

Kaggle: Quora Insincere Questions Classification

文本分类

主讲人：林逸飞

比赛特点



1. 比赛的数据为非脱敏数据，大家可以直观地看到自己在分析什么东西，并进行误差分析
2. 训练数据量足够，大家想用什么模型都可以
3. 竞赛结果为kernel提交，方便没有GPU环境或机器配置不足的同学
4. 比赛配了4个词向量供大家选择，NLP训练营的同学可以学以致用
5. 我试过机器学习和CNN的模型也可以跑出不错的效果，各个训练营的小伙伴们都可以拿自己学的东西试试
6. Kaggle是全球top的数据科学比赛平台，在这里你会和世界各地的数据科学家们交流PK，如果说DC的达观杯是市级赛的话，Kaggle无疑是个世界级的锦标赛。

比赛介绍



An existential problem for any major website today is how to handle toxic and divisive content. Quora wants to tackle this problem head-on to keep their platform a place where users can feel safe sharing their knowledge with the world.

Quora is a platform that empowers people to learn from each other. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. A key challenge is to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

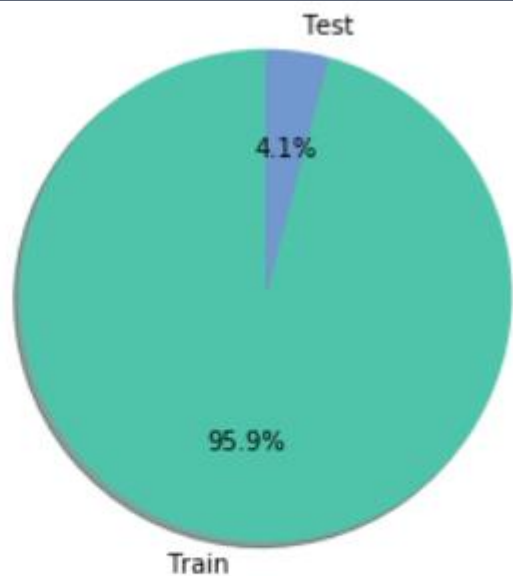


In this competition, Kagglers will develop models that identify and flag insincere questions. To date, Quora has employed both machine learning and manual review to address this problem. With your help, they can develop more scalable methods to detect toxic and misleading content.

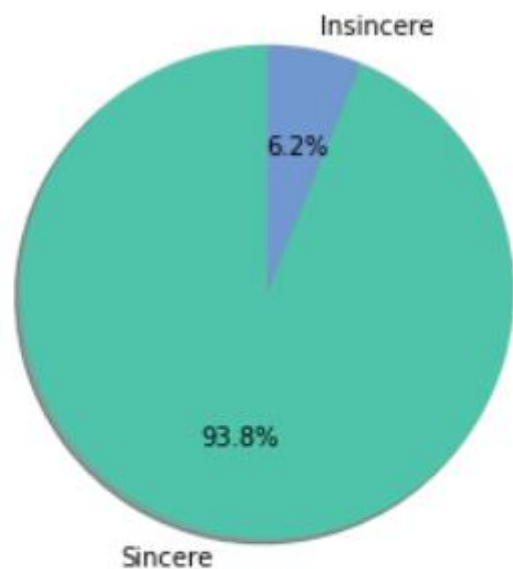
Here's your chance to combat online trolls at scale. Help Quora uphold their policy of "Be Nice, Be Respectful" and continue to be a place for sharing and growing the world's knowledge.

简单来说，就是一个文本二分类任务，识别哪些是正常的问题（0），哪些是不坏好意的问题（1）。

数据介绍



训练数据:1306122条
测试数据:56370条
训练占比:95.86%



负样本数:1225312条
正样本数:80810条
正样本占:6.19%

数据介绍



正常的问题： 确实想知道问题的答案

```
array(['How did Quebec nationalists see their province as a nation in the 1960s?',  
      'Do you have an adopted dog, how would you encourage people to adopt and not shop?',  
      'Why does velocity affect time? Does velocity affect space geometry?',  
      'How did Otto von Guericke used the Magdeburg hemispheres?',  
      'Can I convert montra helicon D to a mountain bike by just changing the tyres?'],  
      dtype=object)
```

