⽂本分类&匹配

# 介绍

## 前⾔

NLP，中⽂叫⾃然语⾔处理，简单来说，是⼀⻔让计算机理解、分析以及⽣成⾃然语⾔的学科， ⽬标是让计算机处理或“理解”⾃然语⾔，以执⾏语⾔翻译和问题回答等任务。

NLP包含⾃然语⾔理解(Natural Language Understanding，NLU) 和⾃然语⾔⽣成

（Natural Language Generation, NLG）两个重要⽅向：

⾃然语⾔理解NLU旨在将⼈的语⾔形式转化为机器可理解的、结构化的、完整的语义表示，通俗来 讲就是让计算机能够理解和⽣成⼈类语⾔。

⾃然语⾔⽣成NLG旨在让机器根据确定的结构化数据、⽂本、⾳视频等⽣成⼈类可以理解的⾃然语

⾔形式的⽂本。

在NLP领域，⽬前平台⽀持⽂本分类、⽂本匹配两类模型；两类模型的功能区别如下：

⽂本分类：搭建定制化的⽂本分类模型，实现⽂本按照类型做⾃动分类

⽂本匹配：搭建定制化的⽂本匹配模型，⽤于判断两句⽂本是否语义⼀致

下⾯是⽂本分类&匹配模型的技术⽂档。对于同样操作⽅式以⽂本分类为例，请悉知。

## 应⽤场景

电商评论调研：应⽤在消费者对产品的反馈信息，或者检测在线评论中的差评信息。

情感分析：提取⽂本的情感内容为⽬标的分类⽅法，利⽤⼀些情感得分指标来量化定性数据。

智能客服：根据⼈工输入与系统问题做匹配，基于相似度给出问题的回答，适⽤于⼈工繁忙等场景。

## 使⽤流程

模型训练的基本流程如下所示（刘旭最终替换）



# 数据准备

## 创建数据集

#### 数据集格式要求

数据上传只⽀持CSV/TXT格式⽂件上传；

