1.对约300万条发票明细数据进行去重，提取出25754条发票明细，见《train2.data》。

2.对25754进行分词、词性标注，提取词性为名词的词语并删除无效数据，共提取出24036条发票明细，见《train2.data.out》。

注：词性标注后少于25754条数据是因为有些明细分词后的名词个数为0，明细可能是测试数据，也可能是生产数据。

3.对24036条发票明细按行去重，再按名词个数进行统计（见train2.data.out.line\_unique）：

单名词：1928 占比：12.187%

双名词：4052 占比：25.613%

三名词：3654 占比：23.097%

四名词及以上：6186 占比：39.103%

总数：15820

注：将发票明细拆分、词性标注后，单词名明细占比12.187%，多名词占比87.813%。

4.