NLP（二十）利用BERT实现文本二分类

原创 jclian Python爬虫与算法 2月12日

收录于话题
#NLP入门系列文章

29个

在我们进行事件抽取的时候，我们需要触发词来确定是否属于某个特定的事件类型，比如我们以政治上的出访类事件为例，这类事件往往会出现“访问”这个词语，但是仅仅通过“访问”这个触发词来判断是否属于出访类事件是不可靠的，比如我们会碰到以下情况：

通过上面的例子，我们知道，像访问速度，访问量这种文档虽然出现了访问，但却不属于政治上的出访类事件。因此，这时候我们需要借助 文本分类 模型来判断，显然，这是一个二分类模型。

本文将会讲述如何利用BERT+DNN模型来判断文档是否属于政治上的出访类事件。

数据集

笔者找了300个文档，里面的文档都含有“出访”这个词语，标签1表示属于政治上的出访类事件，标签0则不是。将数据集分为训练集（250个样本）和测试集（50个样本），比例为5:1，样本不是很多，但借助BERT，我们可以在小样本上取得不错的效果。

训练集（部分）的样本如下：

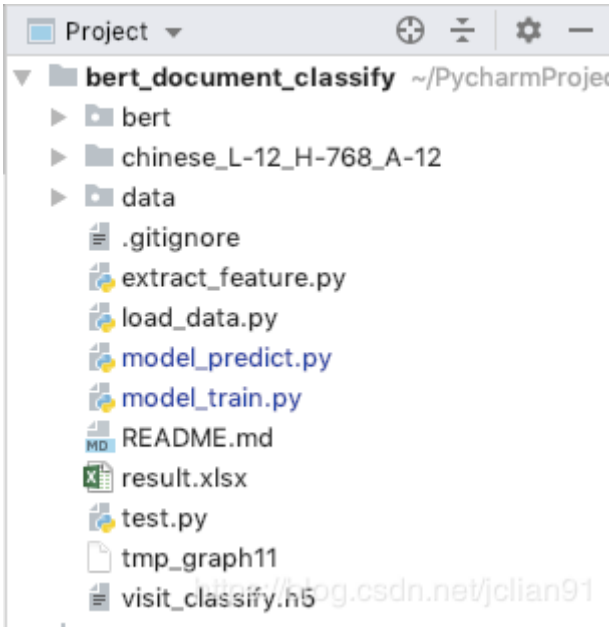
train.txt

```
1 当地时间2月10日，白宫发表声明称，美国总统特朗普及夫人梅拉尼娅将于2月24日至25日访问印度，
0 俄罗斯卫星通讯社11日最新消息，菲律宾总统杜特尔特已下令终止与美国间的《访问部队协定》（VFA
1 据俄罗斯卫星网6日报道，土耳其总统发言人卡林表示，俄罗斯军事代表团将于近日访问安卡拉，讨论
0 先来说说什么是LPDDR5：要知道，手机中有两种内存颗粒，一种就是DRAM也就是大家常说的“运行内
1 在疫情的关键时刻，出现了一件令人感动的事情，让我们明白这才是真正的好朋友，不惧疫情访问我国
0 此次疫情的催化作用下，人们的就医习惯将迎来大改变，线上问诊率也有望大幅提升至10%；多个互联
1 在目前中国处于紧急状态，最需要国际朋友支持的敏感之际，洪森对中国的访问，凸显出柬埔寨王国政
1 当地时间2月3日，就在驻叙利亚的土耳其军队遭到叙政府军炮击的当天，土耳其总统埃尔多安访问乌克
1 当地时间3日，为缓和两国关系，美国高级官员低调前往伊拉克。
1 德国总理默克尔7日结束对南非的国事访问。
0 一场疫情，让线上医疗平台迎来了访问峰值。阿里健康数据显示，其提供在线医疗平台的轻问诊服务上
1 俄罗斯外长拉夫罗夫7日在委内瑞拉首都加拉加斯表示，尽管美国对委实施“非法制裁”，俄罗斯仍将继
0 豫企“在线办公”需求激增，搜索访问量日均增长近7成
0 美光科技今日宣布已交付全球首款量产的 LPDDR5 DRAM 芯片，并将率先搭载于即将上市的小米10智
0 为应对严峻的疫情防控形势，科学有序做好广大居民及返岗返工人员的健康监测服务，经过两天两夜不
0 “武汉加油，买菜有我”。2月7日，由长江日报发起的“社区团购蔬菜”活动正式上线“长江严选”电商平
1 列位欢迎来到照理拍案，前两天美国国务卿蓬佩奥，访问了白俄罗斯，见到了白俄罗斯总统卢卡申科，
1 台湾地区准副领导人赖清德昨天赴美国智库“哈德逊研究所”会谈，知情人士转述与会的美国国防部长
1 据共同社最新报道，日本首相安倍晋三决定将于今年对俄罗斯进行访问，参加庆祝胜利日的庆祝活动。
```

