Datawhale

NLP实践全流程基础演示

潘笃驿

目录

Datawhale

CONTENTS

01 赛题介绍

02 Baseline介绍

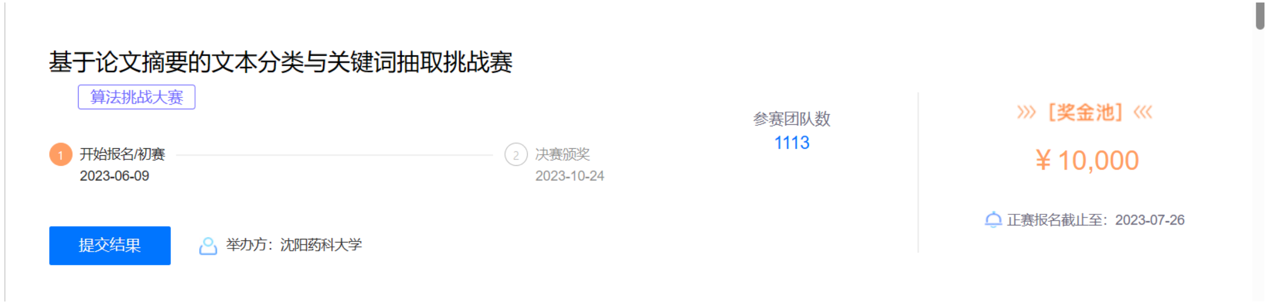
03 Q&A

Datawhale

/01 赛题介绍

赛题介绍

基于论文摘要的文本分类与关键词抽取挑战赛
2023 iFLYTEK A.I.开发者大赛-讯飞开放平台 (xfyun.cn)



基于论文摘要的文本分类与关键词抽取挑战赛

算法挑战大赛

1 开始报名/初赛 2023-06-09

2 决赛颁奖 2023-10-24

参赛团队数 1113

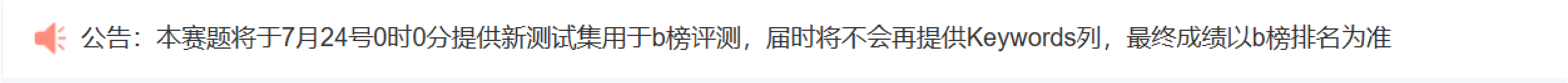
》》［奖金池］《《

￥10,000

正赛报名截止至：2023-07-26

提交结果

举办方：沈阳药科大学



公告：本赛题将于7月24号0时0分提供新测试集用于b榜评测，届时将不会再提供Keywords列，最终成绩以b榜排名为准

- 参与国内顶级竞赛，最佳AI实践机会，**科大讯飞** 与 **阿里云天池** 联合举办
- 理解竞赛实践的**方法论和技巧梳理**，Get**开箱即用**的**Baseline**
- 逐层深入，由浅入深，掌握**底层算法原理**与**最佳实践技巧**

