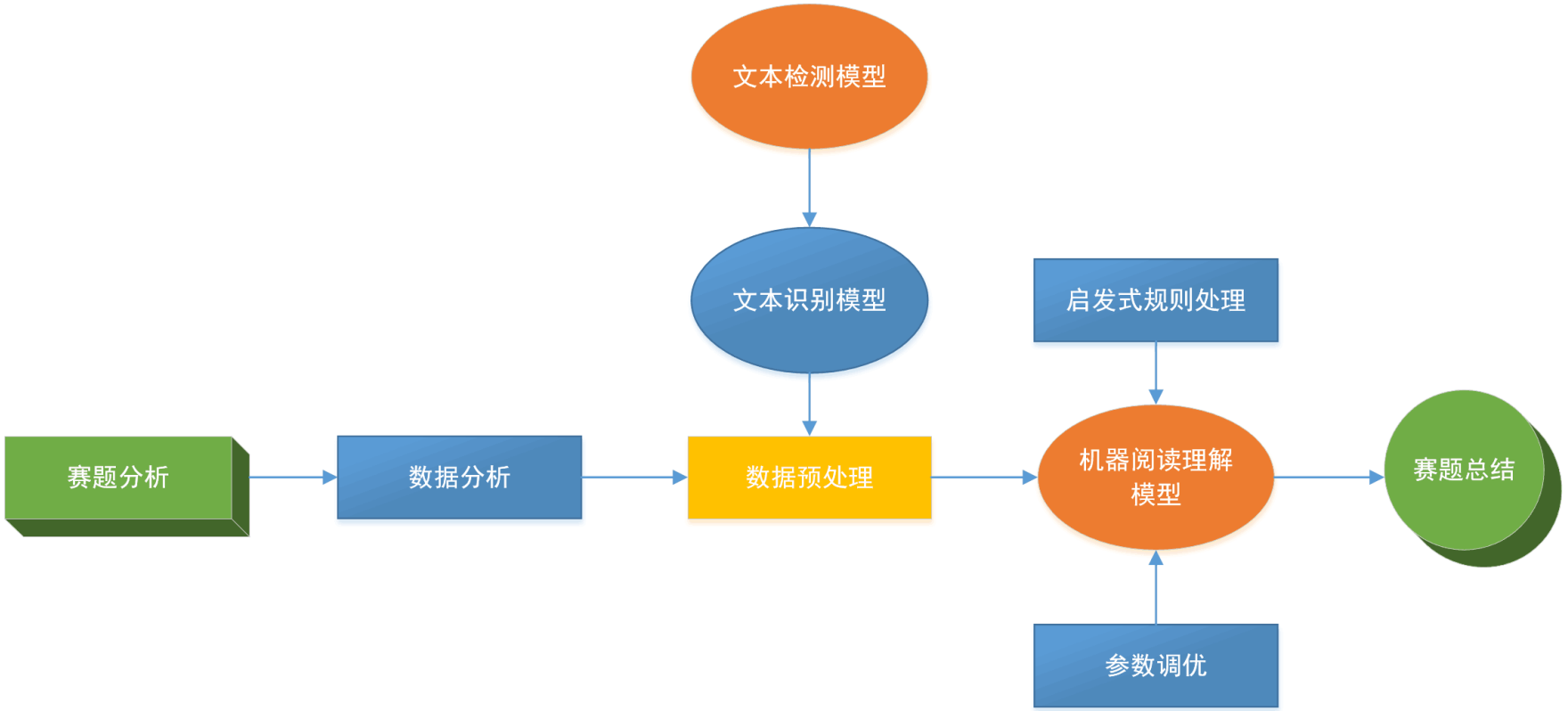
AIWIN

保险文本视觉认知问答

BUSINESS REPOR

顾立辉|陈震辉

系 统 整 体 框 架



目录

CONTENTS

01 赛题分析

PART ONE

赛　题　分　析

赛题要求

- ✓ 以AI为抓手，分析各种类型的文档的数据
- ✓ 对问题和回答进行建模
- ✓ 保证阅读理解一定的正确率，能对一些问题做出解答

赛题难点

- 照片拍摄角度不同，字体混合手写，一张图片可能由多张票据混合，背景噪声影响OCR识别。
- 模型构建和优化

数据

- 文档、说明书、票据、报告等，以图片的形式给出

目标

- 通过AI智能判断所识别文字的内在逻辑，回答关于图片的自然语言问题。

评估指标

- $\frac{1}{N}\sum_{i=0}^{N}\max_{j}\left(s(a_{ij}, o_i)\right)$

数　据　分　析

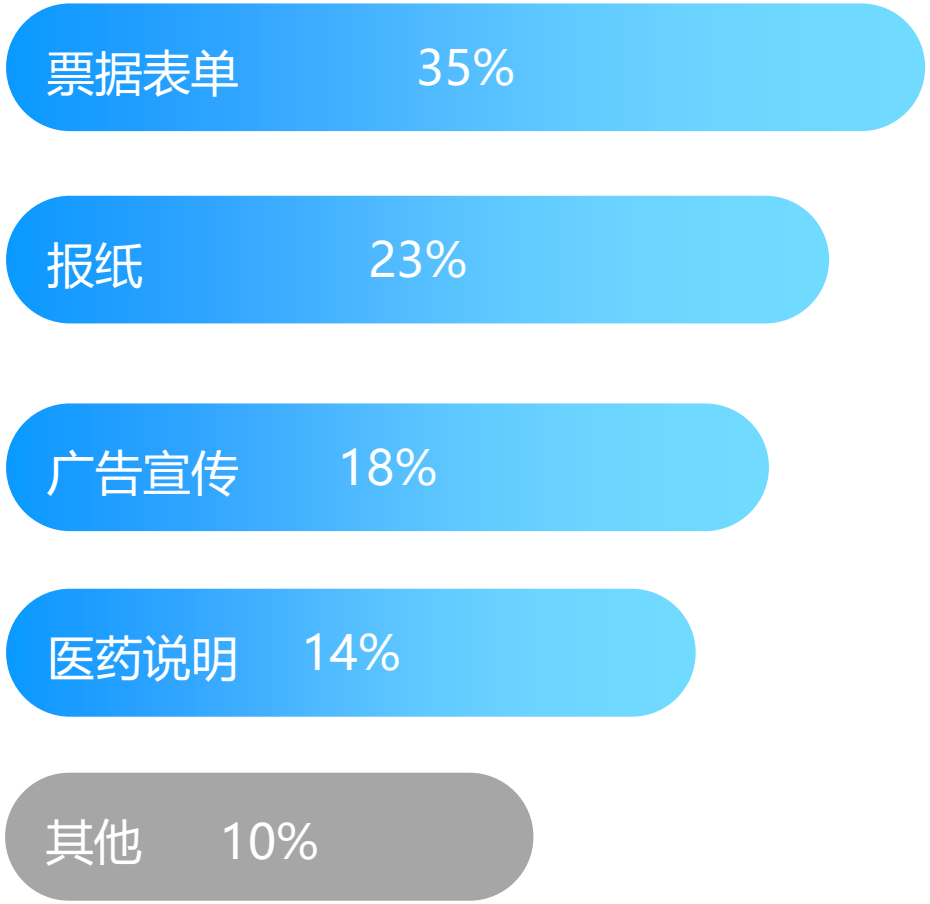
认识数据

index	question_ic	filename	question_text	answer_text
1	Q00001	c850b0d7	这是什么药品?	茶碱缓释片
2	Q00002	c850b0d7	本说明书来源于哪里?	黑龙江鼎恒升药业有限公司
3	Q00003	c850b0d7	本品可通过什么屏障?	胎盘
4	Q00004	c850b0d7	说明书上方正中是什么字?	茶碱缓释片
5	Q00005	c850b0d7	左上角是什么字?	说明书来源：黑龙江鼎恒升药业有限公司
6	Q00006	c850b0d7	老年用药是下一项是什么?	药物相互作用
7	Q00007	c850b0d7	Theophylline Sustainde-release Tablets是药品的什么?	英文名
8	Q00008	c850b0d7	茶碱是指什么?	主要成份
9	Q00009	c850b0d7	198.18是指什么数?	分子量

数据转化


```