训练集部分数据

代码

本项目的结构如下：



项目结构

因为我们这边是小样本量，所以需要用到BERT。又因为是中文，所以需要下载BERT的中文训练文件 chinese_L-12_H-768_A-12 ，这是已经训练好的模型文件。

根据我们在文章NLP（十九）首次使用BERT的可视化指导中的经验，我们需要写代码来调用BERT模型文件，比如tokenizer，padding，masking以及BERT模型产生输出向量等，幸运的是，有人已经帮助我们做好了这件事，我们只需要调用其代码就行了。这部分的代码位于bert文件夹下，读者可以在文章最后的Github地址上找到。因为本文的模型为文本分类模型，所以需要取[CLS]这个token所对应的768维的向量。

接下来，我们先读取数据集，处理成训练集和测试集，脚本为load_data.py，完整的Python代码如下：

```
# -*- coding: utf-8 -*-
# author: Jclian91
# place: Pudong Shanghai
# time: 2020-02-12 12:57
import pandas as pd


# 读取txt文件
def read_txt_file(file_path):
    with open(file_path, 'r', encoding='utf-8') as f:
        content = [_.strip() for _ in f.readlines()]

    labels, texts = [], []
    for line in content:
        parts = line.split()
        label, text = parts[0], ''.join(parts[1:])
        labels.append(label)
        texts.append(text)

    return labels, texts
```

```
file_path = 'data/train.txt'
labels, texts = read_txt_file(file_path)
train_df = pd.DataFrame({'label': labels, 'text': texts})

file_path = 'data/test.txt'
labels, texts = read_txt_file(file_path)
test_df = pd.DataFrame({'label': labels, 'text': texts})

print(train_df.head())
print(test_df.head())

train_df['text_len'] = train_df['text'].apply(lambda x: len(x))
print(train_df.describe())
```

输出结果如下：

```
  label                                               text
0     1  当地时间2月10日，白宫发表声明称，美国总统特朗普及夫人梅拉尼娅将于2月24日至25日访问印...
1     0  俄罗斯卫星通讯社11日最新消息，菲律宾总统杜特尔特已下令终止与美国间的《访问部队协定》(...
2     1  据俄罗斯卫星网6日报道，土耳其总统发言人卡林表示，俄罗斯军事代表团将于近日访问安卡拉，...
3     0  先来说说什么是LPDDR5：要知道，手机中有两种内存颗粒，一种就是DRAM也就是大家常说的“...
4     1  在疫情的关键时刻，出现了一件令人感动的事情，让我们明白这才是真正的好朋友，不惧疫情访问...
  label                                               text
0     1  应巴基斯坦总理伊姆兰·汗、荷兰王国首相吕特、德国联邦政府邀请，国家副主席王岐山将于5月2...
1     1  联邦德国总理默克尔抵达印度进行访问，在雾霾笼罩下的新德里受到军人仪仗队的欢迎。默克尔赞...
2     1  5月6日至12日，省委副书记乌兰率代表团访问韩国、泰国，与韩国国际交流联盟、新村运动中央会...
3     1  国台办发言人马晓光今天（5月22日）表示，新党主席、新中华儿女学会荣誉理事长郁慕明将率台...
4     1  6月13日至15日，联合国反恐事务副秘书长沃伦科夫应邀访问北京和新疆，并与中国外交部副部长...
         text_len
count  250.000000
mean    77.540000
std     36.804493
min     11.000000
25%     47.500000
50%     73.000000
75%    100.750000
max    192.000000
```

