基于word2vec的专利文本自动分类研究

汇报人：

2024-01-13

目录

01

引言

研究背景与意义

专利文本数量激增

随着科技创新的加速，专利文本数量呈现爆炸式增长，手动分类和管理已无法满足需求。

自动分类的重要性

自动分类能够提高专利审查效率，降低人力成本，并为企业和科研机构提供有价值的专利信息。

Word2Vec在文本处理中的应用

Word2Vec是一种高效的词向量生成技术，能够捕捉文本中的语义信息，为专利文本自动分类提供了新的解决方案。

国内外研究现状及发展趋势

国内研究现状

国内在专利文本自动分类方面已有一定研究基础，但大多采用传统的机器学习方法，分类效果有待提高。

国外研究现状

国外在专利文本自动分类方面研究较为深入，采用了深度学习等先进技术，取得了一定成果。

发展趋势

随着深度学习技术的不断发展，结合Word2Vec等词向量生成技术的专利文本自动分类方法将成为未来研究的重要方向。

研究内容、目的和方法

研究内容

本研究旨在探讨基于Word2Vec的专利文本自动分类方法，包括数据预处理、特征提取、分类器设计等关键步骤。

研究目的

通过本研究，期望提高专利文本自动分类的准确性和效率，为企业和科研机构提供更加便捷、准确的专利信息服务。

研究方法

本研究将采用文献调研、实验研究和对比分析等方法，对基于Word2Vec的专利文本自动分类方法进行深入研究。

02

word2vec模型原理及关键技术

word2vec模型概述

要点一

word2vec是一种基于神经网络的词向量表示学习模型

它能够将词语转化为高维向量，进而在向量空间中表示词语的语义信息。

要点二

word2vec模型的应用

它广泛应用于自然语言处理领域，如文本分类、情感分析、机器翻译等。

word2vec模型原理

基于语料库的训练

word2vec模型通过在大规模语料库上进行训练，学习词语的上下文信息，进而生成词向量。

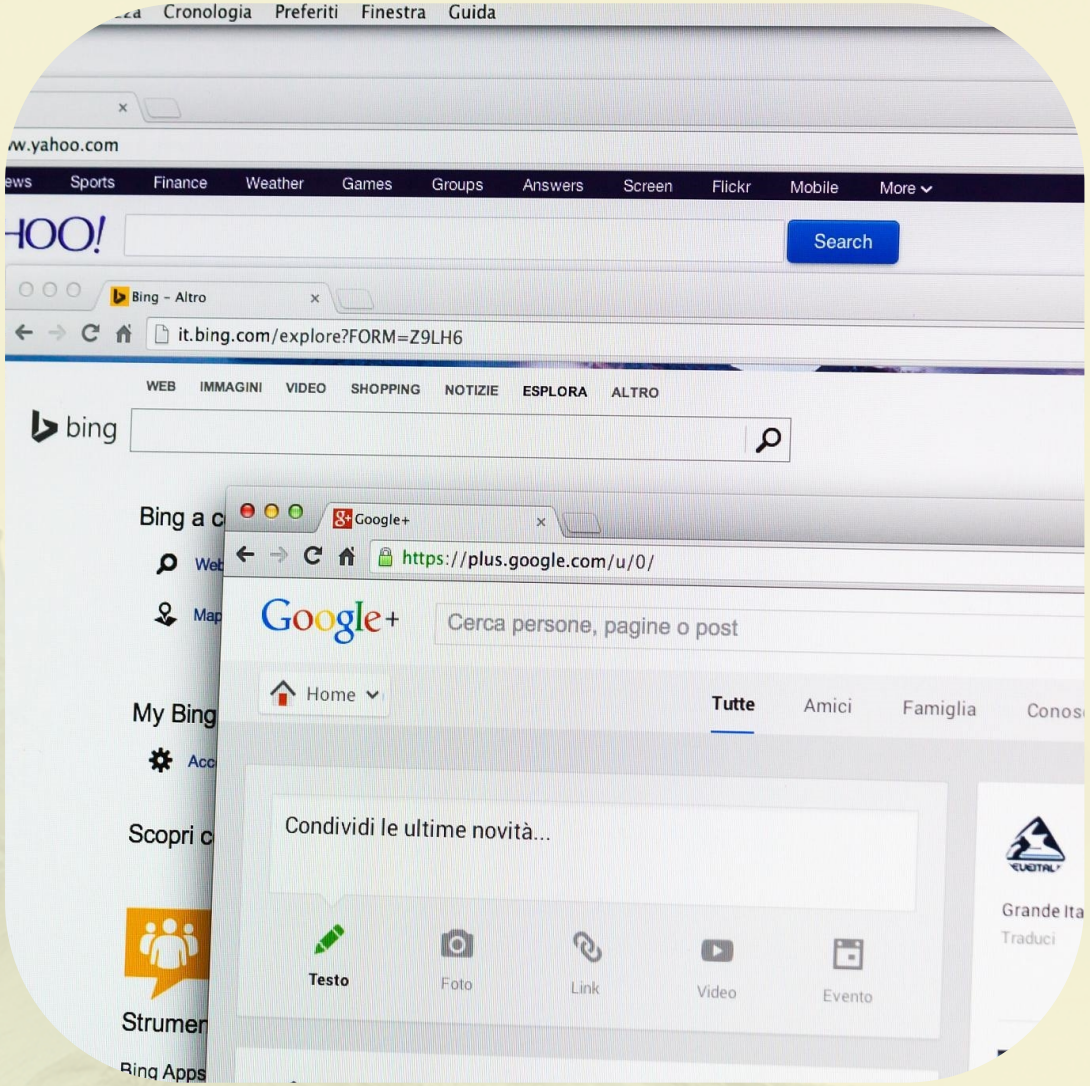
两种主要模型

word2vec包含Skip-gram和Continuous Bag of Words（CBOW）两种主要模型，Skip-gram模型通过当前词预测上下文词，而CBOW模型则通过上下文词预测当前词。

