1.对约300万条发票明细数据进行去重，提取出24992条发票明细，见《train2.1.data》。

2.对24992条明细进行分词、词性标注，提取词性为名词、动词、介词、成语、缩写、形容词、名词修饰的词语并删除无效数据，共提取出23794条发票明细，见《train2.1.data.out》。

注：词性标注后少于24992条数据是因为有些明细分词后的名词个数为0，已被去除。

3.对23794条已拆分、标注的发票明细按行去重【注1】，去重后的数量为15964，再按名词个数进行统计（见train2.1.data.out.line\_unique）：

单名词：2056 占比：12.879%

双名词：3932【注2】 占比：24.631%

三名词：3227 占比：20.214%

四名词及以上：6749 占比：42.276%

总数：15964【注3】

【注1】：按行去重指的是词性标注后的去重，比如华为荣耀V8，华为荣耀V10（2018版），在词性标注后，可能均被标注为“华为”“荣耀”。去重时，两个明细算成1条。

【注2】：彩虹太空杯拆分为： 彩虹 太空杯，拆分后为两个词语，为双名词。单名词则只有一个词语，其他类推。

【注3】：将发票明细拆分、词性标注后，单名词明细占比12.187%，多名词占比87.813%。

**4. 训练集分类1（规则视角）**

货物、劳务、服务、无形资产的一般性描述为：

[某特质/特性][的]货物

或者

货物[特质/特性]

其中，特质/特性可以为型号、规格、材料、材质、版本、修饰词；货物为主名词。特质/特性基本上也是名词，会干扰货物（主名词）的识别。

特质/特性与主名词的组词顺序，没有强制性的约束/规则。比如：

□ 易记账V11.0(带安全锁) === 带安全锁的易记账V11.0

□ 铂光金 荣耀V8 === 荣耀V8 铂光金

**5. 训练集分类2（语句成分视角）**

□ 未登录词

▲ 书名（如：《中国上下五千年》）

▲ 品牌名词（如：金蝶云之家）

▲ 食物名词（如：江小白（酒））

□ 单名词

▲ 可直接搜索（如：水果）

▲ 不可直接搜索（如：马夹）

★ 直意

如： 马夹上位词是服装，而且马夹指的就是服装，而非其他

★ 非直意

如： 汽车，可能指的是玩具汽车

烘焙，指的是面包

□ 多名词

▲ 主名词识别

如：

饼干（胡萝卜味）、胡萝卜味饼干的主名词是饼干

烘焙大礼包 的主名词是烘焙

**6.训练集分析（双名词）**

多名词训练数据占比较多，达87.813%，双名词作为主体名词识别的基础，对双名词的主体名词识别具有重要的意义。

主体名词和修饰性名词的格式有：

n + N （如：战神 鼠标垫、四叶草 耳线）

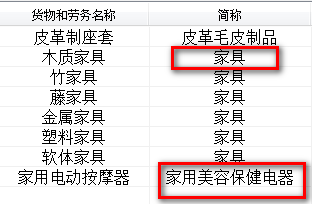
N + n （如：烘焙 大礼包、硬金生肖 狗）

注：n 修饰性名词 N主体名词

**7.训练集分析（三名词）**

物品的修饰词会影响到物品的分类，比如沙发，可以为一般的家具，也可以为税收分类编码中的家用美容保健电器。

如：电动 按摩 沙发，如果去掉修饰性名词“电动”，则可以被分类为“家具”，如果带上“电动”修饰词，则应分类为“家用美容保健电器”。



如：足金 微笑 手链，属珠宝首饰，如果去掉足金，也可分类为小型起重设备。