⽂本数据集⽀持单⽂件上传和多⽂件上传两种⽅式；多⽂件上传将多个CSV/TXT打包成ZIP压缩 包上传，且⽂件格式保持⼀致；CSV⽂件编码⽅式为UTF-8

对于⽂本分类数据集CSV/TXT⽂件第⼀列为分类名称（label）、第⼆列为⽂本数据(text)；对于

⽂本匹配数据集CSV/TXT⽂件第⼀列为分类名称（label）、第⼆、三列为⽂本数据(text)。分类 名称

⽀持中⽂/英⽂/数字/下划线/中划

#### 数据集内容要求

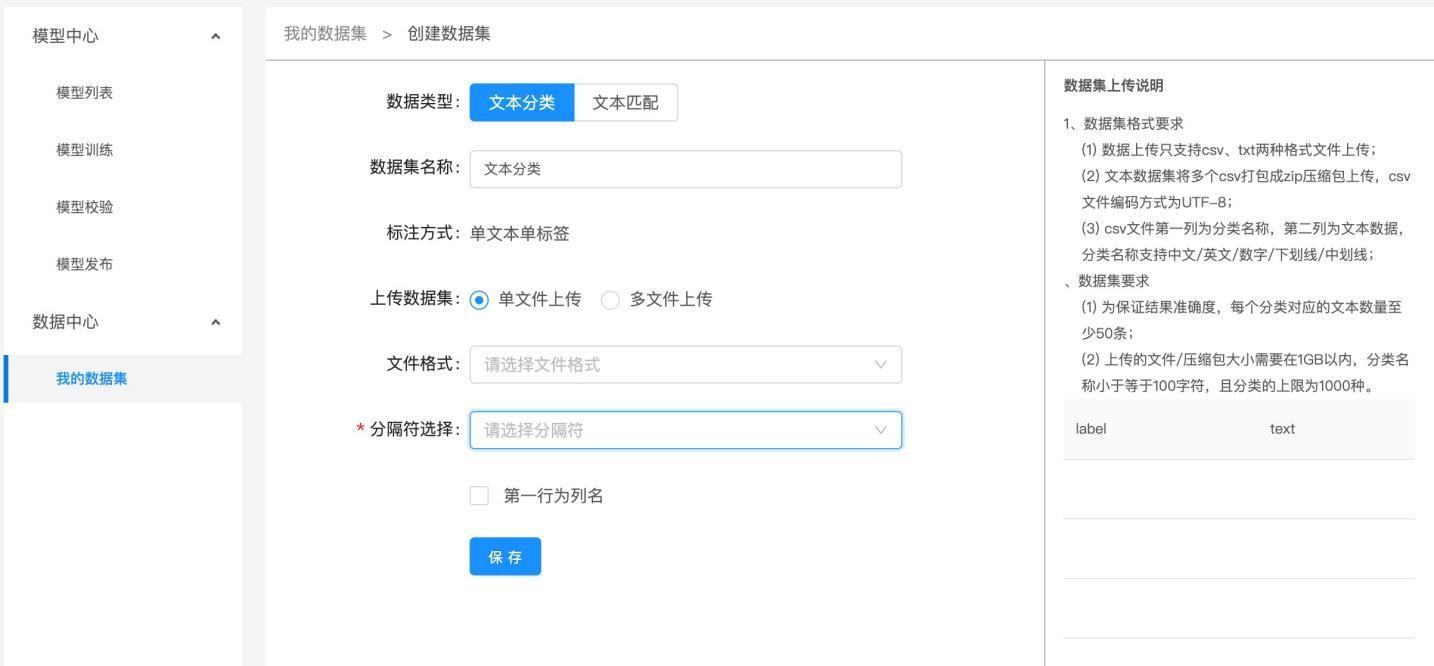
为保证结果准确度，每个分类对应的⽂本数量建议不少于50条。

上传的⽂件/压缩包⼤⼩需要在1GB以内，且分类的上限为1000种。

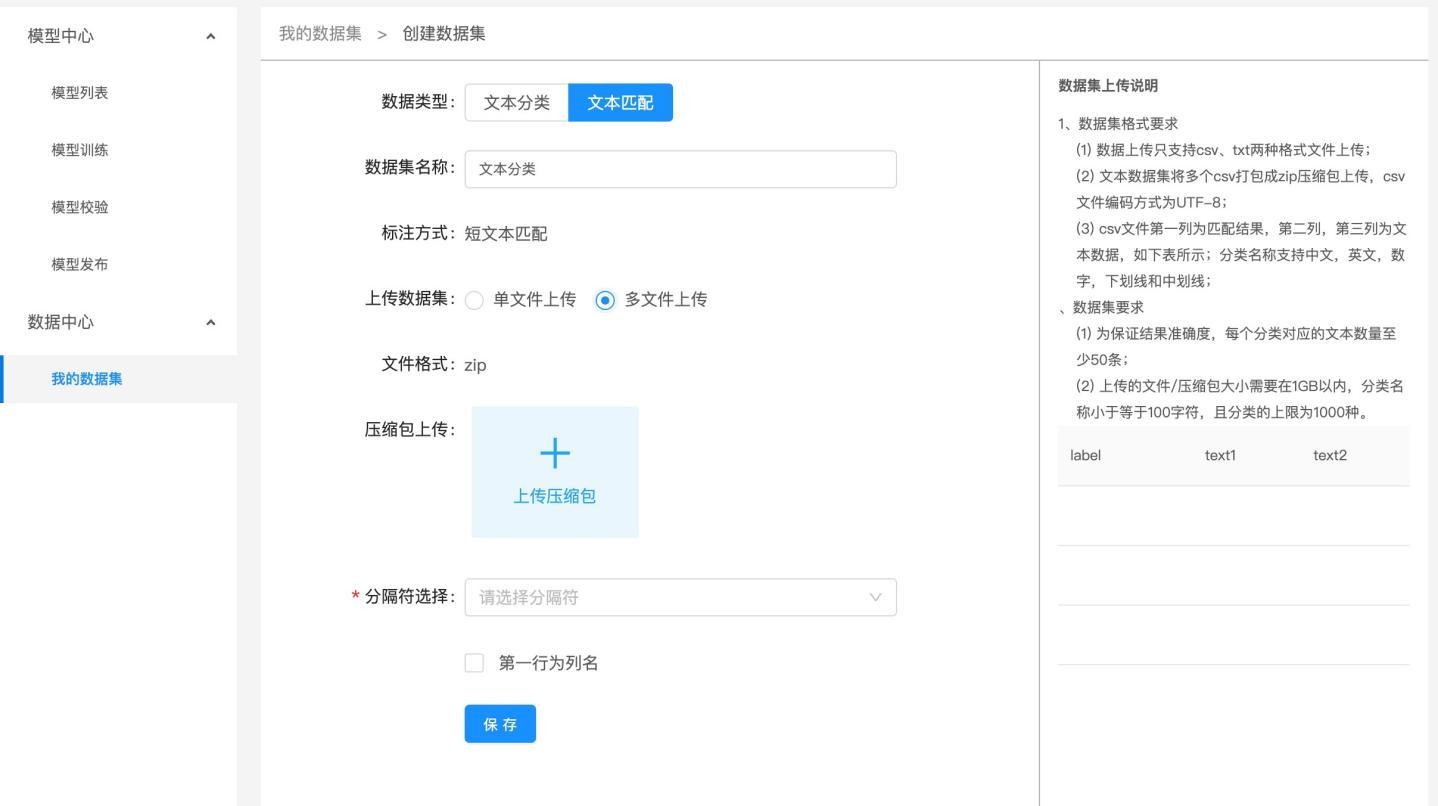
#### 创建数据集

1. 左侧任务导航【模型中⼼】>【我的数据集】>【创建数据集】。
2. 选择数据类型，创建对应的数据集，数据集创建分为两种类型：⽂本分类、⽂本匹配；选择后输入数 据集名称，选择上传数据集⽅式，添加对应格式⽂件，并选择解析⽂件的配置。⽂件上传过程中，请勿 切换其他⻚⾯或关闭。
3. 上传完成点击【保存】，开始解析数据集。

⽂本分类数据集上传（单⽂本上传为例）



⽂本匹配值数据集上传（多⽂本上传为例）



1. 上传数据集后的状态呈处理中状态，处理中的数据集相关信息会在【数据中⼼】>【我的数据集】下 的数据列表中显示。为保证数据集正常解析，请按格式创建数据集（注：zip包下不符合格式要求的内容

