本周任务：

1. 写毕业论文

毕业论文写的还比较少，才写了一章多的内容。

1. NLPCC论文准备材料搜集

目前搜集的对话标准：连贯性，出现说话者与听话者的角色互换

微博话语角色：发话者，受话者，评论者、围观者

不能算作话轮的：

1. 没有会话含义（类似到此一游） 2、转发类 3、水贴广告类

因为有篇论文提到会话提取准确率，在其参考文献中看到一篇微博对话提取的论文 ——2013年EMNLP的会议论文 中科大和华为李航发的

[A Dataset for Research on Short-Text Conversation](https://www.researchgate.net/publication/287781150_A_dataset_for_research_on_short-text_conversation)  正在看，还没看完。

1. 对话提取的是 微博——一条评论 这样形成一对一对的对话
2. 爬取NLP和ML的研究机构或研究者等的微博，做部分预处理（去除10个字以下的微博和5个字以下的评论，去除潜在的广告，取前100条评论）
3. 方法是：对每条微博，分别用三个检索模型各自得到10条，整合为小于等于30条的评论。然后根据以下三个标准，进行标记

**给出的三个标准**：（每个标准给的有例子）

语义相关 Semantic Relevance 与微博语义相关 实体关联：与微博的实体一致（不绝对）

逻辑一致性Logic Consistency

语言行为对齐 Speech Act Alignment 微博要求回答或转发（response 回答或转发）

采用机器学习方法，标记少量数据训练检索模型，大量未标记的数据（1百万条）

还在看它的检索模型。