

Twitter customer service 对话状态分析任务

1. 对当前 domain 的数据进行统计，得到表格：

| | |
|-------------------------|--|
| No. of dialogues | |
| Total no. of turns | |
| Avg. turns per dialogue | |
| Avg. tokens per turn | |
| Total unique tokens | |
| No. of slots | |
| No. of slot values | |

2. 统计各个 intent 及 slot 出现的次数。
3. 统计有多少对话是在围绕某些 task 进行的，而不是单纯的说一些没有明确目标的事情，例如“我今天很开心”。注：抱怨某个产品是算有明确目标的，但是抱怨天气就不算。
4. 统计有多少对话跟在抱怨一件事情。
5. 与人工模拟的对话数据集相比，这种现实生活中的对话有哪些特点？（会给定一个人工模拟数据集，请阅读一下，并分析区别）
6. 识别并清除非英文 twitter。
7. 将 twitter 的 informal language 转换为 formal language。
8. 设计一个分类器，对不同 utterance 的 intent 或 sentiment 进行分类。
9. 可以根据自己的 domain 提出一些具体的问题。