"title" : string "AHEFGLB18921EAAA75R7_20210301111254_2.jpg"
"qas" : [ 9 items
  0 : { 3 items
    "question" : string "图7是表达什么的？"
    "id" : string "Q39453"
    "answers" : [ 1 item
      0 : { 2 items
        "text" : string "美国国债总额迅速增加（十亿美元）"
        "answer_start" : int 427
      }
    ]
  }
```

```
1 : { 3 items
  "question" : string "图8是说明什么的？"
  "id" : string "Q39454"
  "answers" : [ 1 item
    0 : { 2 items
      "text" : string "美元流动性危机解除"
      "answer_start" : int 446
    }
  ]
}
```

数　据　分　析

➢ 数据探索

		mean	min	25%	50%	75%	max
train	context	948	76	640	768	1258	3047
	question	24	7	16	20	26	35
	answer	6	2	5	6	12	34
test	context	894	92	640	768	1258	3047
	question	23	7	17	21	28	33

对日期、金额、电话、邮箱等格式后处理，规格化

6050	Q06050	c4fe8ac6d7	五百丁在五百丁金融大学上学是什么时候?	2015.09-2016.09
6051	Q06051	c4fe8ac6d7	五百丁在五百丁银行广州分行上班是什么时	2014.03-2015.08
6052	Q06052	c4fe8ac6d7	五百丁在五百丁银行银行上班是什么时候?	2013.11-2014.07

