#### 实体提取工具实现方案

实体提取工具是一个用来管理实体提取过程的一个工具，包括准备语料，分词，词语标注，模型训练等过程。

1. 类定义

(1) 定义汽车行业的实体类型和字符，例如定义一个实体类型为汽车，字符为car。

(2) 构建类和实体的关系词典。

词典格式：实体 实体类型 （注：实体和实体类型之间是一个空格）

例如：上汽大通D90 car

轮胎 part

1. 汽车行业大词典

构建汽车行业的大词典用来做分词用，构建完成后加入到jieba的词典里。

词典来源有两种，一种是在网上找一些抓也词表，另一种来源是BOM表。

格式：词语 词频 （注：词语和词频之间是一个空格）

例如：汽车 3

轮胎 3

1. 实体抽取工具
2. 语料管理界面：语料的导入，可以支持语料的批量导入，并将语料保存到MySQL数据库。
3. 类定义界面：用来自定义实体类型的界面，还有类的实体的自定义界面。
4. 标注界面：标注界面分为两个部分，一部分是自动标注，一部分是人工标注，对于批量的语料来说先自动标注一遍，然后再人工校验。包括待处理语料，标注界面，已处理语料。整体流程如下：

自动标注

数据库

待处理语料

人工校验

保存语料

已处理语料

（和python训练程序的接口格式待定）

1. 训练界面：用来训练模型的界面，训练已标记好的语料，得到实体提取的模型并输出模型的一些参数，包括模型保存位置，训练模型所需环境和所用工具版本信息。