赛题任务解读

一、赛事背景

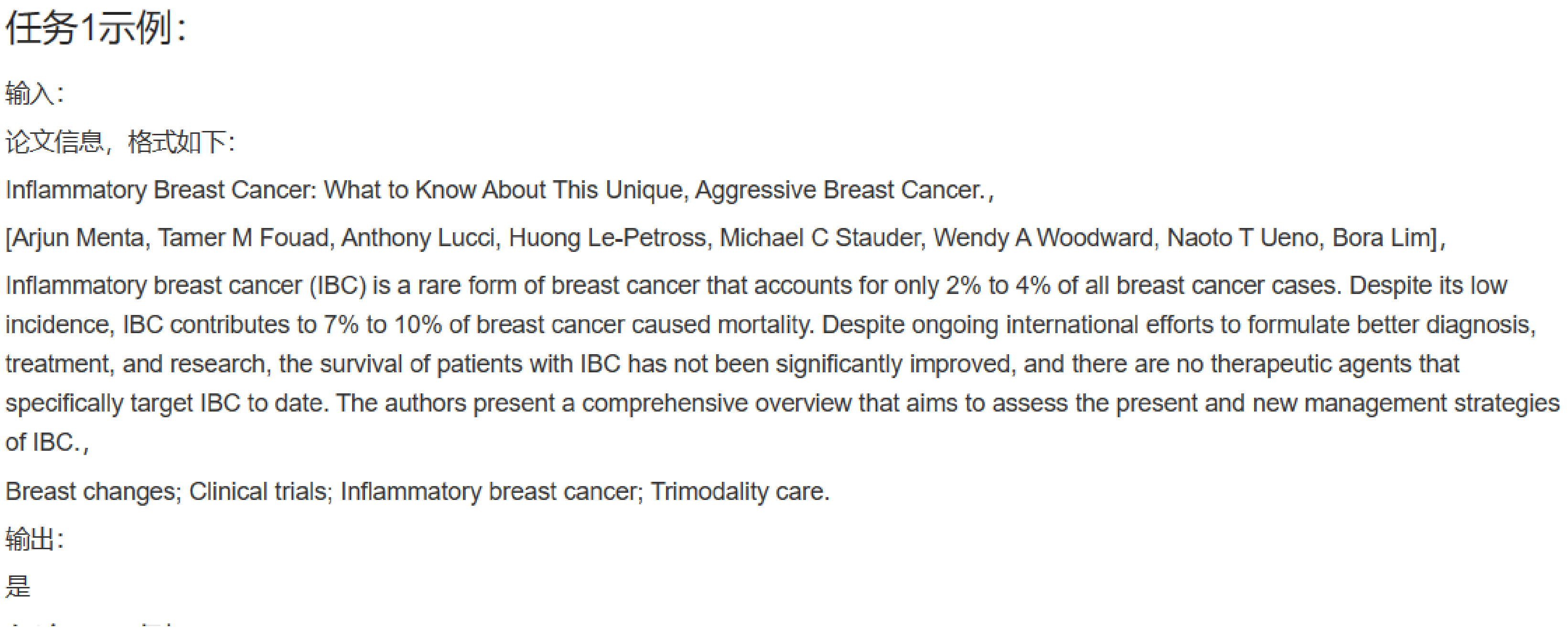
医学领域的文献库中蕴含了丰富的疾病诊断和治疗信息，如何高效地从海量文献中提取关键信息，进行疾病诊断和治疗推荐，对于临床医生和研究人员具有重要意义。

二、赛事任务

本任务分为两个子任务：

1. 机器通过对论文摘要等信息的理解，判断该论文是否属于医学领域的文献。
2. 提取出该论文关键词。

赛题任务解读



任务1示例：

输入：

论文信息，格式如下：

Inflammatory Breast Cancer: What to Know About This Unique, Aggressive Breast Cancer.,

[Arjun Menta, Tamer M Fouad, Anthony Lucci, Huong Le-Petross, Michael C Stauder, Wendy A Woodward, Naoto T Ueno, Bora Lim],