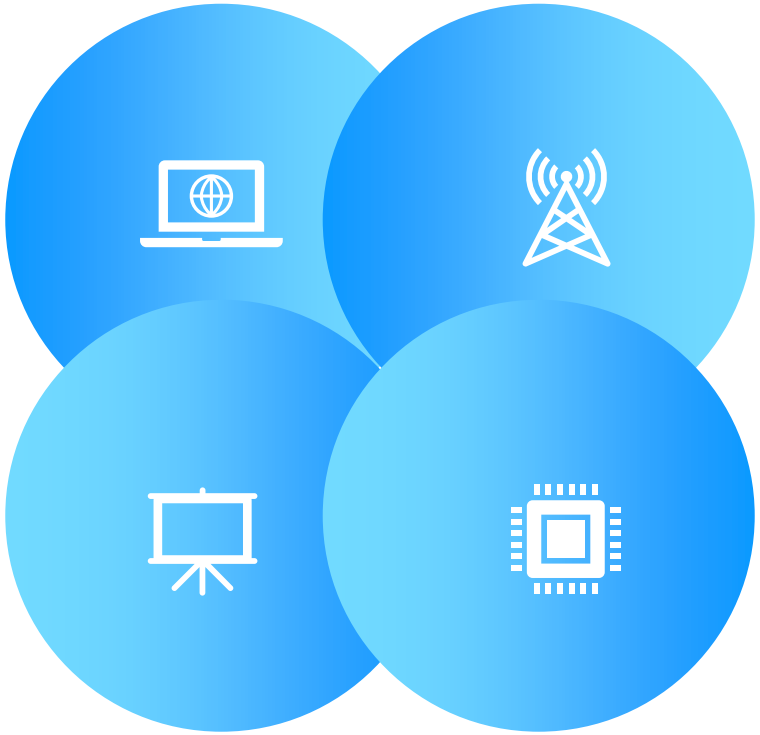
对日期、金额、电话、邮箱等格式后处理，规格化

6279	Q06279	f99f8f8183a	3米延长进水管1根多少钱?	95元
6280	Q06280	f99f8f8183a	1.5米排水管1根多少钱?	52元
6281	Q06281	f99f8f8183a	进水三通多少钱?	30元

优　化　方　向

优化训练流程

调节超参数



更换语义模型

数据处理和转换

02 数据预处理

PART TWO

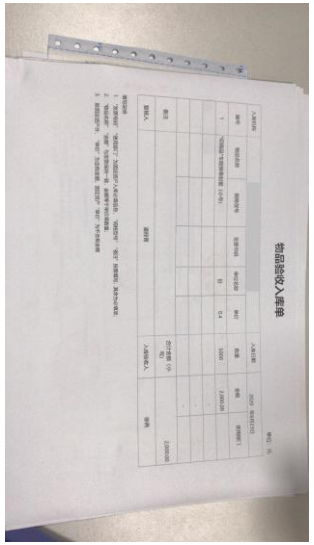
数 据 预 处 理

噪声干扰

无关文本干扰

多票据

要区分票据空间位置

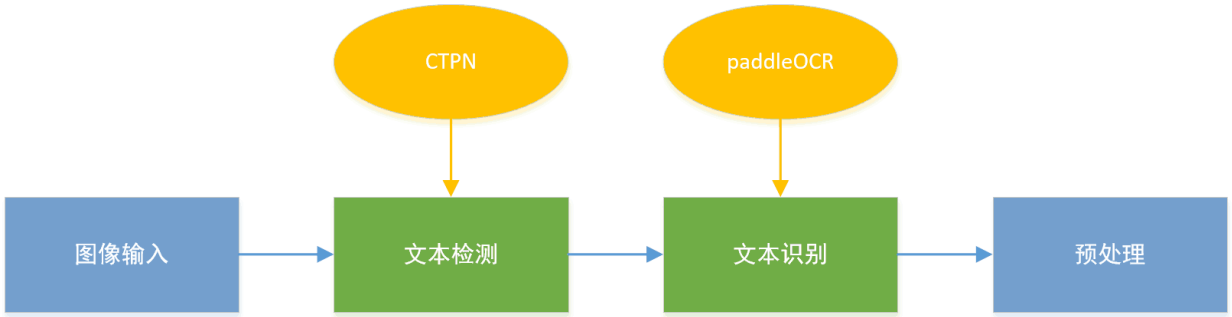


图像校正

对于拍摄角度不同的照片需要进行倾斜校正

滤波

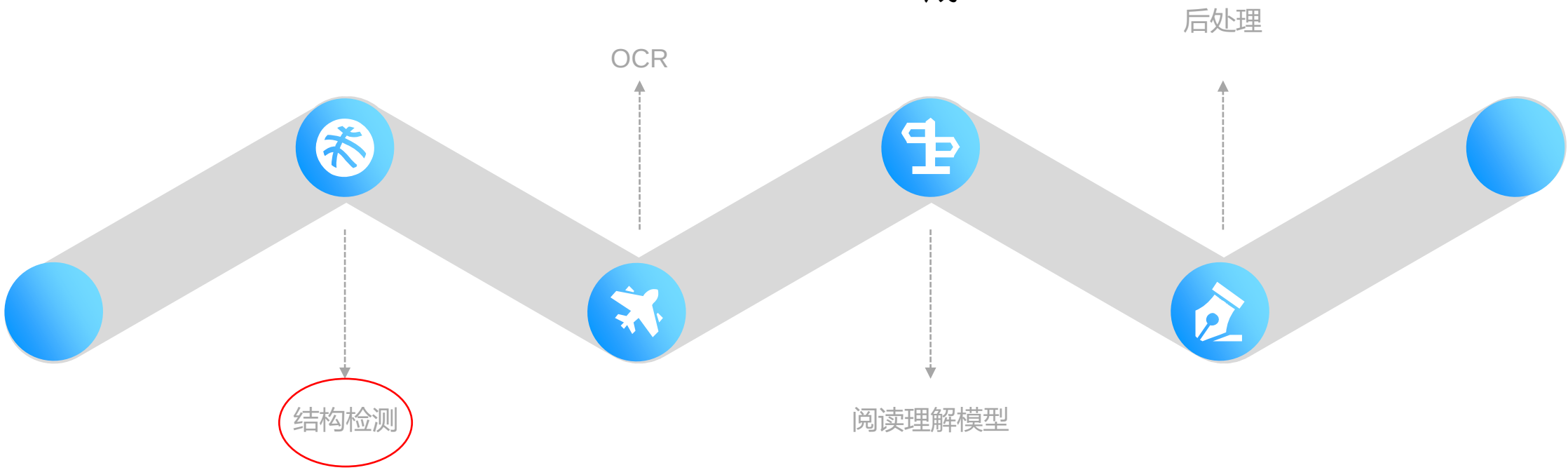
对于噪声（如波浪形纹路）较多的图片需要滤波。



中文场景模型检测CRNN-CTC结构（CNN+RNN+CTC）

- “繁体->简体”、“大写->小写”、“删除空格”、“删除符号”
- 数据增强（调整亮度，调整对比度，调整饱和度，随机旋转）
- 生成用于预测的模型

* 函数 init_eval_parameters：初始化预测参数
* 函数 resize_img：调整图片大小
* 函数 read_image：读取图片并做相应处理
* 函数 infer：对单张图片进行文字识别
* 函数 eval_all：对所有图片进行识别，并生成



数据增强（基于baseline）

```python
def random_brightness(img):#随机调整亮度
    prob = np.random.uniform(0, 1)
    if prob < train_opt['image_distort_strategy']['brightness_prob']:
        brightness_delta = train_opt['image_distort_strategy']['brightne
        delta = np.random.uniform(-brightness_delta, brightness_delta) +
        img = ImageEnhance.Brightness(img).enhance(delta)
    return img
```

```python
def random_contrast(img):#随机调整对比度，进行数据增强
    prob = np.random.uniform(0, 1)
    if prob < train_opt['image_distort_strategy']['contrast_prob']:
        contrast_delta = train_opt['image_distort_strategy']['contrast_d
        delta = np.random.uniform(-contrast_delta, contrast_delta) + 1
        img = ImageEnhance.Contrast(img).enhance(delta)
    return img
```

```python
def random_saturation(img):#随机调整饱和度，进行数据增强
    prob = np.random.uniform(0, 1)
    if prob < train_opt['image_distort_strategy']['saturation_prob']:
        saturation_delta = train_opt['image_distort_strategy']['saturati
        delta = np.random.uniform(-saturation_delta, saturation_delta) +
        img = ImageEnhance.Color(img).enhance(delta)
    return img
```

```python
def rotate_image_0(img):
    """
    图像增强，增加随机旋转角度
    """
    prob = np.random.uniform(0, 1)
    if prob > 0.:
        angle = np.random.randint(-10, 10)
        img = img.convert('RGBA')
        img = img.rotate(angle, resample=Image.BILINEAR, expand=0)
        fff = Image.new('RGBA', img.size, (127, 127, 127, 127))
        img = Image.composite(img, fff, mask=img).convert('RGB')
    return img
```