可以发现，训练数据集的文本长度的75%分位点为100.75，所以我们在模型训练的时候，padding过程中的统一长度取100。

数据预处理之后，我们利用BERT提取文档的特征，每个文档的填充长度为100，对应1个768维的向量，然后用Keras创建DNN来进行模型训练，训练完模型后对测试集进行验证，并保存该模型文件，便于后续的模型预测使用。模型训练的脚本为model_train.py，完整的Python代码如下：

```
# -*- coding: utf-8 -*-
# author: Jclian91
# place: Pudong Shanghai
# time: 2020-02-12 13:37
```

```
import os
# 是否使用GPU训练
# os.environ["CUDA_VISIBLE_DEVICES"] = "4,5,6,7,8"

import numpy as np
from load_data import train_df, test_df
from keras.utils import to_categorical
from keras.models import Model
from keras.optimizers import Adam
from keras.layers import Input, BatchNormalization, Dense
from bert.extract_feature import BertVector

# 读取文件并进行转换
bert_model = BertVector(pooling_strategy="REDUCE_MEAN", max_seq_len=100)
print('begin encoding')
f = lambda text: bert_model.encode([text])["encodes"][0]
train_df['x'] = train_df['text'].apply(f)
test_df['x'] = test_df['text'].apply(f)
print('end encoding')

x_train = np.array([vec for vec in train_df['x']])
x_test = np.array([vec for vec in test_df['x']])
y_train = np.array([vec for vec in train_df['label']])
y_test = np.array([vec for vec in test_df['label']])
print('x_train: ', x_train.shape)

# Convert class vectors to binary class matrices.
num_classes = 2
y_train = to_categorical(y_train, num_classes)
y_test = to_categorical(y_test, num_classes)

# 创建DNN模型
x_in = Input(shape=(768, ))
x_out = Dense(32, activation="relu")(x_in)
x_out = BatchNormalization()(x_out)
x_out = Dense(num_classes, activation="softmax")(x_out)
model = Model(inputs=x_in, outputs=x_out)
print(model.summary())

model.compile(loss='categorical_crossentropy',
              optimizer=Adam(),
              metrics=['accuracy'])

# 模型训练、评估以及保存
model.fit(x_train, y_train, batch_size=8, epochs=20)
model.save('visit_classify.h5')
print(model.evaluate(x_test, y_test))
```

模型训练

在模型训练中，我们创建的DNN模型结构如下：

```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 768)               0
_________________________________________________________________
dense_1 (Dense)              (None, 32)                24608
_________________________________________________________________
batch_normalization_1 (Batch (None, 32)                128
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 66
=================================================================
Total params: 24,802
Trainable params: 24,738
Non-trainable params: 64
_________________________________________________________________
```

模型训练过程中的输出如下：

```
Epoch 1/20

  8/250 [..............................] - ETA: 43s - loss: 1.0427 - acc: 0.3750
250/250 [==============================] - 1s 6ms/step - loss: 0.3345 - acc: 0.8640
Epoch 2/20

  8/250 [..............................] - ETA: 0s - loss: 0.2664 - acc: 0.8750
250/250 [==============================] - 0s 133us/step - loss: 0.2147 - acc: 0.9320

.........(省略部分输出结果).............

Epoch 19/20

  8/250 [..............................] - ETA: 0s - loss: 0.2481 - acc: 0.8750
250/250 [==============================] - 0s 136us/step - loss: 0.0716 - acc: 0.9760
Epoch 20/20

  8/250 [..............................] - ETA: 0s - loss: 0.0149 - acc: 1.0000
250/250 [==============================] - 0s 140us/step - loss: 0.0560 - acc: 0.9800

32/50 [==================>...........] - ETA: 0s
50/50 [==============================] - 0s 4ms/step
[0.3687818288803101, 0.9199999928474426]
```

经过20个epoch的训练，模型在训练集上的准确率为0.9800，在测试集上的准确率约为0.9200，BERT的效果如此惊人，后接简单的DNN模型就能取得如此不错的效果。