不怀好意的问题： 假装提问但其实表达的是自己的恶意观点

```
array(['Has the United States become the largest dictatorship in the world?',  
      'Which babies are more sweeter to their parents? Dark skin babies or light skin babies?',  
      "If blacks support school choice and mandatory sentencing for criminals why don't they vote Republican?",  
      'I am gay boy and I love my cousin (boy). He is sexy, but I dont know what to do. He is hot, and I want to see his di**. What should I do?',  
      'Which races have the smallest penis?'], dtype=object)
```


官网注册



<https://www.kaggle.com/competitions>

或

<https://www.kaggle.com/c/quora-insincere-questions-classification>



Quora Insincere Questions Classification


Detect toxic content to improve online conversations

Featured · a month to go · text data, binary classification

\$25,000

2,932 teams

加入比赛；（我已经报名了所以显示的不一样，报名在蓝色位置）


 Quora · 2,932 teams · a month to go (a month to go until merger deadline)


[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)


Overview

×

Sign in or register with one click:

 Sign in with Facebook

 Sign in with Google

 Sign in with Yahoo

or

Use your Kaggle username or email:
[Register with email »](#)

Sign in

☐ Remember me [Forgot Username / Password](#)

One account per individual.
e.g., If you're joining as a company, please create one account for each participant.

参赛下载数据



切到Data导航栏，点击Download All，将数据都放到一个input文件夹中

A screenshot of the DeepShare competition page. The 'Data' tab is selected in the top navigation bar. The page displays a 'Data Description' section with text about the dataset's content and a 'File descriptions' section listing files like train.csv, test.csv, and sample_submission.csv. At the bottom, there is a 'Data (6 GB)' section with a 'Download All' button circled in red. Below this, there are tabs for 'Data Sources', 'About this file', and 'Columns'.

Overview **Data** Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Data Description

- Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

The training data includes the question that was asked, and whether it was identified as insincere (`target = 1`). The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

Note that the distribution of questions in the dataset should not be taken to be representative of the distribution of questions asked on Quora. This is, in part, because of the combination of sampling procedures and sanitization measures that have been applied to the final dataset.

File descriptions

- train.csv - the training set
- test.csv - the test set
- sample_submission.csv - A sample submission in the correct format
- embeddings/ - (see below)

Data fields

- aid - unique question identifier

Data (6 GB) API kaggle competitions download -c quora... ? **Download All** ✕

| Data Sources | About this file | Columns |
|--------------------------------|--|---------|
| sample_submission... 56.4k x 2 | Sample Submission file, in the correct | aid |

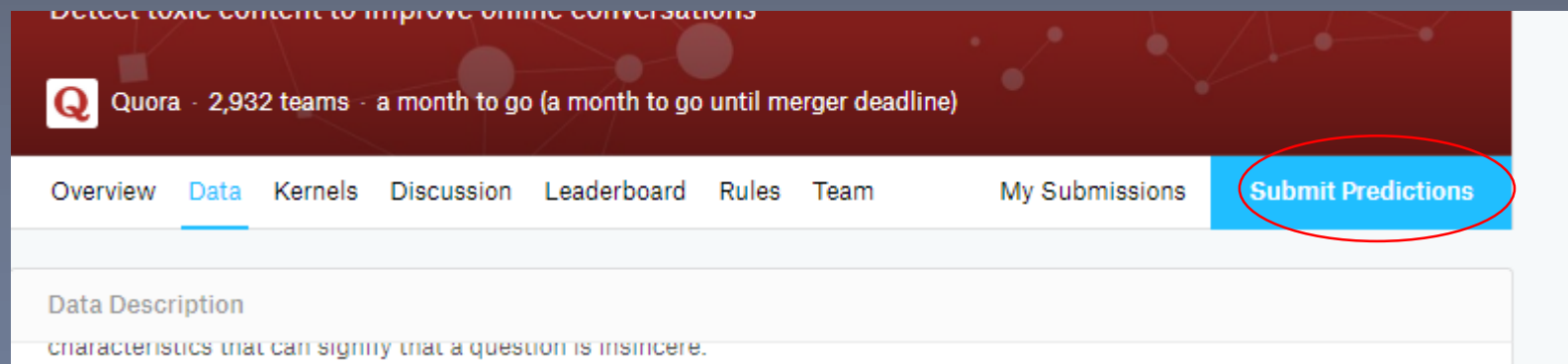
下载的数据集中，embedding为300维的词向量文件。CS224N训练营的同学应该很熟悉，其它训练营的小伙伴们可以不用在意，将它当作特征工程的一个辅助工具就好。