神经网络结构

word2vec模型采用简单的三层神经网络结构，包括输入层、隐藏层和输出层。

关键技术分析



层次softmax

为了提高模型的训练效率，word2vec采用了层次softmax技术，将复杂的归一化问题转化为一系列二分类问题。

负采样

负采样是另一种提高训练效率的技术，它通过随机采样一定数量的负样本，与正样本一起进行训练，从而加速模型的收敛。

参数调优

在word2vec模型的训练过程中，需要进行一系列的参数调优，如学习率、向量维度、窗口大小等，以获得更好的词向量表示效果。

03

专利文本数据预处理

专利文本数据来源及特点

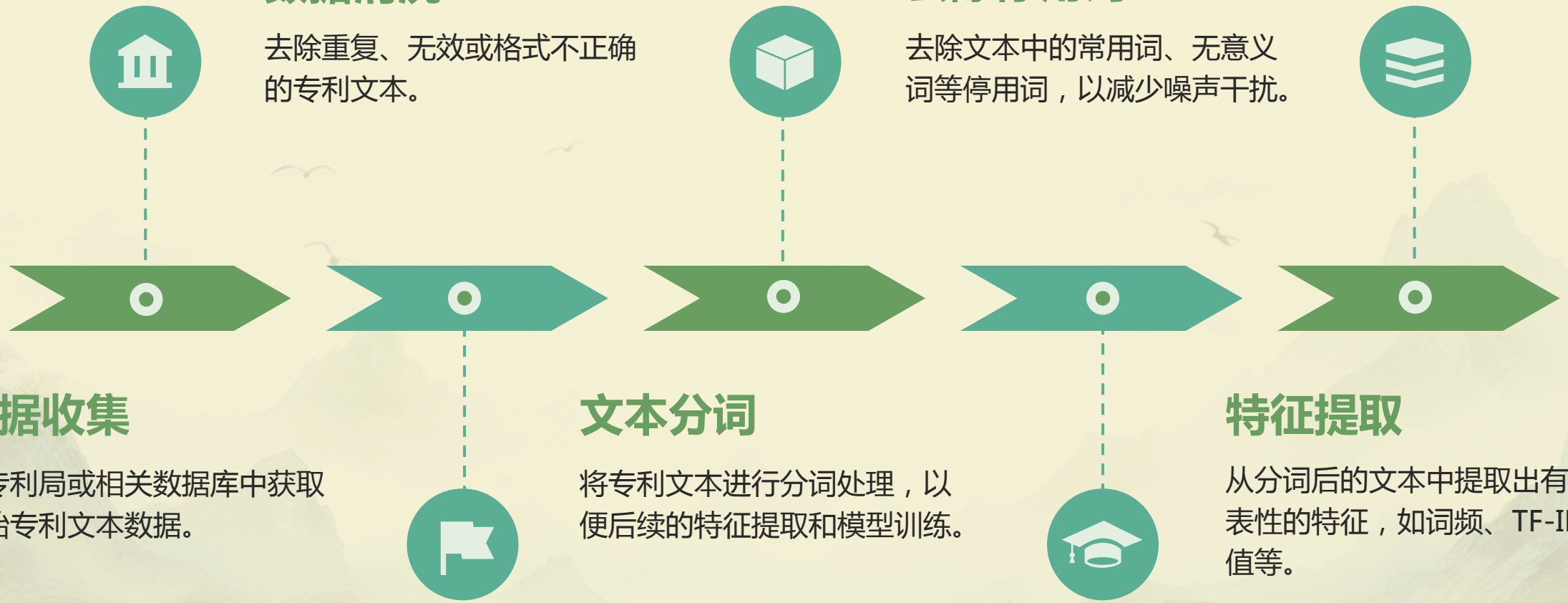
数据来源

专利文本数据通常来自于各个国家的专利局或国际专利组织，如美国专利商标局（USPTO）、欧洲专利局（EPO）和世界知识产权组织（WIPO）等。

文本特点

专利文本具有专业性强、结构固定、用语规范等特点。一般包括标题、摘要、权利要求书、说明书等部分，其中含有大量的技术术语和领域知识。