模型预测

为了再次验证模型的预测效果，笔者从网站上又重新找了20个文档，对其进行预测。预测的脚本为model_predict.py，完整的Python代码如下：

```
# -*- coding: utf-8 -*-
# author: Jclian91
# place: Pudong Shanghai
# time: 2020-02-12 17:33

import pandas as pd
import numpy as np
from bert.extract_feature import BertVector
from keras.models import load_model
load_model = load_model("visit_classify.h5")

# 预测语句
texts = ['在访问限制中，用户可以选择禁用iPhone的功能，包括Siri、iTunes购买功能、安装/删除应用
         'IT之家4月23日消息 近日，谷歌在其官方论坛发布消息表示，他们为Android Auto添加了一项
         '要通过telnet 访问路由器，需要先通过console 口对路由器进行基本配置，例如：IP地址、密
         'IT之家3月26日消息 近日反盗版的国际咨询公司MUSO发布了2017年的年度报告，其中的数据显示
         '应葡萄牙议会邀请，全国人大常委会副委员长吉炳轩率团于12月14日至16日访问葡萄牙，会见副
         '2月26日至3月2日，应香港特区政府“内地贵宾访港计划”邀请，省委常委、常务副省长陈向群赴港
         '目前A站已经恢复了访问，可以直接登录，网页加载正常，视频已经可以正常播放。',
         '难民署特使安吉丽娜·朱莉6月8日结束了对哥伦比亚和委内瑞拉边境地区的难民营地为期两天的访
         '据《南德意志报》报道，德国总理默克尔计划明年1月就前往安卡拉，和土耳其总统埃尔多安进行
         '自9月14日至18日，由越共中央政治局委员、中央书记处书记、中央经济部部长阮文平率领工作代
         'Win7电脑提示无线适配器或访问点有问题怎么办?很多用户在使用无线网连接上网时，发现无线网
         '2019年10月13日至14日，外交部副部长马朝旭访问智利，会见智利外长里韦拉，同智利总统外事
         '未开发所有安全组之前访问，FTP可以链接上，但是打开会很慢，需要1-2分钟才能链接上',
         'win7系统电脑的用户，在连接WIFI网络网上时，有时候会遇到突然上不了网，查看连接的WIFI出
         '联合国秘书长潘基文 8 日访问了日本福岛县，与当地灾民交流并访问了一所高中。',
         '国务院总理温家宝当地时间23日下午乘专机抵达布宜诺斯艾利斯，开始对阿根廷进行正式访问。
         '正在中国访问的巴巴多斯总理斯图尔特 1 5 日在陕西西安参观访问。',
         '据外媒报道,当地时间10日,美国白宫发声明称,美国总统特朗普将于2月底访问印度,与印度总理
         '2月28日，唐山曹妃甸蓝色海洋科技有限公司董事长赵力军等一行5人到黄海水产研究所交流访问
         '2018年7月2日，莫斯科孔子文化促进会会长姜彦彬，常务副会长陈国建，在中国著名留俄油画大
         ]

labels = []

bert_model = BertVector(pooling_strategy="REDUCE_MEAN", max_seq_len=100)

# 对上述句子进行预测
for text in texts:

    # 将句子转换成向量
    vec = bert_model.encode([text])["encodes"][0]
    x_train = np.array([vec])

    # 模型预测
    predicted = load_model.predict(x_train)
    y = np.argmax(predicted[0])
    label = 'Y' if y else 'N'
    labels.append(label)

for text,label in zip(texts, labels):
    print('%s\t%s'%(label, text))

# 将结果保存为xlsx文件
df = pd.DataFrame({'句子':texts, "是否属于出访类事件": labels})
df.to_excel('./result.xlsx', index=False)
```