文　本　检　测

·12· 副刊天地 参考消息 2020年4月13日

【美国《纽约时报》网站4月11日报道】题：他本该预见即将发生的事情：特朗普

美媒披露 特朗普如何一步一步输掉战"疫"

病毒应对失败的背后（记者　埃里克·利普顿　戴维·桑格　玛吉·哈伯曼　迈克尔·希尔　马克·马泽蒂　朱利安·巴恩斯）

1月28日晚，美国退伍军人事务部的高级医疗顾问卡特·迈彻尔博士给分散在政府各部门和各大学中的一群公共卫生专家发去电子邮件："不管你怎样去应对，局面都将会是糟糕的。这场疫情的预期规模已经看起来让人难以置信。"

疫情竟被置于次要位置

在美国发现首例武汉病毒感染病例一周之后，迈彻尔就在敦促美国公共卫生官僚机构的高层人士们警觉起来，为可能采取远为极端的措施做好准备。现在，人们预计这场疫情大流行将夺走数万美国人的生命。

发出这种呼吁的并非只有他一个人。整个1月，特朗普政府内部形形色色的人物——从白宫高级顾

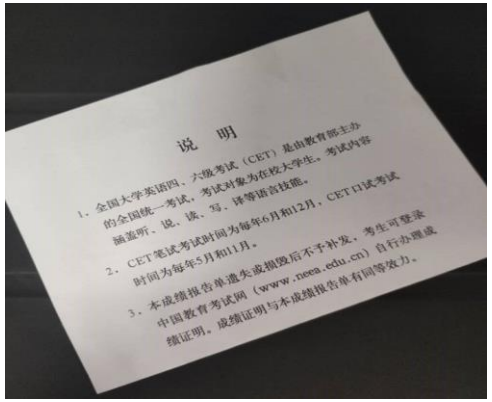
4月11日，两名男子戴着口罩从美国首都华盛顿的国会大厦附近走过。（刘杰　摄）

息上面，而在他们的意见被总统勉强接受之前，至关重要的几个星期白白过去了——在这段时间里，病毒的扩散很大程度上是畅通无阻的。

当特朗普在3月中旬最终同意建议在全国范围实施社交疏远措施，从而实际上使经济多半陷入停摆时，在他的一些最亲密的助手看来，他似乎垂头丧气、一蹶不振。一名助手形容他对于这场危机的呈现方式感到"抑郁"和"困惑"。他曾寄希望帮助自己赢得连任竞选的经济突然间变得一团糟。

这位助手说，他只得通过举行每天一次的白宫媒体吹风会来重拾自己的气势。在吹风会上，他常常寻求改写过去几个月的历史。他一度宣称，自己"在这场疫情被称为大流行之前，就感觉到这是一场大流行"，并在另一次吹风会上坚称自己不得不充当"国家的拉拉队长"，这仿佛解释了他为什么未能让公众准备好应对正在到来的状况。

各种措施计划被搁置一旁



说　明

1．全国大学英语四、六级考试（CET）是由教育部主办的全国统一考试，考试对象为在校大学生。考试内容涵盖听、说、读、写、译等语言技能。

2．CET笔试考试时间为每年6月和12月，CET口试考试时间为每年5月和11月。

3．本成绩报告单遗失或损毁后不予补发，考生可登录中国教育考试网（www.neea.edu.cn）自行办理成绩证明。成绩证明与本成绩报告单有同等效力。

·12· 副刊天地 参考消息 2020年4月13日

【美国《纽约时报》网站4月11日报道】题：他本该预见即将发生的事情：特朗普

美媒披露 特朗普如何一步一步输掉战"疫"

病毒应对失败的背后（记者　埃里克·利普顿　戴维·桑格　玛吉·哈伯曼　迈克尔·希尔　马克·马泽蒂　朱利安·巴恩斯）

1月28日晚，美国退伍军人事务部的高级医疗顾问卡特·迈彻尔博士给分散在政府各部门和各大学中的一群公共卫生专家发去电子邮件："不管你怎样去应对，局面都将会是糟糕的。这场疫情的预期规模已经看起来让人难以置信。"

疫情竟被置于次要位置

在美国发现首例武汉病毒感染病例一周之后，迈彻尔就在敦促美国公共卫生官僚机构的高层人士们警觉起来，为可能采取远为极端的措施做好准备。现在，人们预计这场疫情大流行将夺走数万美国人的生命。

发出这种呼吁的并非只有他一个人。整个1月，特朗普政府内部形形色色的人物——从白宫高级顾

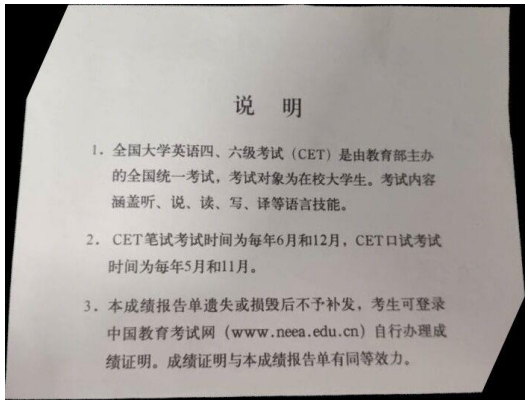
4月11日，两名男子戴着口罩从美国首都华盛顿的国会大厦附近走过。（刘杰　摄）

息上面，而在他们的意见被总统勉强接受之前，至关重要的几个星期白白过去了——在这段时间里，病毒的扩散很大程度上是畅通无阻的。

当特朗普在3月中旬最终同意建议在全国范围实施社交疏远措施，从而实际上使经济多半陷入停摆时，在他的一些最亲密的助手看来，他似乎垂头丧气、一蹶不振。一名助手形容他对于这场危机的呈现方式感到"抑郁"和"困惑"。他曾寄希望帮助自己赢得连任竞选的经济突然间变得一团糟。

