机器学习工程师纳米学位

毕业项目开题报告

项目题目

自然语言处理-文档归类

一、项目背景

自然语言指人类使用的语言，如汉语、英语等，而语言是思维的载体，是人际交流的工具。人类历史上以语言文字形式记载和流传的知识占到了知识总量的 80%以上。20 世纪，随着计算机的诞生和信息技术的迅速发展，如何让计算机实现人们希望的语言处理功能？如何让计算机实现海量语言信息的自动处理和有效利用？针对这两个问题的研究，计算语言学（Computational Linguistics）应运而生，而自然语言处理（natural language processing，NLP）则是其中的一个研究领域。

自然语言处理就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。自然语言处理研究的内容包括机器翻译、信息检索、自动文摘、文档归类、问答系统、信息过滤、语言教学、文字识别、文字编辑和自动校对等。在该项目中，研究的是文档归类（Document categorization），也叫文本自动分类或信息分类，其目的就是计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。解决该问题的关键是将文本转换为计算机能够识别和计算的形式，然后就可以用合适的模型来实现。文档归类可以应用到图书管理、内容管理、信息监控等领域。

二、问题描述

文档归类的关键是将文本转换为计算机能够识别和计算的形式，这样才能方便计算机发挥其强大的计算优势。因此我们的问题就是：文本应该用什么样的形式表示？才能有利于模型识别和计算。

三、输入数据

使用到的数据集为 20newsgroups。20newsgroups 数据集是用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约 20,000 左右的新闻组文档，均匀分为 20 个不同主题的新闻组集合。项目中使用 sklearn.datasets.fetch_20newsgroups 获取该数据集。该数据集包含新闻内容和新闻主题类别，对其进行词袋子模型表示后，对数据集进行分类模型的训练和评估。然后将该数据集利用 Word2Vec 方式即词向量模型表示，训练后进行简单评测。

项目中还用到了 text8 数据包。数据集通过 http://mattmahoney.net/dc/text8.zip 获取。text8 来源于 enwiki8，enwiki8 是从 wikipedia 上得到的前 100，000，000 个字符；而 text8 就是把这些字符当中各种奇怪的符号，非英文字符全都去掉，再把大写字符转化成小写字符，

把数字转化成对应的英语单词之后，得到的。将该数据集利用 Word2Vec 方式即词向量模型表示，训练后进行简单评测。

四、解决办法

使用词袋子模型表示文本。对输入数据的文本进行预处理，分词，得到单词的集合，建立词典，然后文本就可以用特征词的频率向量或者加权词频 TF-IDF 向量表示。在该模型下，每个词是独立存在的，没有上下文的关联性。20newsgroups 数据集包含了新闻内容和新闻主题类别，可以使用监督分类模型进行训练和评估。

使用 Word2Vec 方式即词向量模型表示文本。对输入数据的文本进行预处理，分词。然后使用 Word2Vec 模型进行训练。Word2Vec 的基本思想是把自然语言中的每一个词，表示成一个统一意义统一维度的短向量。而 Word2Vec 注重的是词与上下文的关联性。训练后的评估也是检测单词之间的相似性和关联性。

五、基准模型

词袋子模型采用的是 sklearn.feature_extraction.text 下的 CountVectorizer 来表示特征词的频率向量和 TfidfVectorizer 来表示加权词频 TF-IDF 向量。对于分类模型则是利用 sklearn 的 tree.DecisionTreeClassifier、svm.SVC、naive_bayes.MultinomialNB 进行文本分类的训练、预测和评估。

Word2Vec 模型采用的是 gensim.models 的 word2vec 来进行数据的训练和评估。

六、评估指标

词袋子模型的评估指标采用的是 accuracy_score。即模型根据预测的准确率的高低进行评估。

Word2Vec 模型的评估指标是对几条简单的单词进行相似度预测，例如'woman', 'man'等。然后利用 Google 公开的 20000 条左右的语法与语义化训练样本'questions-words.txt'进行最终评估，'questions-words.txt'中每一条遵循 A is to B as C is to D 这个格式。

七、设计大纲

获取到数据集后，首先观察和分析数据，对数据进行预处理。当输入数据较多时，会造成特征向量的维度变得很大。通过数据预处理，过滤掉一些无关紧要的信息，可以达到降维的目的。数据预处理包括去掉标点符号、单词大小统一成小写、对同一词不同形式（如单复数）进行统一、过滤掉停用词。在测试集上的操作效果如下：

状态	开始	转小写	停用词过滤	同一词统一
vocabulary length	138412	115822	115678	108970

数据预处理后分别使用 CountVectorizer 和 TfidfVectorizer 进行词袋子模型表示，然后利用 DecisionTreeClassifier、SVC、MultinomialNB 监督分类模型进行训练、预测和评估。之后挑选出合适的模型进行最终的模型调优，最后进行最终的评估。

将上述预处理的数据使用 Word2Vec 模型进行训练，然后进行评估。然后使用 text8 数据包使用 Word2Vec 模型进行训练，然后进行评估。对评估效果进行比较分析。

八、参考文献

【1】宗成庆，《自然语言理解》PPT，中科院自动化研究所模式识别国家重点实验室
【2】Steven Bird, Ewan Klein & Edward Loper，《PYTHON 自然语言处理》，O'REILLY
【3】网站“我爱自然语言处理”，CSDN、知乎、简书等一些网络相关资源