**税务领域本体在税务文本中的应用**

**自然语言处理**

文本分类是自然语言处理的一项主要应用。良好的文本分类能够极大的提高文本信息检索的效率。现在流行的自然语言处理工具有：Stanford University 基于Java开发的一组自然语言处理工具包；Python语言开发的NLTK(Natural Language Toolkit)，最新的版本是3.2.1，已经支持Stanford Word Segmenter，完全支持Unicode编码；还有针对中文的自然语言处理工具包SnowNLP。这些自然语言处理工具都有丰富的文本分类功能。图7.4是一文本分类的框架流程。

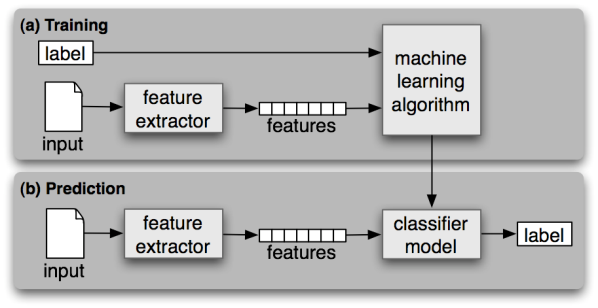


Fig. 7.1 Flowchart of Text Classifier

图 6.1文本分类的流程[102]

由于中文并不像英文那样用空格把每一个单词分开，而是由阅读者根据语义自己分开每个词语。而多数自然语言处理软件都是在分词的基础上处理分析的。英语文本由一个一个的单词组成，自然的就分词了。中文就需要先分词，然后再做其他的分类、检索、统计、特征提取等处理。现在比较流行的中文分词软件工具包有JIEBA, 中科院的NLPIR/ICTCLAS2016分词, Stanford Word　Segmenter，清华大学的THULAC[103]。SnowNLP中文处理工具包内置中文分词工具，使用Python语言[104]。图8.2是使用SnowNLP工具包中文分词的一个示例。

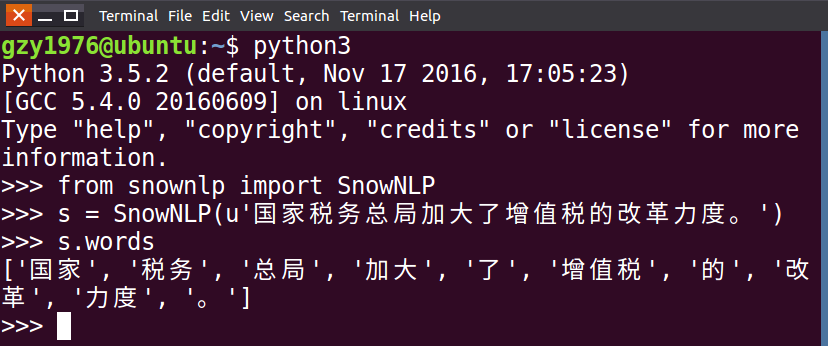


图 6.2 SnowNLP工具包中文分词

Fig. 7.2 SnowNLP Toolkit

前面提到的自然语言处理工具都提供了机器学习的功能，能够进行分词、词性标注、文本分类、提取关键词、提取摘要、文本相似度分析；支持N-gram, HMM, Naïve Bayes, TextRank等算法。这些工具中NLTK提供了丰富的文本分类算法外，还支持scikit-learn机器学习算法包[105]。本文主要使用NLTK研究机器学习和领域知识本体的文本分类方法。

**机器学习**

Scikit-learn[106]

**本体文本分类**

文本分类在NLTK中的具体实现流程见图7.6。从输入中用特征提取器提取文本特征，如：统计的词频、语法特征、关键字等， 通过机器学习算法训练出分类模型，通过这一模型对需要分类的文本进行分类。这种方法存在一些问题，例如：有些词汇出现的频率不高，但是非常重要，如每个案例中会出现“合计”，“共计”等表示汇总的词语，一般也就出现一次，但是这些词语后面紧跟的数字是整个案例的合计数。然而合计数在税务评估案例的分类和查询中非常重要。统计的词频、语法特征、关键字等特征提取的方法没有考虑到文本的语义。应用本体则可以有效在设计特征提取器时，加入对语义的分析，以提高文本分类的准确率。

本文提出的算法是使用Protégé 构建的领域本体，通过Jena读取需要分类的文本，并将领域本体为基础处理、推理、分析作为NLTK分类器的特征提取器。具体的实现如图8.3所示。

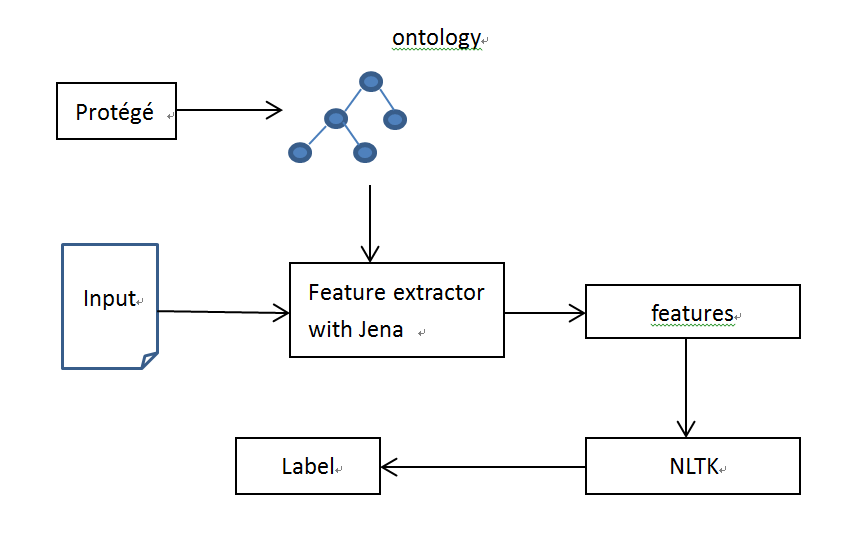


图 6.3

Fig. 7.3

**税务稽查报告分类**

税务稽查是税务机关依法对纳税人、扣缴义务人和其他税务当事人履行纳税义务、扣缴义务及税法规定的其他义务等情况进行税务检查和处理工作的行政执法行为。税务稽查是税务工作的重要环节，是其他环节的强大后盾。在加强税收征管、税收立法、提高社会纳税遵从度方面有着不可替代的作用。税务稽查选案是指按照一定的方式、方法和要求，对纳税人各项信息数据进行采集、分析、筛选，最终为税务稽查的实施确定具体检查对象的过程。 选案是税务稽查案件的来源，是税务稽查的第一道程序，选案准确的高低直接影响到稽查职能的发挥。现在的税务稽查部门普遍使用的选案方法是“分行业税负分析”，近年来形成了结合纳税额度的“唯税负论”。只是按照企业税负的高低或统计学意义上的异常来选案。这种选案方法虽然简单易用，但是在发现高价值目标、指导检查方面明显不能满足需要。而且有些企业根据税务局的这一特点，通过调整账面信息调高税负率来逃避检查。

**本章小结**