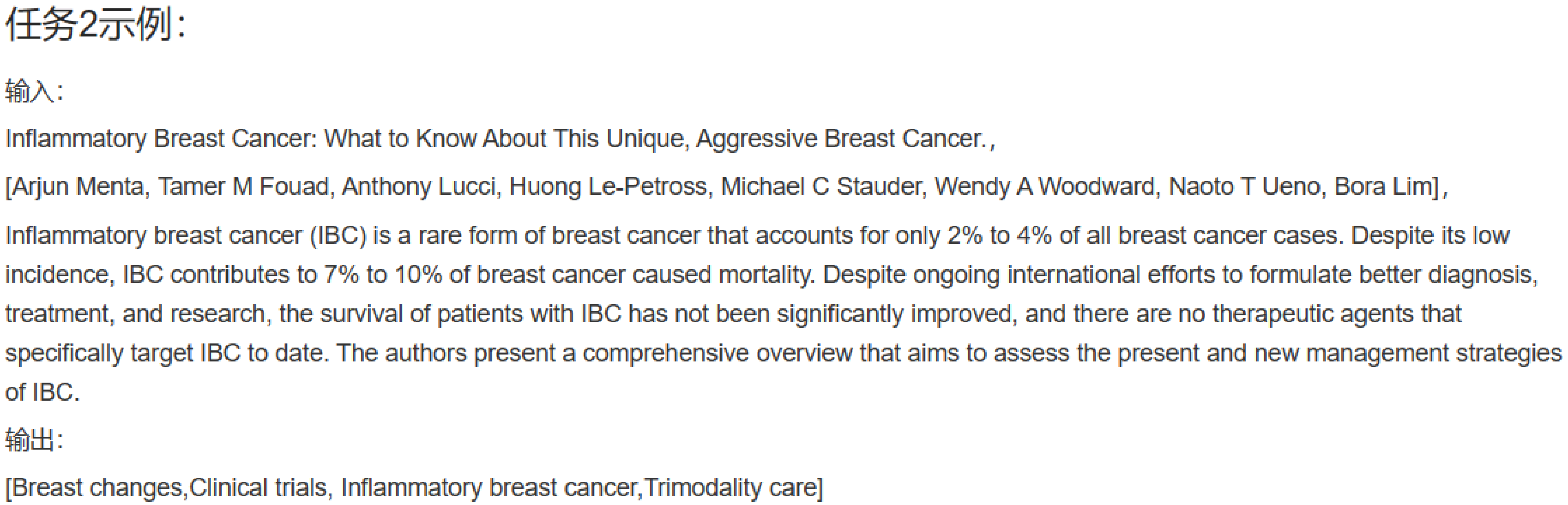
Inflammatory breast cancer (IBC) is a rare form of breast cancer that accounts for only 2% to 4% of all breast cancer cases. Despite its low incidence, IBC contributes to 7% to 10% of breast cancer caused mortality. Despite ongoing international efforts to formulate better diagnosis, treatment, and research, the survival of patients with IBC has not been significantly improved, and there are no therapeutic agents that specifically target IBC to date. The authors present a comprehensive overview that aims to assess the present and new management strategies of IBC.,

Breast changes; Clinical trials; Inflammatory breast cancer; Trimodality care.

输出：

是

赛题任务解读



任务2示例：

输入：

Inflammatory Breast Cancer: What to Know About This Unique, Aggressive Breast Cancer.,

[Arjun Menta, Tamer M Fouad, Anthony Lucci, Huong Le-Petross, Michael C Stauder, Wendy A Woodward, Naoto T Ueno, Bora Lim],