这位助手说，他只得通过举行每天一次的白宫媒体吹风会来重拾自己的气势。在吹风会上，他常常寻求改写过去几个月的历史。他一度宣称，自己"在这场疫情被称为大流行之前，就感觉到这是一场大流行"，并在另一次吹风会上坚称自己不得不充当"国家的拉拉队长"，这仿佛解释了他为什么未能让公众准备好应对正在到来的状况。

各种措施计划被搁置一旁



说　明

1．全国大学英语四、六级考试（CET）是由教育部主办的全国统一考试，考试对象为在校大学生。考试内容涵盖听、说、读、写、译等语言技能。

2．CET笔试考试时间为每年6月和12月，CET口试考试时间为每年5月和11月。

3．本成绩报告单遗失或损毁后不予补发，考生可登录中国教育考试网（www.neea.edu.cn）自行办理成绩证明。成绩证明与本成绩报告单有同等效力。

03 模型构建

PART TWO

模　型　构　建

预训练模型
- bert-base-chinese ☆
- ernie ☆
- roberta
- electra-base

ernie在训练中引入的DLM能有效地提升模型对文本相似度的建模能力，综合考虑以及训练情况，选用ernie作为预训练模型

➢ 提交格式转换


```
"root" :  8615 items
  [  100 items
    0 : {  2 items
      "answer" : string "九年十二月"
      "questionId" : string "Q00001"
    }
    1 : {  2 items
      "answer" : string "财政部、中国证券监督管理委员会"
      "questionId" : string "Q00002"
    }
    2 : {  2 items
      "answer" : string "中兴华会计师事务所李尊农首席合伙人"
      "questionId" : string "Q00003"
    }
```

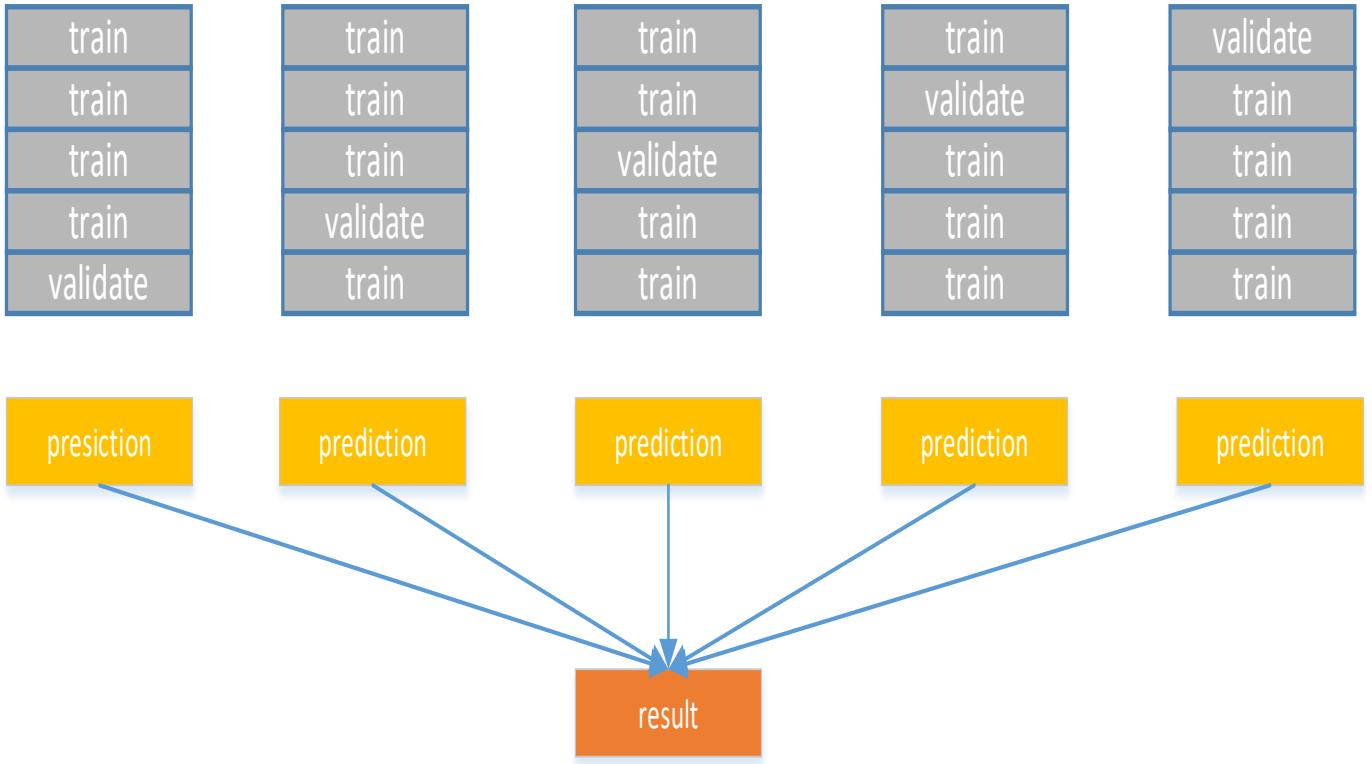

model	F1_score(epochs=1)
Bert-base-chinese	67.23
ernie	67.45

模　型　优　化

➢ 优化参数

```
############参数配置###############
# 模型名称
MODEL_NAME = "bert-base-chinese"
fold_num=5  #采用五折交叉验证
# 最大文本长度
max_seq_length = 512
# 文本滑动窗口步幅
doc_stride = 128
# 训练过程中的最大学习率
learning_rate = 3e-5
# 训练轮次
epochs = 8
# 数据批次大小
batch_size = 8
# 学习率预热比例
warmup_proportion = 0.1
# 权重衰减系数，类似模型正则项策略，避免模型过拟合
weight_decay = 1e-4
```