模型预测的结果会输出，同时也会保存至Excel，文件的内容如下：

句子	是否属于出访类事件
在访问限制中，用户可以选择禁用iPhone的功能，包括Siri、iTunes购买功能、安装/删除应用等，甚至还可以让iPhone变成一台功能手机。以下是访问限制具体可以实现的一些功能	N
IT之家4月23日消息 近日，谷歌在其官方论坛发布消息表示，他们为Android Auto添加了一项新功能：可以访问完整联系人列表。用户现在可以通过在Auto的电话拨号界面中打开左上角的菜单访问完整的联系人列表。值得注意的是，这一功能仅支持在车辆停止时使用。	N
要通过telnet 访问路由器，需要先通过console 口对路由器进行基本配置，例如：IP地址、密码等。	N
IT之家3月26日消息 近日反盗版的国际咨询公司MUSO发布了2017年的年度报告，其中的数据显示，去年盗版资源网站访问量达到了3000亿次，比前一年（2016年）提高了1.6%。美国是访问盗版站点次数最多的国家，共有279亿次访问；其后分别是俄罗斯、印度和巴西，中国位列第18。	N
应葡萄牙议会邀请，全国人大常委会副委员长吉炳轩率团于12月14日至16日访问葡萄牙，会见副议长费利佩、社会党副总书记卡内罗。	Y
2月26日至3月2日，应香港特区政府“内地贵宾访港计划”邀请，省委常委、常务副省长陈向群赴港考察访问，重点围绕“香港所长、湖南所需”，与特区政府相关部门和机构深入交流，推动湖南与香港交流合作取得新进展。	Y
目前A站已经恢复了访问，可以直接登录，网页加载正常，视频已经可以正常播放。	N
难民署特使安吉丽娜·朱莉6月8日结束了对哥伦比亚和委内瑞拉边境地区的难民营地为期两天的访问，她对哥伦比亚人民展现的人道主义和勇气表示赞扬。	Y
据《南德意志报》报道，德国总理默克尔计划明年1月就前往安卡拉，和土耳其总统埃尔多安进行会谈。	Y
自9月14日至18日，由越共中央政治局委员、中央书记处书记、中央经济部部长阮文平率领工作代表团对希腊进行工作访问。	Y
Win7电脑提示无线适配器或访问点有问题怎么办?很多用户在使用无线网连接上网时，发现无线网显示已连接，但旁边却出现了一个黄色感叹号，无法进行网络操作，通过诊断提示电脑无线适配器或访问点有问题，且处于未修复状态，这该怎么办呢?下面小编就和大家分享下Win7电脑提示无线适配器或访问点有问题的解决方法。	N
2019年10月13日至14日，外交部副部长马朝旭访问智利，会见智利外长里韦拉，同智利总统外事顾问萨拉斯举行会谈，就智利举办亚太经合组织（APEC）第二十七次领导人非正式会议等深入交换意见。	Y
未开发所有安全组之前访问，FTP可以链接上，但是打开会很慢，需要1-2分钟才能链接上	N
win7系统电脑的用户，在连接WIFI网络网上时，有时候会遇到突然上不了网，查看连接的WIFI出现“有限的访问权限”的文字提示。	N
联合国秘书长潘基文 8 日访问了日本福岛县，与当地灾民交流并访问了一所高中。	Y
国务院总理温家宝当地时间23日下午乘专机抵达布宜诺斯艾利斯，开始对阿根廷进行正式访问。	Y
正在中国访问的巴巴多斯总理斯图尔特 1 5 日在陕西西安参观访问。	Y
据外媒报道,当地时间10日,美国白宫发声明称,美国总统特朗普将于2月底访问印度,与印度总理莫迪进行战略对话。	Y
2月28日，唐山曹妃甸蓝色海洋科技有限公司董事长赵力军等一行5人到黄海水产研究所交流访问。黄海水产研究所副所长辛福言及相关部门负责人、专家等参加了会议。	Y
2018年7月2日，莫斯科孔子文化促进会会长姜彦彬，常务副会长陈国建，在中国著名留俄油画大师牟克教授的陪同下，访问了莫斯科国立苏里科夫美术学院，受到第一副校长伊戈尔·戈尔巴秋克先生接待。	Y

excel文件中的内容

所有预测的文档完全正确！

预测

本项目已开源，Github地址为：https://github.com/percent4/bert_doc_binary_classification 。

通过笔者自己的试验，BERT在小标注样本量的效果确实很不错，后续我们还将继续接触BERT！

感谢大家的阅读~