NLPartners-PDF文档关键信息自动高亮

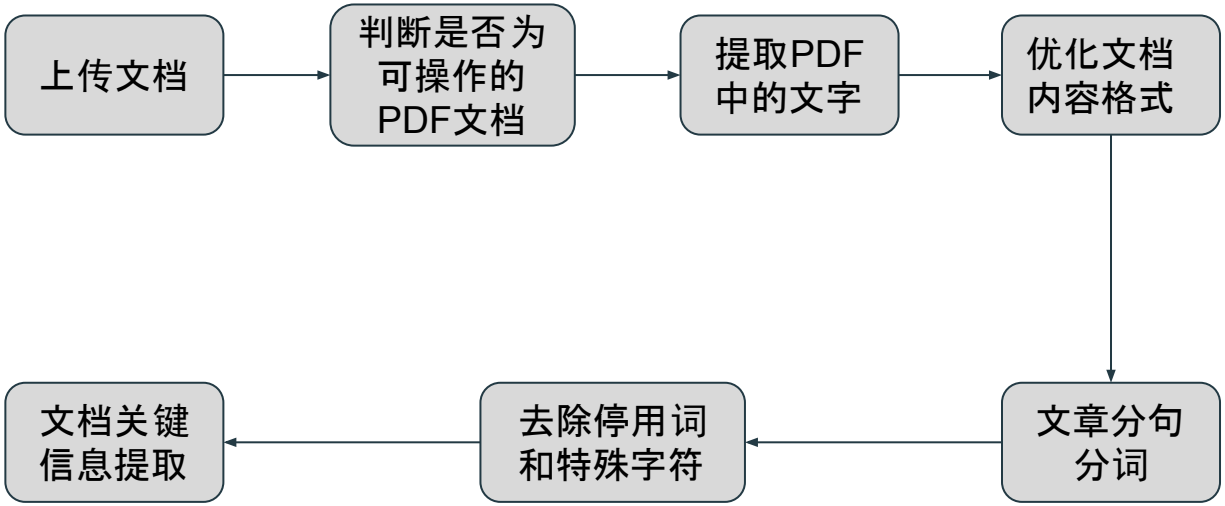
----2020.05.27

目录

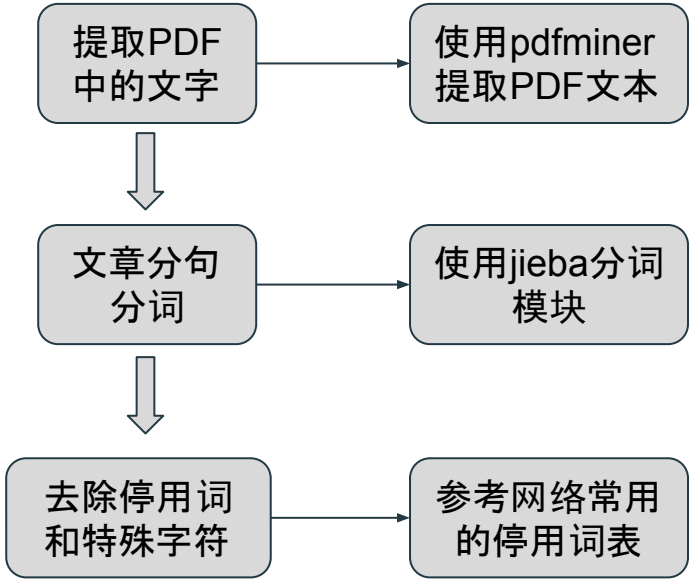
项目介绍

本项目的主要目的是将上传的PDF进行解析，然后对文档中的关键信息进行高亮显示。上传了文档后，会对PDF进行判断，判断是否是可解析的文档，当是可解析的文档时，后台将对文档的内容进行抽取，然后提取文中的关键信息，最后将这些关键信息高亮之后返回到页面显示，同时页面可操作将高亮的文档保存到本地。

文档处理


上传文档
判断是否为可操作的PDF文档
提取PDF中的文字
优化文档内容格式
文章分句分词
去除停用词和特殊字符
文档关键信息提取

文档处理



PDFMiner 是一个PDF文档文本提取工具

jieba是一款常用的中文分词组件

<哈工大停用词表>和<百度停用词表>

关键信息提取

TextRank算法是一种抽取式的无监督的文本摘要方法，其由PageRank而来，最终有公式

$$S(v_i) = (1-d) + d \sum_{(j,i)\in\varepsilon} \frac{w_{ji}}{\sum_{v_k \in out(v_j)} w_{jk}} S(v_j)$$