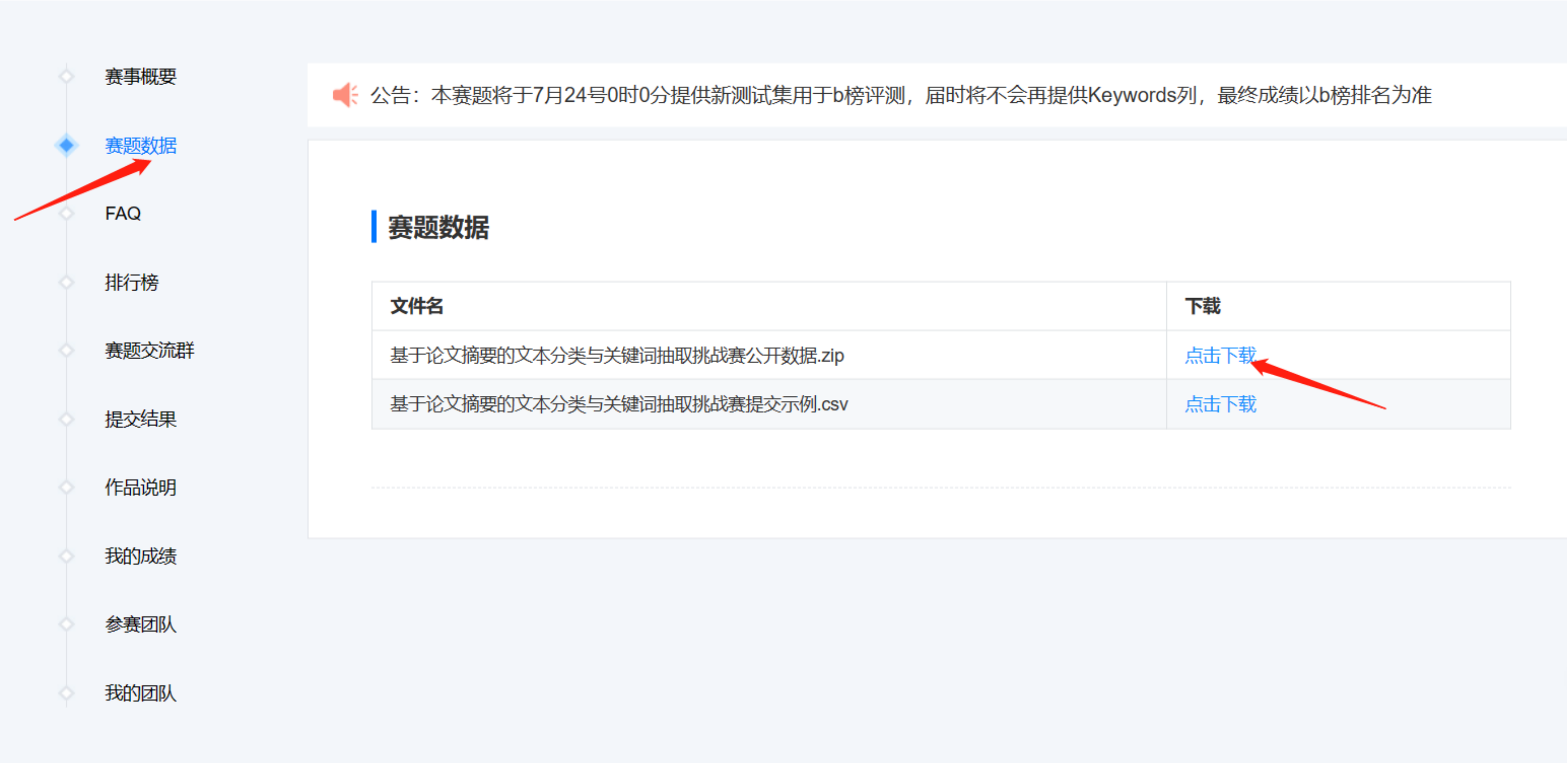
Inflammatory breast cancer (IBC) is a rare form of breast cancer that accounts for only 2% to 4% of all breast cancer cases. Despite its low incidence, IBC contributes to 7% to 10% of breast cancer caused mortality. Despite ongoing international efforts to formulate better diagnosis, treatment, and research, the survival of patients with IBC has not been significantly improved, and there are no therapeutic agents that specifically target IBC to date. The authors present a comprehensive overview that aims to assess the present and new management strategies of IBC.

输出：

[Breast changes,Clinical trials, Inflammatory breast cancer,Trimodality care]

