NLU数据集需求

NLU模块	需要的数据	数量	优先级
意图识别	游戏中另外需要的意图种类	根据需求确定	高
意图识别	每一类意图的文本	50条/意图	高
NER	游戏命名实体确定	根据需求确定	高
NER	带游戏命名实体的文本标注	20W条	高
情感分析	游戏中带情感倾向的文本	20W条	中
spam detection	敏感及其他文本的标注	1W条/种类	中
指代消解	游戏聊天文本的指代消解标注	5W条	中
文本补全	/	/	/
opinion detection	/	/	/
topic detection	/	/	/
domain detection	/	/	/

意图识别需求：

- 需要确定新的意图种类，目前存在的意图有身份设定、日常用语、小技能等，更多请看
  - 数量：根据实际需求确定
- 需要每一类意图的数据集，举例: 意图：爱不爱妈妈 数据：喜欢娘亲吗；爱娘亲吗；爱娘亲不；...
  - 数量：每一类意图50条左右文本，最好能包含游戏中的真实文本

NER需求

- 确定在游戏中需要的命名实体列表（举例，游戏中的实体可能包括：装备、技能等）
- 包含游戏实体的真实文本并进行命名实体标注，举例： 文本：护腕（标注为装备实体）精啄，帮会点会心。 会心（标注为技能实体）高了 秒伤提升非常明显 ，有时候铁画能爆个3W多。
  - 数量：总文本数20W条，每个实体对应的文本不小于1000条

情感分析需求

- 为了学习在游戏语境下的情感，所以需要带情感倾向的游戏文本，举例：
  - 愤怒（情感）：削就削，老子90用2200幸运装打木桩都能8000多了，该不该削你心里没点哈数？有tm什么开不开心的？自己玩好你自己的，别他🐴成天来装小学生好吗，神相偷你家钱

了?

Spam Detection

***这一块目前硕芃负责情感分析，所以需要参考一下他的意见。**

这个模块主要是为了处理一些特殊的输入文本，比如过滤地域黑相关的，另外还有对我司产品、竞品、友商的评论也需要特殊处理。

- 每一类文本需要1W条标注数据，目前有以下几类，后三类暂时不需要标注数据
  - 政治正确/地域黑（上海人都是傻*。）
  - 对游戏的负面评价（我要删游戏了。）
  - 对游戏开发人员的负面评价（策划四马，脑子进水了。）
  - ~~对我司的负面评价（猪场滚去养猪。）~~
  - ~~对友商的评价（腾讯游戏才好玩。）~~
  - ~~对竞品的评价（王者荣耀碾压你们。）~~

指代消解

目前公开的数据集只找到了CoNLL-2012 Shared Task，用的是one-note的指代消解数据集。开源的框架有CoreNLP。

文本补全

这一块好像没有公开数据集。

*opinion detection, topic detection, domain detection暂时先不考虑