将会被忽略掉）。



1. 解析成功后的数据集可以正常参与训练。

## 数据管理

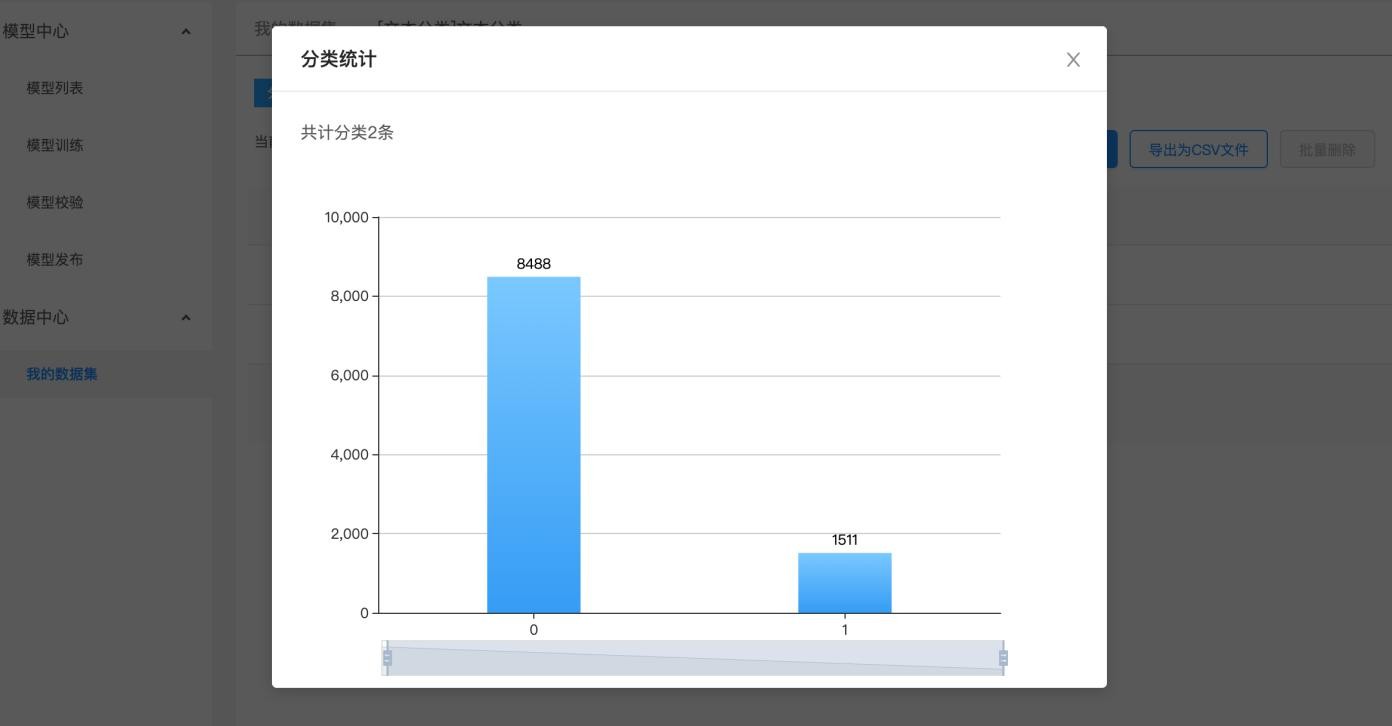
### 前⾔

对数据管理的操作以分类举例，⽂本匹配同。

### 操作流程

* 1. 在【我的数据集】中，除了可以创建新的数据集之外，还可以对已上传的数据集进⾏查看、修改名 称、补充数据等操作。
