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。文档归类中的类别有很多，因此属于监督学习中的多分类问题。我们的问题是：数据预处理过程中，文本应该用什么样的形式表示？才能有利于分类模型训练和预测。

三、输入数据

主要数据集为 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，因此适用于监督分类模型。我们将该数据集分别用词袋子模型和 Word2Vec 方式即词向量模型表示，然后用监督分类模型进行训练和预测。

辅助数据集为 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。该数据集可用于 Word2Vec 方式即词向量模型的辅助训练。

四、解决办法

使用词袋子模型表示文本：数据预处理后，对文本建立词典，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。使用词袋子模型处理后的数据可以直接进行监督分类模型的训练。

使用 Word2Vec 方式即词向量模型表示文本：数据预处理后，使用 Word2Vec 模型进行词向量表示。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。然后进行词向量“相加取平均”处理，具体做法是，每条新闻由若干个词组成，每个词在 word2vec 中都有由一个长度为 N 的词向量表示，且这个词向量的位置是与词的语义相关联的。对于每一条新闻，将这条新闻中所有的词的词向量加和取平均，这样既能保留句子中所有单词的语义，又能生成一个蕴含着这句话的综合语义的“句向量”。这样就可以进行监督分类模型的训练。

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。

Word2Vec 模型采用的是 gensim.models 的 word2vec 来进行词向量的预处理。

监督分类模型则是利用 sklearn 的 tree.DecisionTreeClassifier、svm.SVC、naive_bayes.MultinomialNB 进行文本分类的训练、预测和评估。

六、评估指标

使用 log_loss 损失函数作为评估指标。损失函数越小，模型的鲁棒性就越好。

辅助评估指标是 accuracy_score。即模型根据预测的准确率的高低进行评估。

七、设计大纲

获取数据集；观察和分析数据集；对数据集进行预处理。我们的数据集还是比较多的，因此特征向量的维度比较大。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。在训练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

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。

使用 Word2Vec 模型进行词向量表示，进行简单评测，若效果不佳则词向量模型使用 text8 数据包进行辅助训练。然后对词向量“相加取平均”处理，根据解决办法中所描述的方法将其生成一个蕴含着这句话的综合语义的“句向

量”。利用 DecisionTreeClassifier、SVC、MultinomialNB 监督分类模型进行训练、预测和评估。

对比分析文本表示模型，每种分类模型的表现，然后挑选出最优模型进行调优，并进行最终预测。

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