Cmy:就是我们上课的时候说的那个，我们是做聊天辅助app。

Yc：啊，聊天辅助app

Cmy：您不是说让我们做两个么，最好可以的话。一个是识别上下文，还有一个是用爬虫爬数据。第一点我们想的是。因为我们一开始想的是，不是给它直接反馈一句话就是想给用户一个

Yc：稍微等一下哈

你说

Cmy：上课的时候您不是提意见说，能识别它（聊天内容）的上下文嘛，要考虑它的语境，还有一个就是爬虫嘛。然后我们就是想，第一个的话。因为我们是想给用户一个提示。就是给他一个话题，不是说给他一句话。就是主要还是想让他手动干预，让他自己打（字），我们只是提供一个话题、一个提示。所以这样的话我觉得好像就是，也不用上下文去。

Yc：那问题就在这儿。你给他一个提示，他聊着聊着，你还给他吗？他聊的过程中还给他再提示吗？

Sbq：除非他在问。

Yc：他在问？在你的系统里面问？是不是啊？在你的系统怎么问的？

Cmy：emmm我们就是……

Yc：这不是一样的吗？这就好像你在跟他对话。就很简单了。我今天要跟女朋友聊，你有什么建议，那你给他话题咯。那过一会儿，聊不下去了，有没有更多的。你再给他提供一下啦。你是不是也要把前面那句分析一下的。你说对不对了。否则你怎么和他讲，然后你跟他说，跨度很大的，一会儿聊电影，一会儿聊那个。他又不分析你前面的，结果发现。他肯定要知道你，首先你要，他经过这个，几句对话，他要判断。你是哪个。所以说，从实现的角度来说啊，举个例子。你肯定要搞一个类似于树状的结构的er图（）。你自己，你们要能够有一个判断啊。比如说，我最开始对这个人一点都不了解。他要跟我对话，我首先要知道，这个理论，那么你可能会问到啊，我想要跟女朋友聊，那你就知道了，他要跟女朋友聊。一个分类。然后你得问问他，你是男的还是女的。好了，男的跟女的聊天，女的跟女的聊也都不一样的啊。也都叫里面的啊

Sbq：这个用之前就定好了。

Yc：就已经选好了？那不管怎么样你前面肯定有个导航一样的东西。知道他的身份，而且知道两个聊的对象的基本信息，对不对？是不是？

G11：嗯嗯嗯

Yc：这个实际上是个分类信息。对不对？然后你只对搞了一个很大的东西，或者说搞了一个数据库一样的东西。然后来根据这些分类信息来找到那个合适他聊的那个话题。对不对？是这个意思吧？

G11：对

Sbq：我们想的

Yc：那你想的是啥？你得跟我说啊，我是这样理解的

Lsw：一开始我们设想的时候是这样的。跟对方聊天，比如客户还是什么的时候，聊天聊着聊着突然话题继续不下去了。然后我们就是通过这个软件，把它打开，然后输入那个。我们自己输入一些信息。比如客户的一些个人信息，或者是我们上文聊天中的某些关键词。我们自己手动输进去。然后它那个把搜索的结果拿出来，然后我们看到合适的信息了，我们就把它复制上去，或者我们自己重新组织一下语言打一遍再发出去，变成一个新的话题。

Yc：就是你在你的那个系统里面已经有，构造了很多话题以及聊法还是什么呢？

Lsw：不是不是，这个是我们搜索出来的，就是，额

Yc：是，已经，已经按照特定的关键字把它搜索出来的，实时的还是事先已经搜好的？

G11：实时的

Yc：实时的去搜？搜完了以后你还要把它处理好，找到合适他（的），是这个意思吧？

Lsw：嗯嗯

Yc：app上吗

Lsw：嗯对

Yc：你实时搜app上怎么搜？调用百度去搜吗？

Lsw：就，目前我们想的是爬虫去爬

Yc：爬虫。不是搜诶。爬虫是把整个网站爬下来。搜是精确找到哪个位置上有我想要的东西。你爬的话，你把所有东西，不管它合不合适的东西全爬下来了。

Llw：所以说，本来应该是用爬虫先把他微博上的东西爬下来，然后对这个人大概的分析他是需要，他是怎么样的人，或者他对什么感兴趣，然后再去搜这些信息，然后再搜出来反馈给用户。