  2. 点击数据集操作中的【查看】，可以按照分类查看及原始数据查看两种形式；分类查看可针对每个 分类进⾏查看操作。



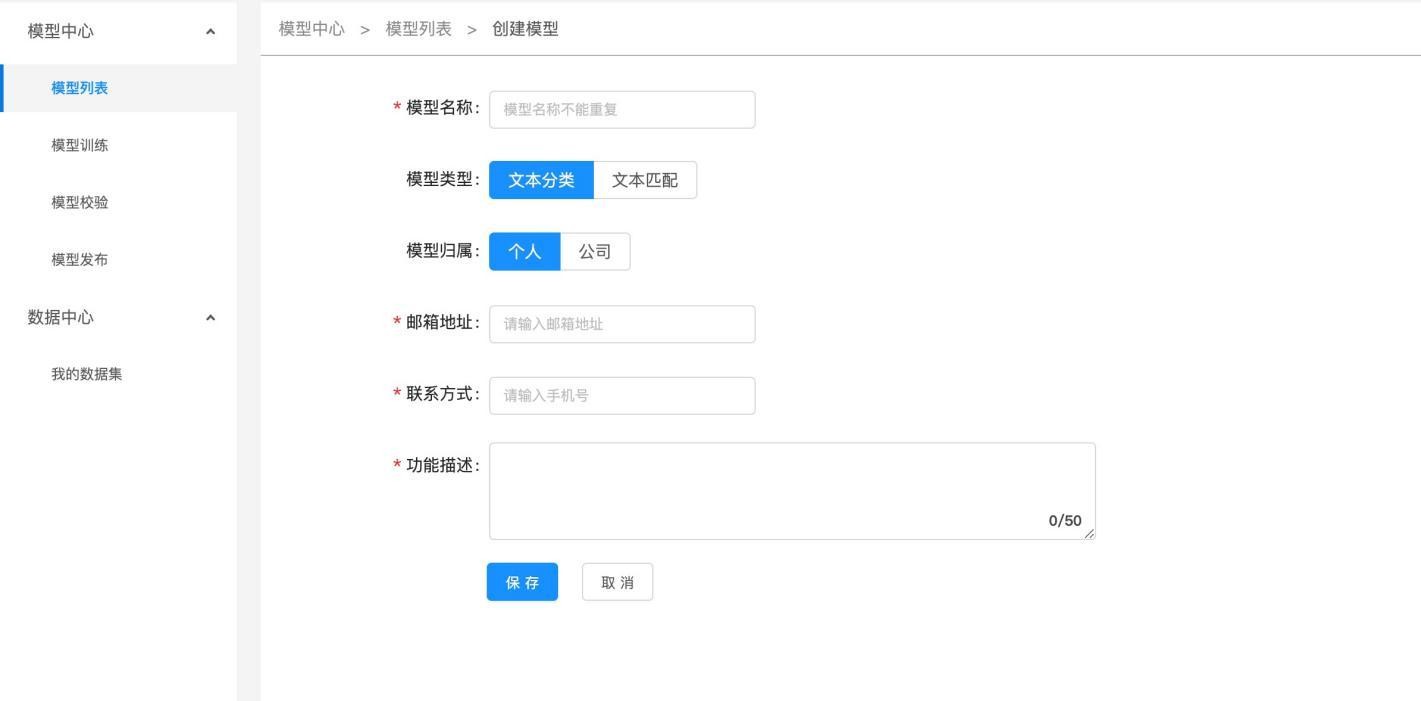
* 1. 分类查看时可以查看分类统计分析图表，也可以将解析出来的⽂件导出为CSV⽂件。 分类统计如下图：