提交介绍

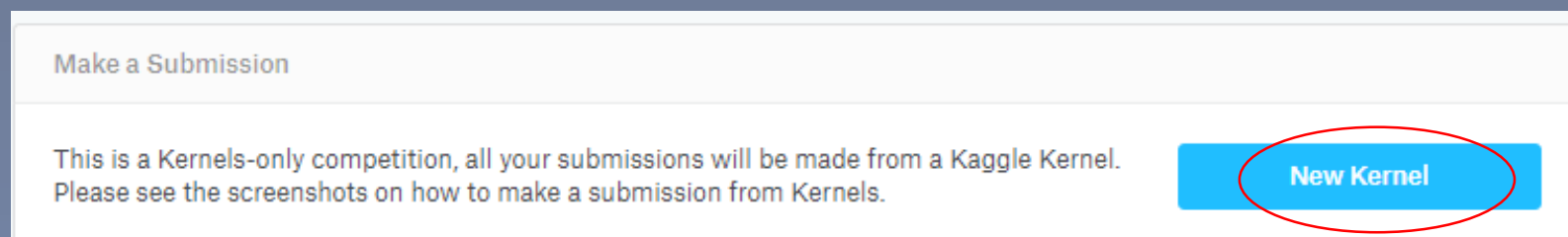


这个比赛是一个kernel only的比赛，和大家以前参加的达官杯比赛不一样。我们并不是提交一个预测好的csv文件，而是提交一个完整的运行代码到kaggle的云服务器上面，云服务器上面会自动运行。具体操作如下：

点击“submit predictions”



点击“New Kernel”



提交介绍



点击“Script” 提交建议选择Script模式

Select Kernel Type

Script

```
import numpy as np # linear algebra
import pandas as pd # data processing,
# Input data files are available in the
from subprocess import check_output
print(check_output(["ls", "../input"]))
# Any results you write to the current
```

- Python, R, RMarkdown
- Runs all the code, every time
- Ideal for fitting a model and competition submissions
- Shares code for review and RMarkdown reports

Notebook

Introduction

```
# Loading in the training data
train = pd.read_csv("../train.csv")
```

- Jupyter Notebooks in Python or R
- Runs cells of code and Markdown
- Ideal for interactive data exploration and polished analysis
- Shares insights through code & commentary