赛题数据

Datawhale

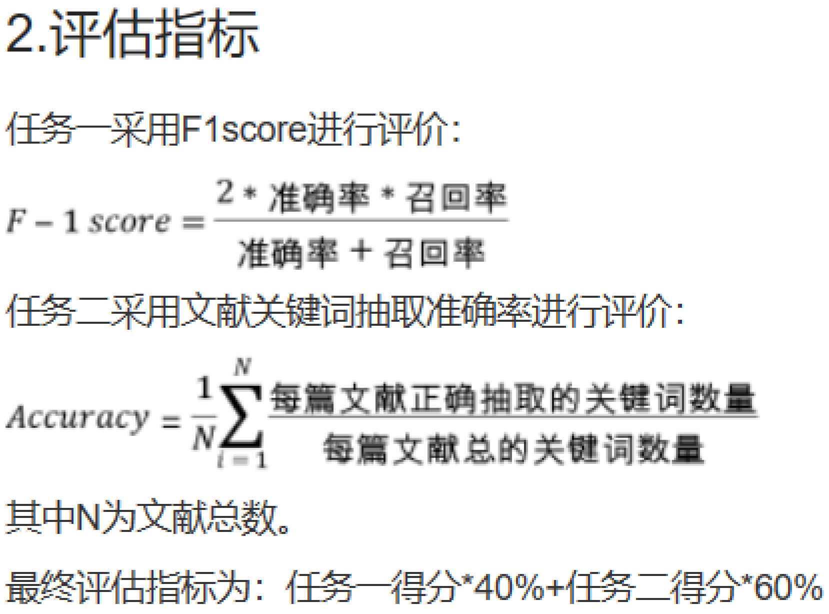


公告：本赛题将于7月24号0时0分提供新测试集用于b榜评测，届时将不会再提供Keywords列，最终成绩以b榜排名为准

赛题数据

文件名	下载
基于论文摘要的文本分类与关键词抽取挑战赛公开数据.zip	点击下载
基于论文摘要的文本分类与关键词抽取挑战赛提交示例.csv	点击下载

评价指标



2.评估指标

任务一采用F1score进行评价:

$$F-1\ score=\frac{2*准确率*召回率}{准确率+召回率}$$

任务二采用文献关键词抽取准确率进行评价:

$$Accuracy=\frac{1}{N}\sum_{i=1}^{N}\frac{每篇文献正确抽取的关键词数量}{每篇文献总的关键词数量}$$

其中N为文献总数。

最终评估指标为：任务一得分*40%+任务二得分*60%

/02 Baseline介绍

Baseline介绍

本次赛事baseline提供三个版本

1.利用传统的特征提取方法（如TF-IDF/BOW）结合机器学习模型求解任务
2.使用预训练的BERT模型进行建模求解任务
3.微调ChatGLM2-6b模型求解任务

在线版本Baseline体验

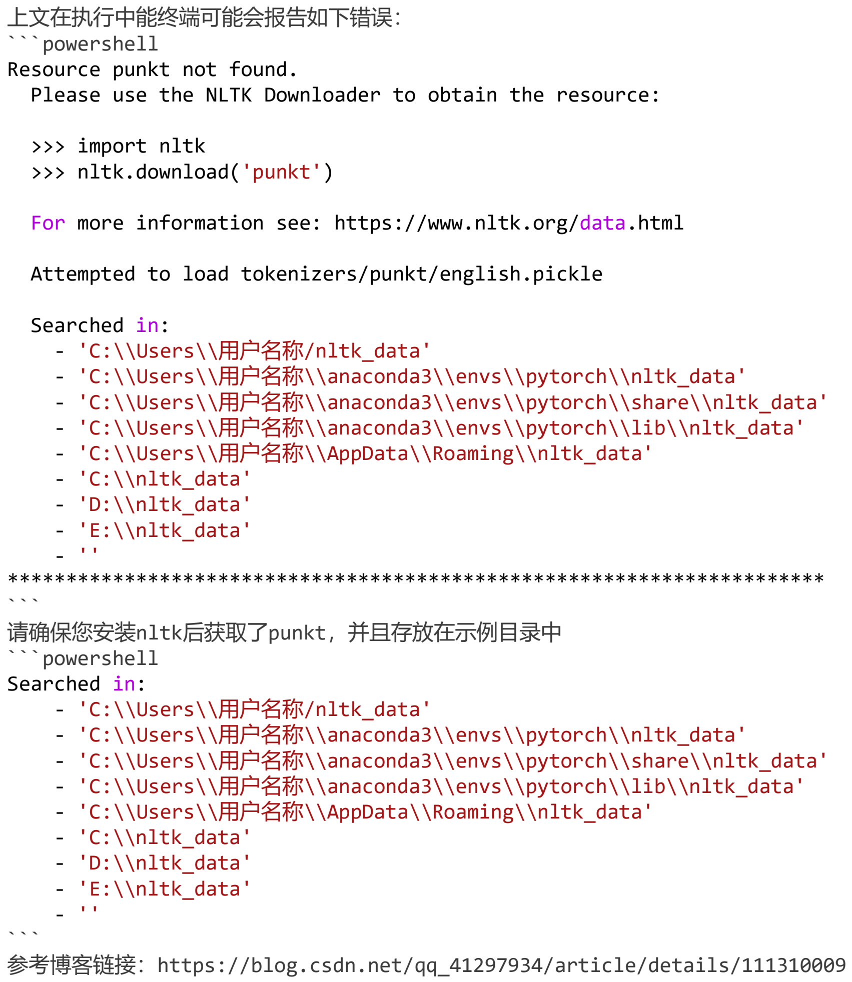
在线体验地址：手把手打一场NLP赛事 - 飞桨AI Studio (baidu.com)