点击【导出CSV⽂件】操作，通过浏览器直接按分类导出⽂本解析结果。

* 1. 点击数据集下分类的【查看】操作，可查看每个分类的详细内容。

# 模型管理

## 模型创建

在导航【模型列表】下点击【创建模型】，填写模型名称、模型类型、联系⽅式等信息即可创建模型

## 模型列表

模型创建成功后可以在【模型列表】中查看刚才创建的模型，模型列表可以筛选展示的模型类型。



### 模型版本查看

模型训练状态可以在【模型列表】下对应模型查看版本训练状态；训练完成的模型可进⾏发布、校验、 或查看训练集的操作；



训练完成后，点击【完整评估报告】，可看到模型效果，以及详细的模型评估报告；⻅下【模型评估结 果】

点击【查看训练集】按钮可查看训练集下的对应分类、⽂本数及每个分类的训练效果。

点击【查看详情】可以查看对应分类下的图⽚详情。



### 模型评估结果

#### 整体评估

在这个部分可以看到模型训练整体的情况说明，包括基本结论、准确率、F1-score、精确率、召回率等。这部分模型效果的指标是基于训练数据集，随机抽出部分数据不参与训练，只参与模型效果评估计

算。所以当数据量较少时，参与评估的数据会更少，⽆法完全准确体现模型效果。

查看模型评估结果时，需要思考在当前业务场景，更关注精确率与召回率哪个指标。是更希望减少误识

别，还是更希望减少漏识别。前者更需要关注精确率的指标，后者更需要关注召回率的指标。同时F1- score可以有效关注精确率和召回率的平衡情况，对于希望准确率与召回率兼具的场景，F1-score越接近1效果越好。