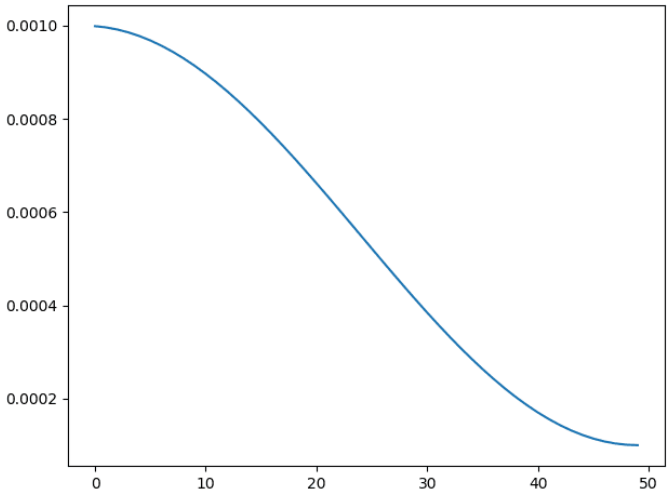
➢ 五折交叉融合

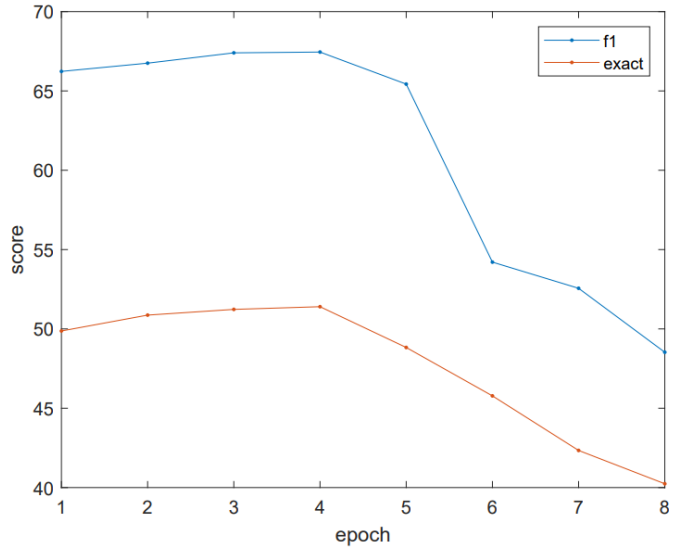


参　数　调　优

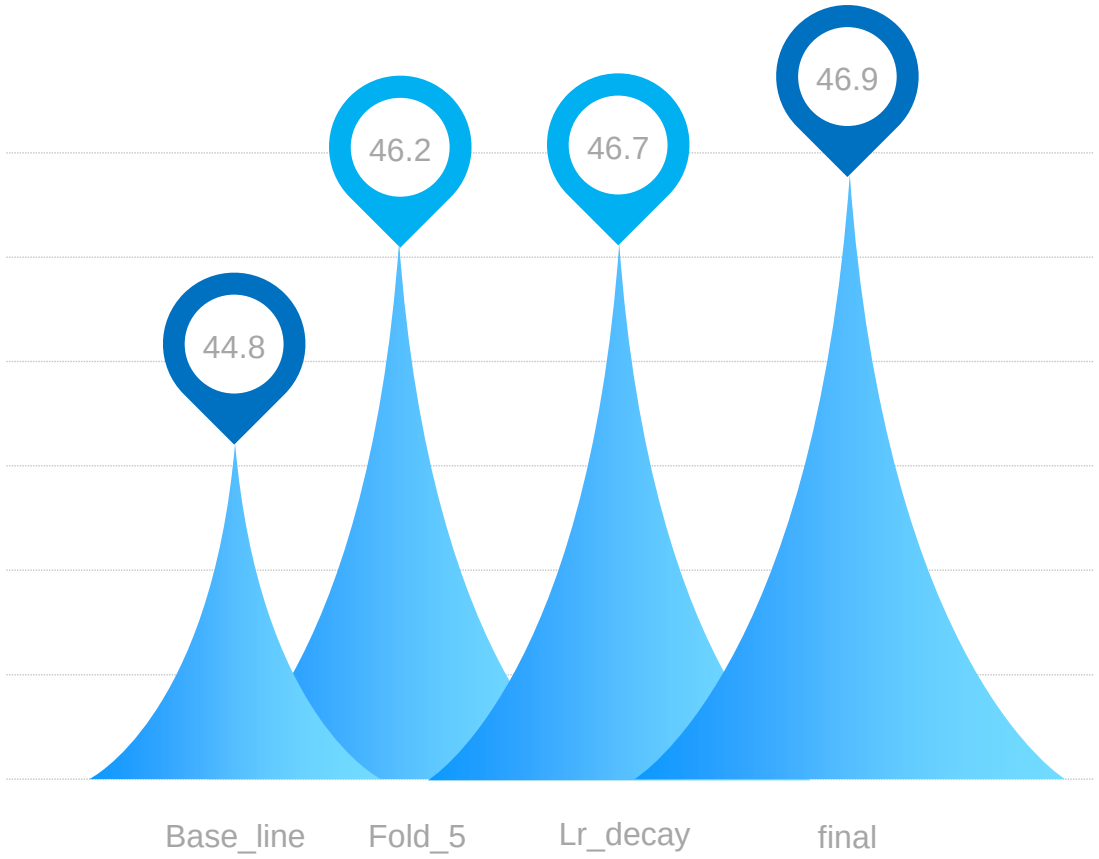
	fold_num	Max_seq_length	Doc_stride	Learning_rate
优化前默认参数	3	256	64	1e-1
优化后参数	5	512	128	3e-5

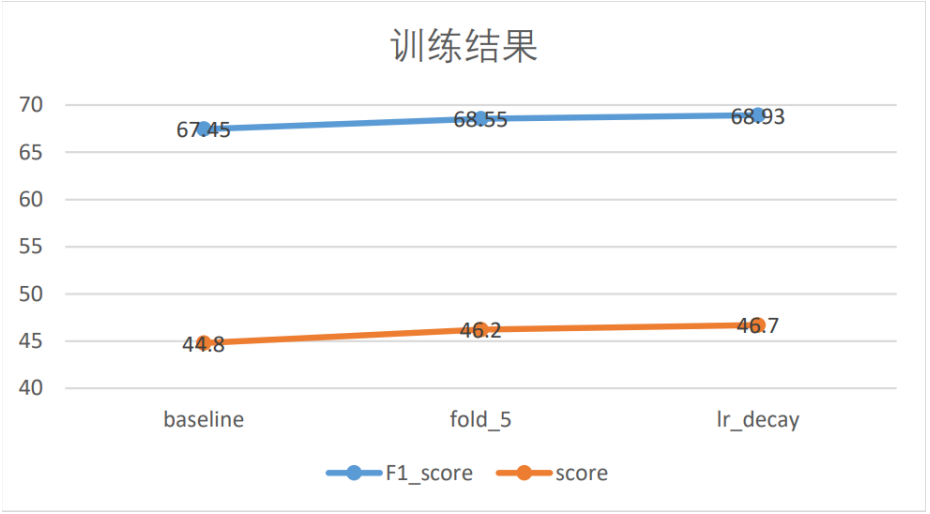
	epochs	Bacth_size	Warmup_proportion	Weight_decay
优化前默认参数	1	4	1	1e-4
优化后参数	4	8	0.1	1e-4





