1. 基本情况描述

在行业版本中，可能面对的情况是，客户只提供了少量的qa列表。

而想要得到好的效果，就需要有大量的qa尽可能的来覆盖整个业务。

这个需求目前来看，很难通个客户编写qa或在使用过程中对qa的维护来完成。

特别是在实施初期，无法提升的效果会严重影响客户的使用兴趣。

其实这个需求的解决办法最有效的只有一个，让程序在网上或客户的资料中自动获得需要的qa。

这样就有几个问题要解决：

1. 爬取的qa未必是完全格式化的，如何解决清洗的问题。
2. 假设清洗完成，如何解决没有回答、或回答不正确的问题。
3. 假设每一个question都有answer，但是有q重复的或意思相近的。如何找到相近的q，解决聚类的问题。
4. 假设聚类的问题解决，但这个聚类是以某种规则或机器学习的方法完成的，其聚类的结果未必和业务相关。那么如何解决answer与业务相关的分类问题。

将Q聚类

按A分类

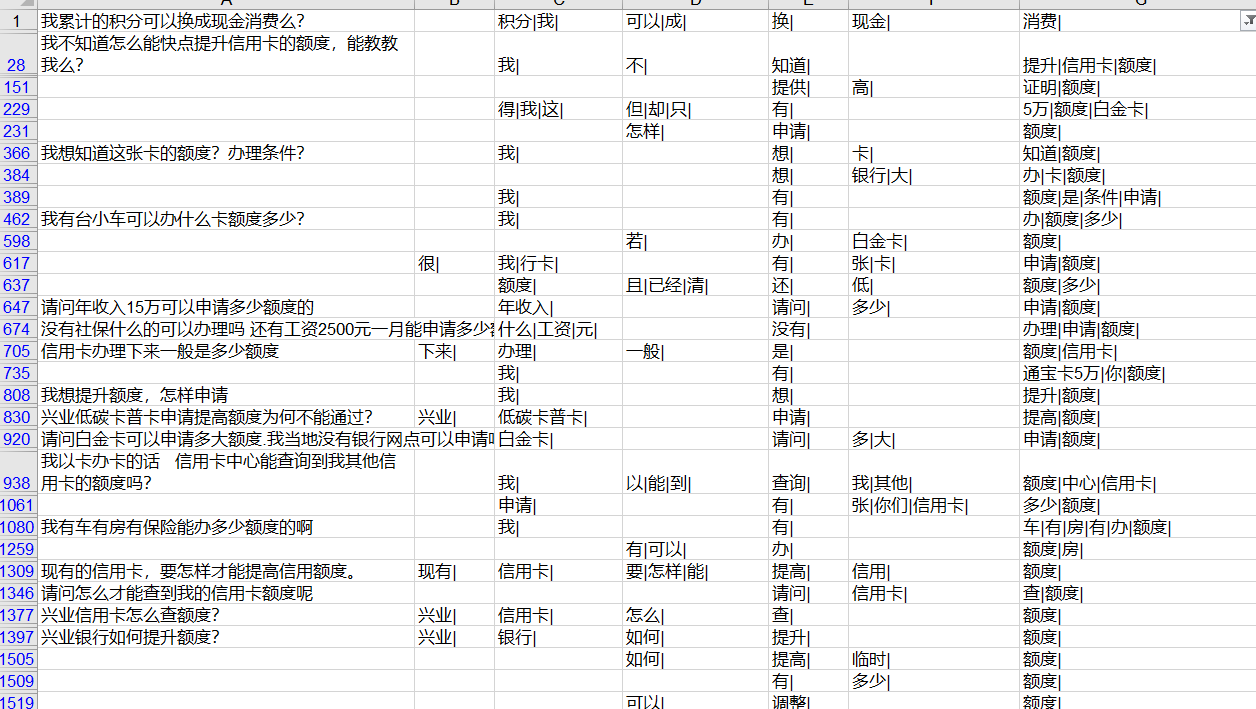
如何清洗

如何得到准确的answer

这一系列的问题中，可能暂时不需要解决的问题有：“如何清洗”，“如何得到准确的answer”。因为answer是有限集合，如何有不太干净的question，似乎也不太要紧。

需要有限解决的问题是:“将Q聚类”，可以认为Q的聚类就是一种清洗，聚的效果越好，后续的分类，可能花点人工都可以搞定。当然效果差点也可以，延伸的工作可以对answer再进行一次聚类，可以认为是对聚类的迭代。

下图是对Q的ltp分析效果



数据供参考，对execl不太会用，有点乱。

下图是对A的ltp分析效果

**参见文件**

需要近似完美解决的问题是：“按A分类”，因为这个是直接面对客户的，直接产品化的部分。主要有以下几点：

用户answer的编写、组织

如何发布

日志标注

运行期间的调整

1. Question聚类的解决方案
2. 使用ltp对question进行句法分析
3. 以主谓宾（目前的观测结果是，谓宾对句子的意思描述更加具体）的形式把对应的词集合取出。
4. 找一个算法进行聚类（应该要用到W2V）
5. Answer分类的解决方案