注意：若想要更充分了解模型效果情况，建议发布模型为API后，通过调⽤接⼝批量测试，获取更准确 的模型效果。

整体评估结果如下图（以⽂本分类为例）



#### 详细评估

在这个部分可以看到上述训练效果具体的数据情况及识别错误⽂本信息。



⽂本匹配评估结果如下



# 模型训练

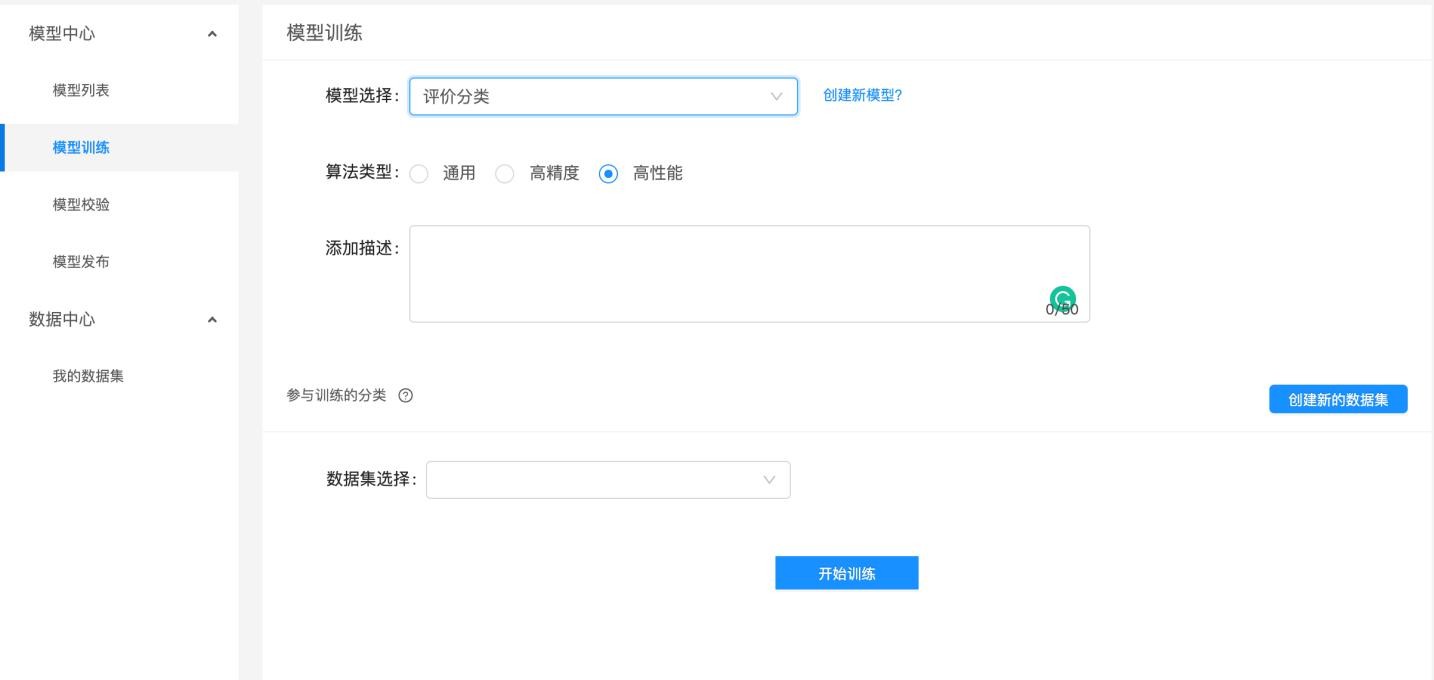
【模型训练】通过模型版本列表中【训练新版本】即可训练当前选中的模型，也可通过导航【训练模 型】中，进⾏模型训练

## 模型选择

选择参与此次训练的模型。如果您是通过模型版本列表中【训练新版本】训练模型，模型名称会默认带 过来，您也可以进⾏修改。

## 算法类型选择

算法类型分为：通⽤、⾼精度、⾼性能；不同算法类型对应不同模型，可以根据业务场景需要，选择符 合⾃身要求的类型。



## 训练分类选择

「数据集选择」下拉框选择参与训练的数据集，默认选择该数据集下全部分类参与训练，可以通过【查 看分类详情】操作选择参与训练的分类。

注意：⾄少选择2个及以上的分类才可以开始训练。



选择所需的分类后点击【确认】按钮可以添加选中的分类。

## 模型训练

点击【开始训练】，则开始训练模型；开始训练之后会跳转到相对应的模型下的版本列表中查看训练状 态；

训练时间与数据量⼤⼩有关，请耐⼼等待。



训练中的模型也可以停⽌训练，点击操作中的【停⽌训练】可停⽌训练；

# 模型校验

通过模型版本列表操作栏中点击【校验】即可校验当前版本的模型，也可通过导航【校验模型】中，选 择模型及版本进⾏校验。



选择好对应的模型、版本，即可点击【启动模型校验服务】，启动模型校验的时间与模型有关，⼤约需 要等待1分钟即可启动成功。



模型校验可以通过直接在左侧输入框内输入⽂本进⾏单⽂本校验，也可以通过点击【批量校验】，上传

⽂件进⾏多⽂本校验。

选择好⽂件，可点击【上传】，将⽂件上传并进⾏校验； 校验模型选择的⽂件越⼤所需时间越⻓；



校验成功之后会返回识别结果，点击对应的⽂本数据，可在右侧查看对应⽂本的校验结果。 点击【很多】查看该⽂本数据的详细信息。

# 模型发布

## 操作说明

训练完成后，可将模型部署在云服务上。

模型发布可以通过选择「模型列表」下版本对应【发布】操作或者在左侧导航点击【模型发布】进⾏。 云服务发布需要添加服务名称、接⼝地址、或可添加其他要求。接⼝地址不可重复。

云服务提交申请之后会跳转⾄对应模型下版本列表，查看发布状态，发布状态为已发布可对该模型进⾏ 服务详情查看或者下线。