训　练　结　果





	F1_score	Score
baseline	67.45	44.8
fold_5	68.55	46.2
lr_decay	68.93	46.7

04 赛题总结

PART TWO

提　交　排　名

比赛详情　阶段　参赛提交　排名结果　论坛交流　队伍管理

报名训练阶段　第 1 轮排名阶段　第 2 轮排名阶段　终选答辩阶段

阶段描述

5月12日将提供测试集，接受提交，开展第1轮排名，勿忘提交打包为ZIP

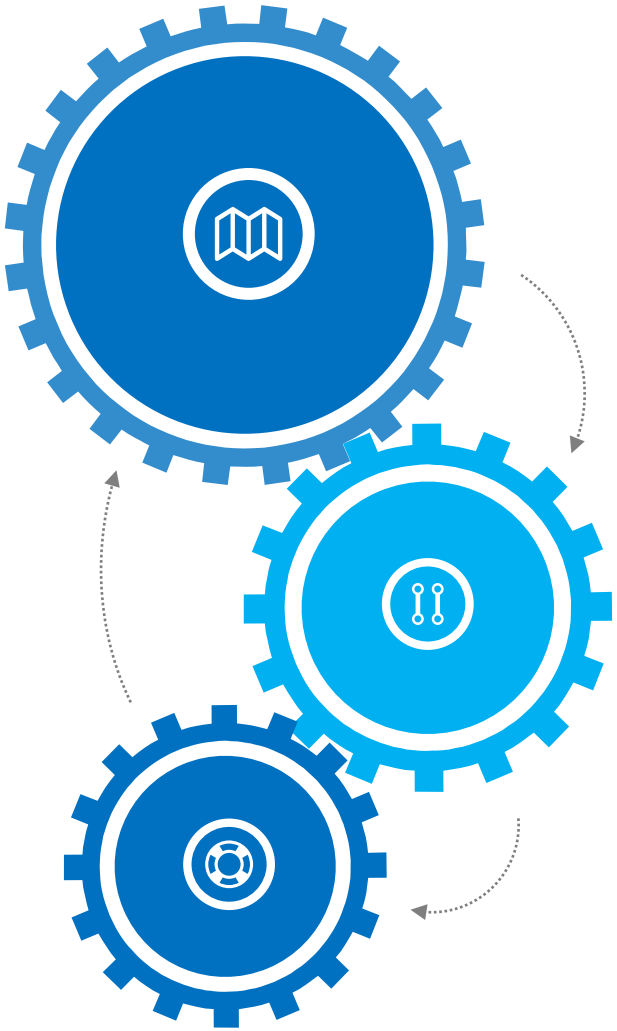
每日最大提交数： 1

最大提交总数： 999

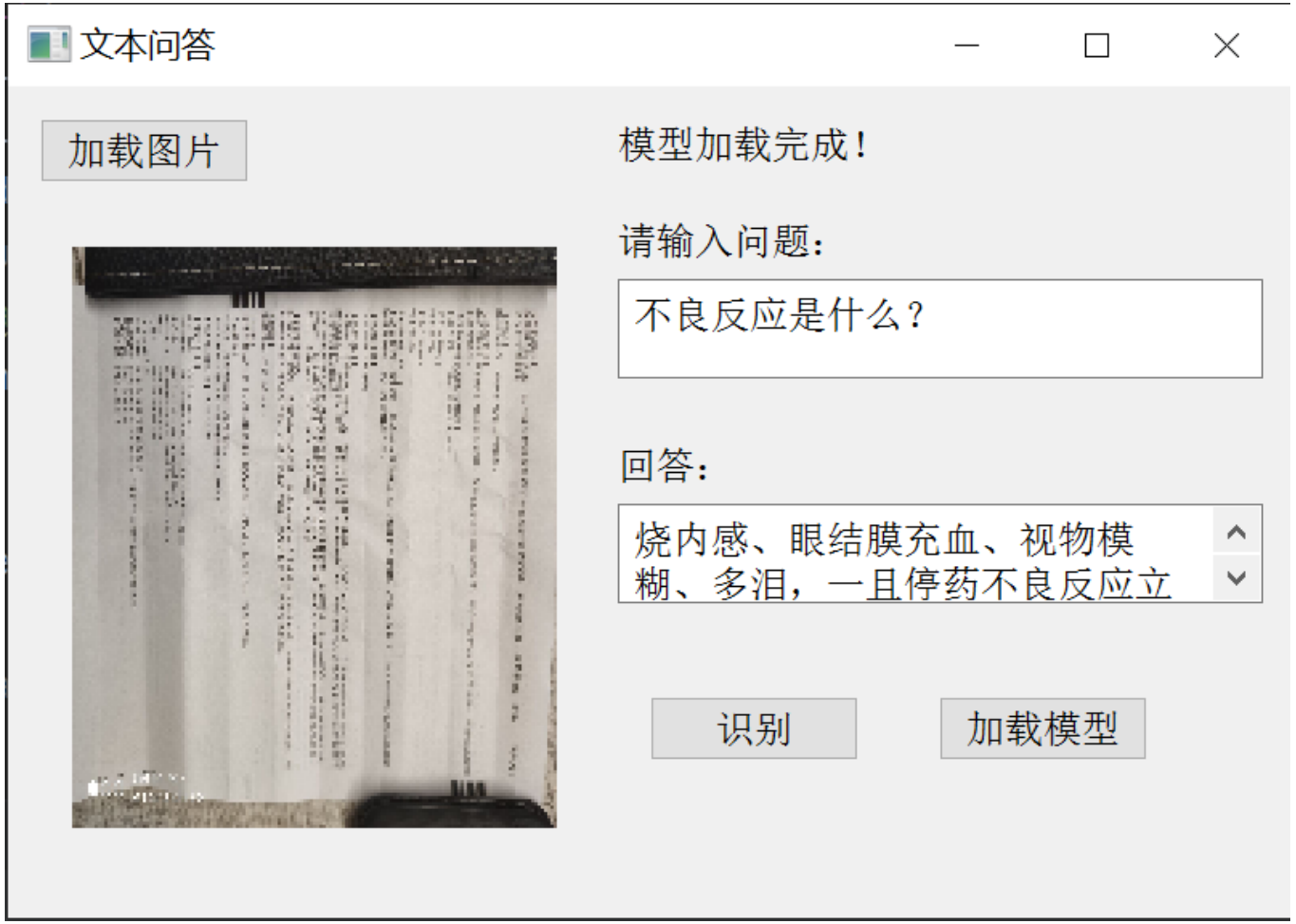
Download CSV

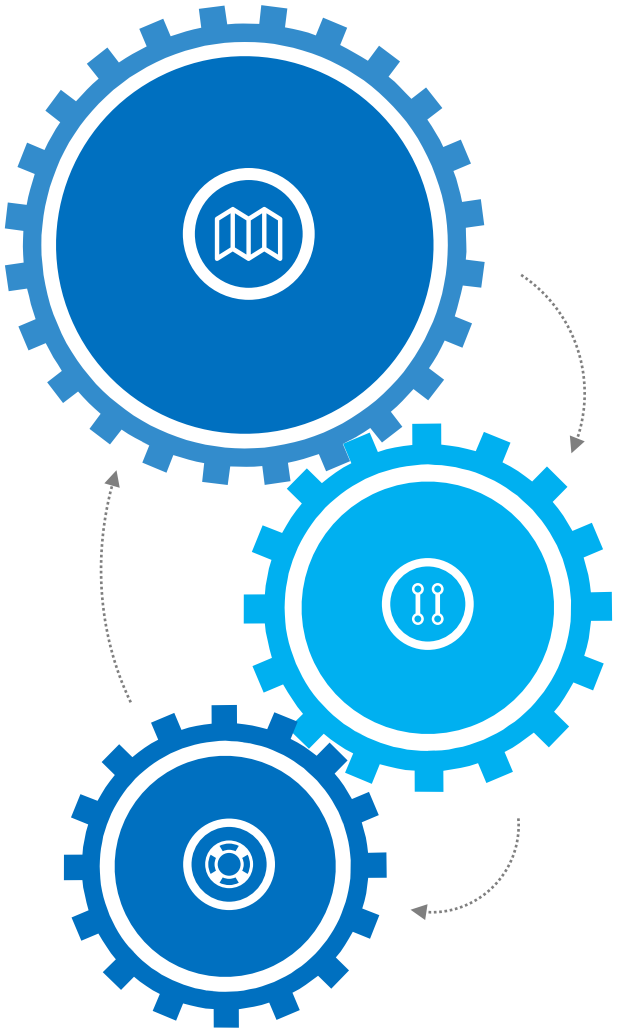
Results

#	用户	提交	最后提交日期	团队名称	Score ▲	Char_Match ▲	Exact_Match ▲
1	zytx121	9	05/15/21		0.533 (1)	0.603 (1)	0.462 (1)
2	bushou	7	05/15/21		0.526 (2)	0.594 (2)	0.458 (2)
3	cccccczk	3	05/15/21		0.525 (3)	0.594 (3)	0.455 (3)
4	KangChenfei	6	05/15/21		0.518 (4)	0.584 (4)	0.452 (4)
5	sun188	4	05/15/21		0.514 (5)	0.579 (6)	0.449 (5)
6	Terence	2	05/14/21		0.512 (6)	0.580 (5)	0.445 (6)
7	song	4	05/15/21	深度不学习	0.508 (7)	0.577 (7)	0.439 (7)
8	Max	3	05/16/21	SWHL	0.497 (8)	0.558 (8)	0.436 (8)
9	mikezhang95	3	05/14/21		0.478 (9)	0.529 (11)	0.427 (9)
