# Maluuba dataset

链接地址：https://datasets.maluuba.com/Frames/dl

该数据集基于假期预定的场景——具体来说，查找航班和宾馆。Maluuba让两个人在聊天室中对话并收集了这些数据。一个人扮演用户，另一个人充当计算机。用户试图查找特价机票，另一个充当聊天机器人的人使用数据库检索信息。

对于该数据集的意图识别更加困难的，用户经常改变谈话主题。你可能同时讨论去滑铁卢、蒙特利尔、多伦多的计划。Maluuba发布的数据集侧重于进行同时涉及多个主题的对话。

该数据集包含1369个有关旅行规划的多轮对话，共有19986个行对话，每个多轮对话约14.6行对话。

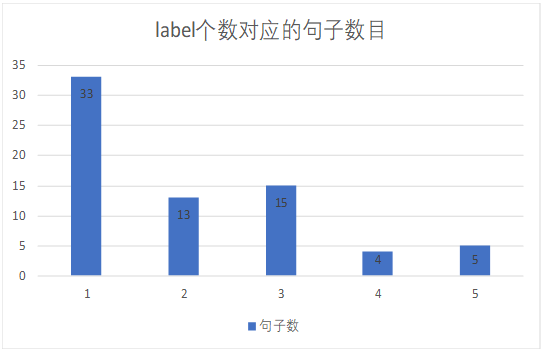
通过查看一些数据集的意图标注，如：

句子：tell me about the weather

意图：weather-query（Siri or Cortana）

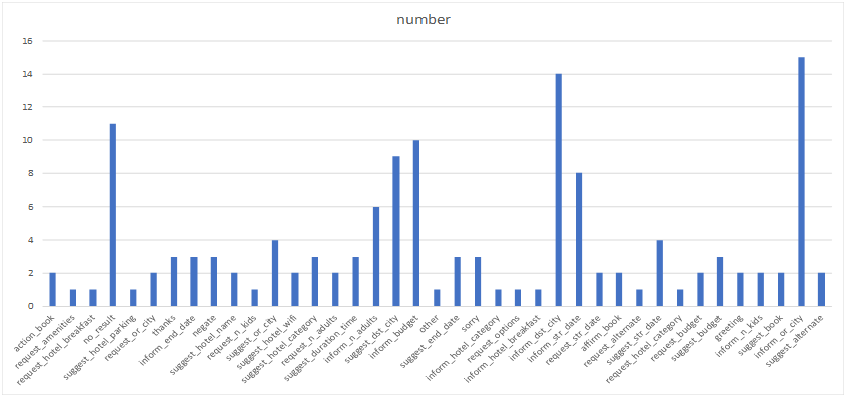
通过人工标注的50句对话，5个对话片段，其分析如下：

每个句子的label个数如下：



Label的分布如下：

|  |  |
| --- | --- |
| label | number |
| action\_book | 2 |
| request\_amenities | 1 |
| request\_hotel\_breakfast | 1 |
| no\_result | 11 |
| suggest\_hotel\_parking | 1 |
| request\_or\_city | 2 |
| thanks | 3 |
| inform\_end\_date | 3 |
| negate | 3 |
| suggest\_hotel\_name | 2 |
| request\_n\_kids | 1 |
| suggest\_or\_city | 4 |
| suggest\_hotel\_wifi | 2 |
| suggest\_hotel\_category | 3 |
| request\_n\_adults | 2 |
| suggest\_duration\_time | 3 |
| inform\_n\_adults | 6 |
| suggest\_dst\_city | 9 |
| inform\_budget | 10 |
| other | 1 |
| suggest\_end\_date | 3 |
| sorry | 3 |
| inform\_hotel\_category | 1 |
| request\_options | 1 |
| inform\_hotel\_breakfast | 1 |
| inform\_dst\_city | 14 |
| inform\_str\_date | 8 |
| request\_str\_date | 2 |
| affirm\_book | 2 |
| request\_alternate | 1 |
| suggest\_str\_date | 4 |
| request\_hotel\_category | 1 |
| request\_budget | 2 |
| suggest\_budget | 3 |
| greeting | 2 |
| inform\_n\_kids | 2 |
| suggest\_book | 2 |
| inform\_or\_city | 15 |
| suggest\_alternate | 2 |



举个例子：

User:Hi I'd like to go to Caprica from Busan, between Sunday August 21, 2016 and Wednesday August 31, 2016

意图：inform\_dst\_city（提出目的地） inform\_or\_city（提出出发地） inform\_str\_date（提出出发时间） inform\_end\_date（提出结束时间）

Wizard：And what would be your maximum budget for this trip?

意图：request\_budget（询问预算）

User:Actually it's unlimited for this trip

意图：inform\_budget（提出预算）

总结：1.句子的意图不仅仅与本句话相关，还与上文有关，必须结合上文分析，才能准确理解该句的意图。2.数据中，无关句子较少，基本为打招呼和感谢。3.文中部分标签带有依赖性，比如inform\_dst\_city(目的地)和inform\_or\_city(出发地)相互关联，inform\_str\_date(开始时间)和inform\_end\_date(结束时间）相互关联。4.label的分布不均衡，有些label的样本多，有些label的样本少。