本地Baseline体验

Datawhale

Baseline代码地址：AI夏令营 - NLP实践教程 - 飞书云文档 (feishu.cn)
将代码部分复制到本地后运行

本地Baseline体验需要注意的问题



上文在执行中能终端可能会报告如下错误：

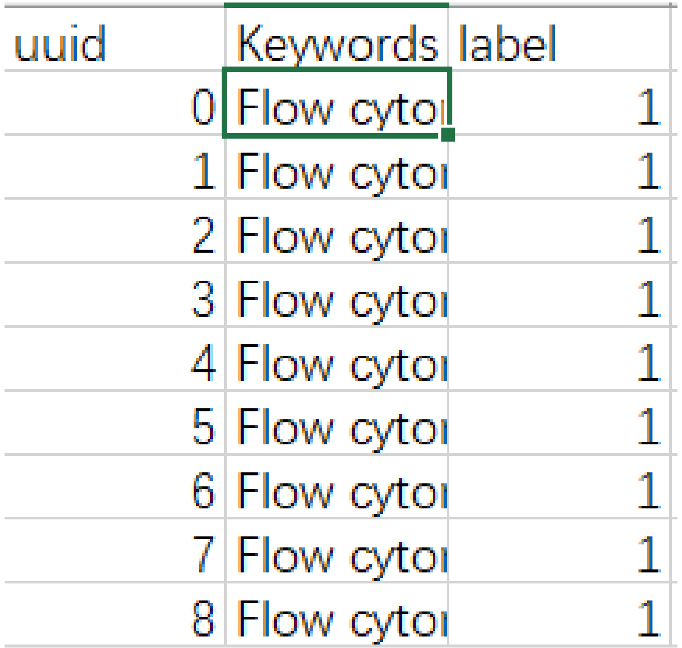
~~~
```powershell
Resource punkt not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('punkt')

  For more information see: https://www.nltk.org/data.html

  Attempted to load tokenizers/punkt/english.pickle

  Searched in:
    - 'C:\\Users\\用户名称/nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\share\\nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\lib\\nltk_data'
    - 'C:\\Users\\用户名称\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
    - ''
**********************************************************************
```
~~~

请确保您安装nltk后获取了punkt，并且存放在示例目录中

~~~
```powershell
Searched in:
    - 'C:\\Users\\用户名称/nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\share\\nltk_data'
    - 'C:\\Users\\用户名称\\anaconda3\\envs\\pytorch\\lib\\nltk_data'
    - 'C:\\Users\\用户名称\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
    - ''
```
~~~

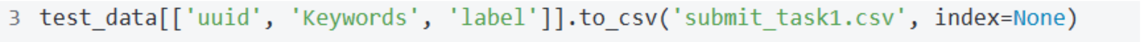
参考博客链接：https://blog.csdn.net/qq_41297934/article/details/111310009

结果提交



uuid	Keywords	label
0	Flow cyto	1
1	Flow cyto	1
2	Flow cyto	1
3	Flow cyto	1
4	Flow cyto	1
5	Flow cyto	1
6	Flow cyto	1
7	Flow cyto	1
8	Flow cyto	1

最后提交结果要求包含uuid，keywords 以及label列



```
test_data[['uuid', 'Keywords', 'label']].to_csv('submit_task1.csv', index=None)
```

Datawhale

/03 Q&A

Thank you