Yc：嗷，你是这样想。按照你现在这样想，你得告诉这个软件，我要聊的这个人的微博在什么地方，然后这个软件就把这个人的微博撸一遍，从头到尾，然后分析一下。然后大概，大概归归类这个人大概是个什么样的，喜欢什么样的。然后建议你大概……他关注的主题是这些。是这个意思是不是？

G11：嗯嗯嗯

Yc：（思考ing）这是个分类。互联网信息分类。实际上本质是个分类。所以说最关键的算法是那个，还不是爬，就是爬来的那个信息进行处理。就是归类嘛。你要归出来，你这些类要事先是已经分好的。把它归到某一类上去。对不对？应该是这样子的。你不可能新产生出这样一个类。对不对？那个要求太高了，做不出来。这个要求也高。你知道做这个东西的话，工作量有2块。一个是你要做的app，这个还是简单的，因为这个app做的东西，它，反正输入的和你后台系统连的是吧，那你后台系统的关键是你拿到app要马上去爬。爬来了以后要马上去进行分析，然后再归类。然后这个东西就是，不仅仅是要爬了，分析归类。它难就难在要实时地给他。这个不能让他等好久。

G11：对，就是这个问题。

Yc：所以，我个人感觉呢，你不要做那么实时的。就是说很简单，就是说你，你用预约机制。就是类似于预约机制。我给你的信息都不能实时去做。就是很简单，你期望要跟什么样的人的聊，那你给我一个（他的信息）。然后呢，我大概，比方说一天或者两天以后再给你信息。能做到这个就行，先做到这个再说。因为，其实，你就相当于把技术路线全部走通哇。你要实时的就是你要把它速度处理快。对不对嘛？所以我担心你，先把它走通再说。那个速度快要求很高，其实不仅仅依赖于你的软件还依赖于你的硬件。是不是啦，所以处理速度也要够快才行。所以先把这个东西搞搞好，我知道你要做什么。可以呀。那就是，其实有两块。一块是简单的app，输入（不是下载？？？）一个，然后我注册一下，然后我预约一个，我要聊的这个人大致是怎么样的，然后你把他……然后下次……

Llw：老师！那，那个，数据分类那一块您有没有什么好的建议什么的，数据分类那个有点懵。

Yc：先去爬虫爬来，我估计你们就找个现成的爬虫。

Llw：emmm那应该不是

Yc：自己写？（深思）这个，不太容易。爬来的话。因为它也涉及到专门的数据结构啊等等的……我建议你们可以借鉴。先借鉴。实在不行我允许你们去找个现成的爬虫按钮，就那个蚂蚁，去爬来。关键就是爬来的数据怎么分析。这个是……因为，实际上是，就算是微博它也是html文档。对不对啦？你要识别它的结构然后要分词然后筛选然后计算。这些也是有现成的。有了现成以后，关键是你怎么分类？这有分类算法的。我建议你们，你们准备用什么语言？

G11：python

Yc：诶，好。Python有很多的机器学习方法，库，现成的。我建议你们，你们是不是在学？

G11：嗯

Yc：我建议你们去学中国大学慕课上面的那个，北京理工大学的那个，嵩天那个老师的。嵩就是嵩山啊，一个山字一个高。其实你查Python都是这个人的。很多的，这个最有名的这个排在最前面。然后他这里面，然后。嗳，然后记牢，你查python关键字的话，嵩天下面除了python的基础以外还有python图像识别，python机器学习。然后你去看python机器学习里面就有python做分类算法。可以吧？我就给你们先说到此。可以，这个还是蛮有挑战。好，那我们先到这里。

G11：谢谢老师！

Yc：所以这个实现上面的话在可行性上面也要写，原来想的也要别的算法也要，比较一下。Ok，好了。