在右边将GPU设置为ON

Settings

Sharing

Private, 0 collaborators

Language

Python

Docker

Latest available

GPU BETA

GPU on

Internet BETA

Internet blocked

Custom packages are not supported for GPU instances

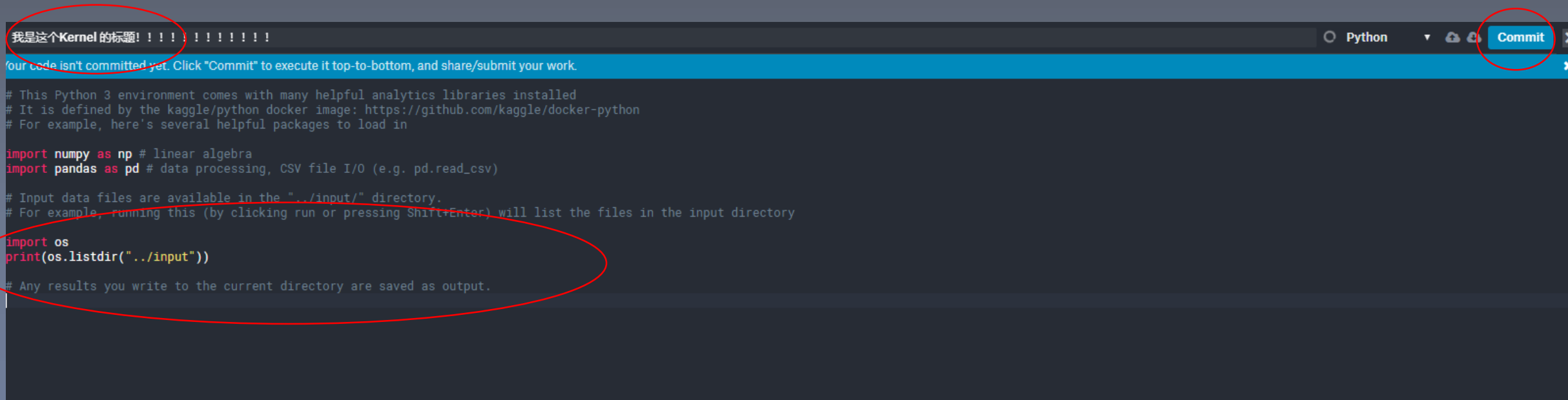
Docs

> API

提交介绍

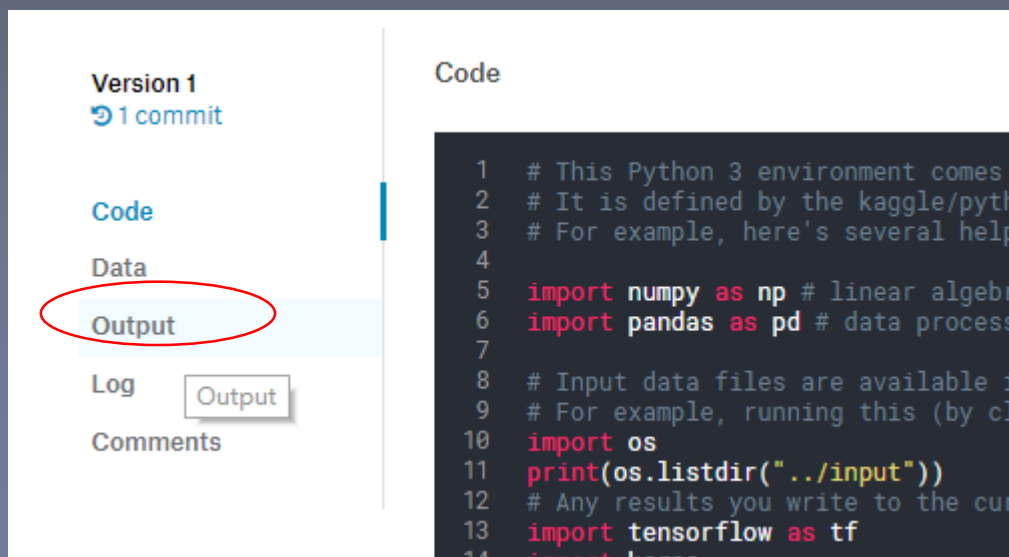


设定一个标题，黏贴进我们提供的baseline模型，点击commit。运行完成之后，点击open version



提交介绍

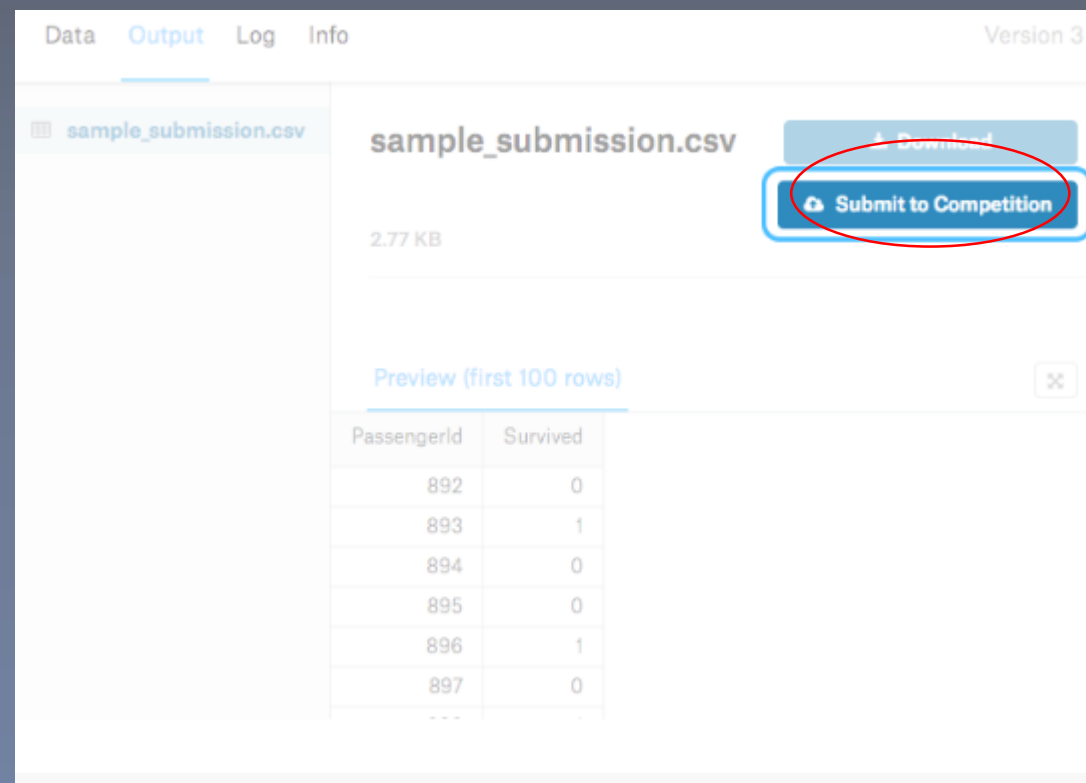
点击导航窗格的“Output”



The screenshot shows the Kaggle interface with the 'Output' tab selected in the left sidebar. The sidebar also includes 'Version 1', '1 commit', 'Code', 'Data', 'Log', and 'Comments'. The 'Output' tab is highlighted with a red circle. The main area displays a code editor with Python code for data processing and machine learning.

```
1 # This Python 3 environment comes
2 # It is defined by the kaggle/pyth
3 # For example, here's several help
4
5 import numpy as np # linear algebr
6 import pandas as pd # data process
7
8 # Input data files are available i