数据预处理流程



数据收集

从专利局或相关数据库中获取原始专利文本数据。

数据清洗

去除重复、无效或格式不正确的专利文本。

文本分词

将专利文本进行分词处理，以便后续的特征提取和模型训练。

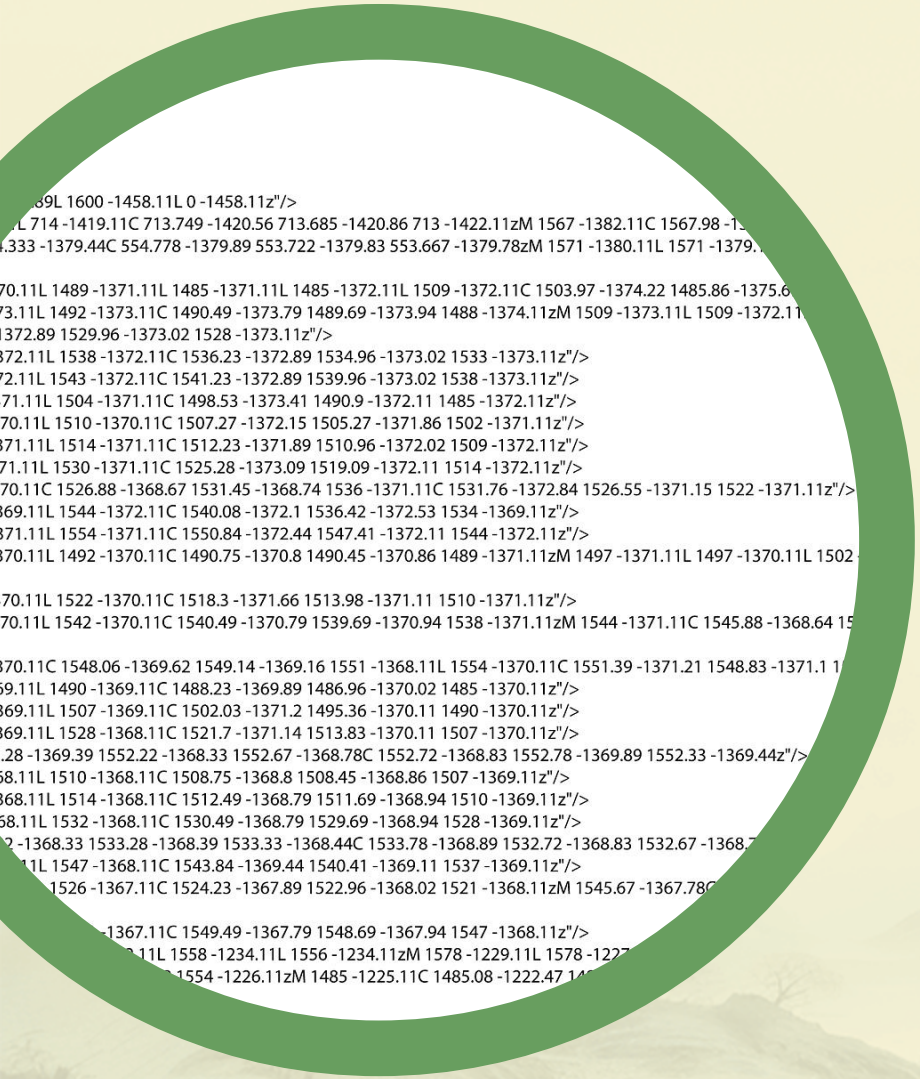
去除停用词

去除文本中的常用词、无意义词等停用词，以减少噪声干扰。

特征提取

从分词后的文本中提取出有代表性的特征，如词频、TF-IDF值等。

预处理结果展示



01

分词结果

展示经过分词处理后的专利文本，可以看到文本被切分成了一个个独立的词语。

02

去除停用词结果

展示去除停用词后的文本，可以看到文本中的常用词和无意义词被去除，文本更加干净。

03

特征提取结果

展示提取出的特征词及其对应的权重或重要性得分，这些特征将用于后续的模型训练和分类任务。

04

基于word2vec的专利文本特征提取