10	wangw	2	05/14/21		0.474 (10)	0.541 (9)	0.406 (10)
11	glh9803	3	05/16/21	glh9803	0.468 (11)	0.532 (10)	0.405 (11)
12	Langouste	5	05/16/21		0.448 (12)	0.512 (12)	0.384 (12)
13	insteadzou	1	05/15/21		0.340 (13)	0.382 (13)	0.297 (13)
14	finlay	4	05/14/21	阿水	0.184 (14)	0.210 (14)	0.158 (14)
15	maxmon	2	05/16/21		0.061 (15)	0.100 (15)	0.021 (15)



图形界面构建


文本问答
加载图片
模型加载完成！
请输入问题：
不良反应是什么？
回答：
烧内感、眼结膜充血、视物模糊、多泪，一旦停药不良反应立
识别
加载模型



赛　题　总　结

飞桨

➢ 对Paddle平台的使用体验

- 提供免费的GPU算力
- 给出baseline流程
- Jupyter notebook调试方便
- 可视化方便参数优化

赛　题　总　结

➢ 心得体会

本次机器学习的课程设计通过参加了一个实际的人工智能创新大赛，对大赛的流程有了一定的了解。

刚开始接触赛题时有点害怕，不知从何下手，但是paddle平台已经给出了baseline的流程，解决了我们开始时的一系列的困惑。从一开始跑通模型获得成就感，到后面为提交格式而发愁，第一次成功提交得到排名结果的喜悦，后面逐步提升训练精度的满足感。非常感谢paddle平台，免费提供了GPU算力，平且给了我们展示自己的机会和平台。

参加本次比赛，对paddle上的一系列NLP自然语言处理模型了解了很多，并且为了提高训练精度和排名，也查阅了大量的文本检测模型优化方法。虽然最终精度有所提高，但是排名却并没有提升，略有遗憾。但是没有关系，本次积累的经验对自己而言更加宝贵，希望以后能多多参与此类比赛，是一个非常好的实践机会。

AIWIN

谢谢

顾立辉|陈震辉