9 # For example, running this (by cl
10 import os
11 print(os.listdir("../input"))
12 # Any results you write to the cur
13 import tensorflow as tf
14 import keras
```

最后在点击一下“submit to competition”就好了!



The screenshot shows the Kaggle interface with the 'sample_submission.csv' file selected. The file size is 2.77 KB. The 'Submit to Competition' button is highlighted with a red circle. Below the file information, there is a preview of the first 100 rows of the CSV file.

| PassengerId | Survived |
|-------------|----------|
| 892 | 0 |
| 893 | 1 |
| 894 | 0 |
| 895 | 0 |
| 896 | 1 |
| 897 | 0 |
| ... | ... |

提分策略



BaseLine模型是Facebook在2016推出的FastText模型，其优点是简单计算快速，方便大家跑通模型。但是效果却不见得能很好，大家在跑通模型之后可以尝试做一下几件事情：

1. 在本地划分数据集为训练和开发集，不必每次都在kernel上面运行然后提交；
2. 尝试不同的模型，比如SVM\RF\LSTM\GRU\CNN等等；
3. 初学者参加比赛不要求以分数为主，而是多尝试学习几个模型，你用的东西必须要学会它；
4. 在初期不要使用模型融合，而应该调整单模型，对它进行偏差、方差和误差的分析，进步更快；
5. 先保证你的训练集F1得分在0.75甚至0.85以上，再去想办法加强它的泛化能力；（课本知识没学扎实不急参加小测）
6. 多在讨论区里面和世界各地的小伙伴交流学习，尝试复现论文应用于比赛；
7. 看看你的模型都将哪些句子预测错了（非脱敏数据的好处）；

