predict(texts)]这一住了，是什么原因呢？

qq_25666147

问下 我编译 最后的一个 提示 NameError: name 'classifier' is not define
d 需要import 什么不

qq_25666147 Re: 2017-05-16 12:08发表

是我粗心了 应该吧 第前面训练的 要导入
添加 classifier = fasttext.load_model('news_fasttext.model.
bin', label_prefix='__label__')

qq_25666147 4楼 2017-05-16 11:

我编译最后一个 提示 NameError: name 'classifier' is not defined 请问
要怎么解决

lxg0807 Re: 2017-10-18 15:07发表

回复qq_25666147：重新更正了一下代码，最后一个只是一
部分代码

末世大金矿

您好。处理文本的时候应该过滤特殊符号和停用词。

zljcrazy

对语料的量有参考值吗？好像小样本，效果特别差

qq_26599263

您这篇文章写的太棒了，最近我也在看fasttext，但是并不知道进行完
本分类后有什么具体的应用

lxg0807 Re: 2016-11-08 15:04发表

回复qq_26599263：或许可以做新闻推荐，这个是个分类模型，可以根据你的任务来做。

qq_26599263 Re: 2016-11-30 11:44发表

回复lxg0807：博主有没有做过用fasttext进行标签预测的测试，标签预测时测试集的tags也是用__label__做前缀吗

发表评论

用 户 名： weixin_35068028

评论内容：

提交

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场