Baseline 模型

1-45

```
1 # This Python 3 environment comes with many helpful analytics libraries installed
2 # It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
3 # For example, here's several helpful packages to load in
4
5 import numpy as np # linear algebra
6 import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
7
8 # Input data files are available in the "../input/" directory.
9 # For example, running this (by clicking run or pressing Shift+Enter) will list the files in the input directory
10 import os
11 print(os.listdir("../input"))
12 # Any results you write to the current directory are saved as output.
13 import tensorflow as tf
14 import keras
15 from keras import Model
16 from keras.preprocessing.text import Tokenizer
17 from keras.preprocessing.sequence import pad_sequences
18 K = keras.backend
19 L = keras.layers
20
21 maxlen = 72 #一个句子补全或保留至72个单词就好
22
23 def load_and_prec():
24     """
25     读取数据集和词表
26     """
27     df_train = pd.read_csv("../input/train.csv")
28     df_test = pd.read_csv("../input/test.csv")
29     ## fill up the missing values
30     train_X = df_train["question_text"].fillna("#_#").values
31     test_X = df_test["question_text"].fillna("#_#").values
32     ## Tokenize the sentences
33     tokenizer = Tokenizer()
34     tokenizer.fit_on_texts(list(train_X))
35     train_X = tokenizer.texts_to_sequences(train_X)
36     test_X = tokenizer.texts_to_sequences(test_X)
37     ## Pad the sentences
38     x_train = pad_sequences(train_X, maxlen=maxlen)
39     x_test = pad_sequences(test_X, maxlen=maxlen)
40
41     ## Get the target values
42     y_train = df_train['target'].values
43
44     return x_train, x_test, y_train, tokenizer.word_index
45
```

46-84



deepshare.net

深享网

```
46
47
48 def load_glove(word_index, max_features):
49     """
50     读取词向量
51     """
52     EMBEDDING_FILE = '../input/embeddings/glove.840B.300d/glove.840B.300d.txt'
53     def get_coefs(word, *arr):
54         return word, np.asarray(arr, dtype='float32')
55     embeddings_index = dict(get_coefs(*o.split(" ")) for o in open(EMBEDDING_FILE, encoding='utf8'))
56
57     all_embs = np.stack(embeddings_index.values())
58     emb_mean, emb_std = all_embs.mean(), all_embs.std()
59     embed_size = all_embs.shape[1]
60
61     embedding_matrix = np.random.normal(emb_mean, emb_std, (max_features, embed_size))
62     for word, i in word_index.items():
63         if i >= max_features:
64             continue
65         embedding_vector = embeddings_index.get(word)
66         if embedding_vector is not None:
67             embedding_matrix[i] = embedding_vector
68     return embedding_matrix
69
70
71 def FastText(embedding_matrix):
72     """
73     定义FastText模型
74     """
75     ipt = L.Input(shape=(maxlen,))
76     x = L.Embedding(input_dim=embedding_matrix.shape[0],
77                    output_dim=embedding_matrix.shape[1],
78                    weights=[embedding_matrix],
79                    trainable=False)(ipt)
80     x = L.GlobalAveragePooling1D()(x)
81     out = L.Dense(1, activation='sigmoid')(x)
82     model = Model(inputs=ipt, outputs=out)
83
84     return model
```


Baseline 模型

85-104

```
85
86 x_train, x_test, y_train, word_index = load_and_prec()
87 max_features = len(word_index)+1
88 embedding_matrix = load_glove(word_index, max_features)
89
90 model = FastText(embedding_matrix)
91 model.compile(
92     loss = 'binary_crossentropy',
93     optimizer = 'adam',
94 )
95 hist = model.fit(x_train, y_train,
96                 batch_size = 256,
97                 epochs = 1,
98                 verbose = 1,)
99
100 #####提交#####
101 y_pred = model.predict(x_test)
102 sub = pd.read_csv('../input/sample_submission.csv')
103 sub.prediction = y_pred > 0.2
104 sub.to_csv("submission.csv", index=False)
```

眼过千遍
不如
手过一遍



deepshare.net

深享网

联系我们：

电话：18001992849

邮箱：service@deepshare.net

QQ：2677693114



公众号



客服微信