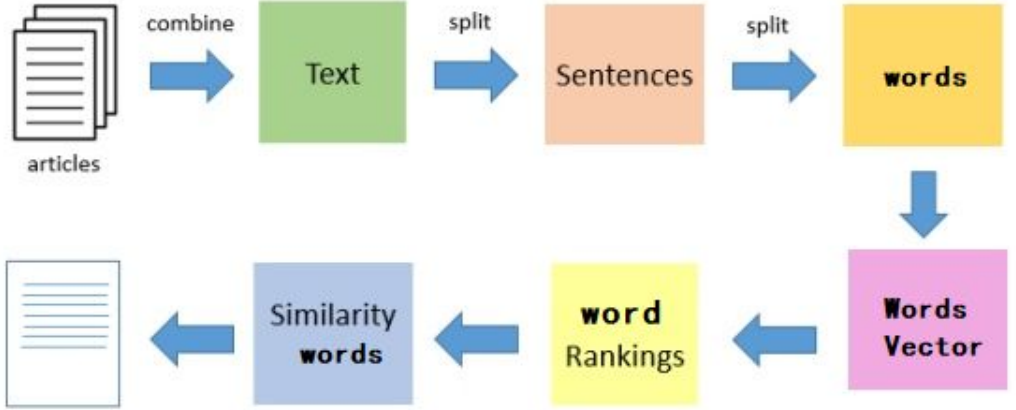
用S(vi)定义vi这个词的TextRank值; 用 ∈来表示所有其他可以链接到v□这个词的集合那么便可以用*(j,i)*来定义集合中的其中一个（由词 v□链接到词vi）

TextRank中一个单词i的权重取决于与在*i*前面的各个点*j*组成的 *(j,i)*这条边的权重，以及*j*这个点到其他其他边的权重之和。

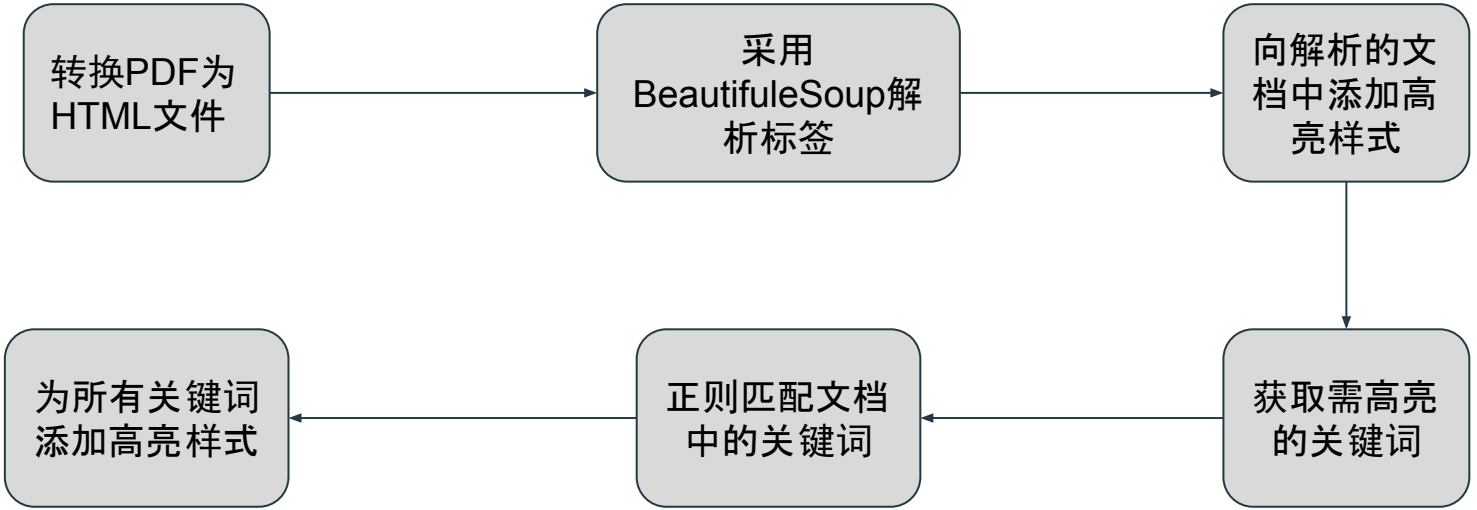
关键信息提取

文档信息提取的流程大致如下：

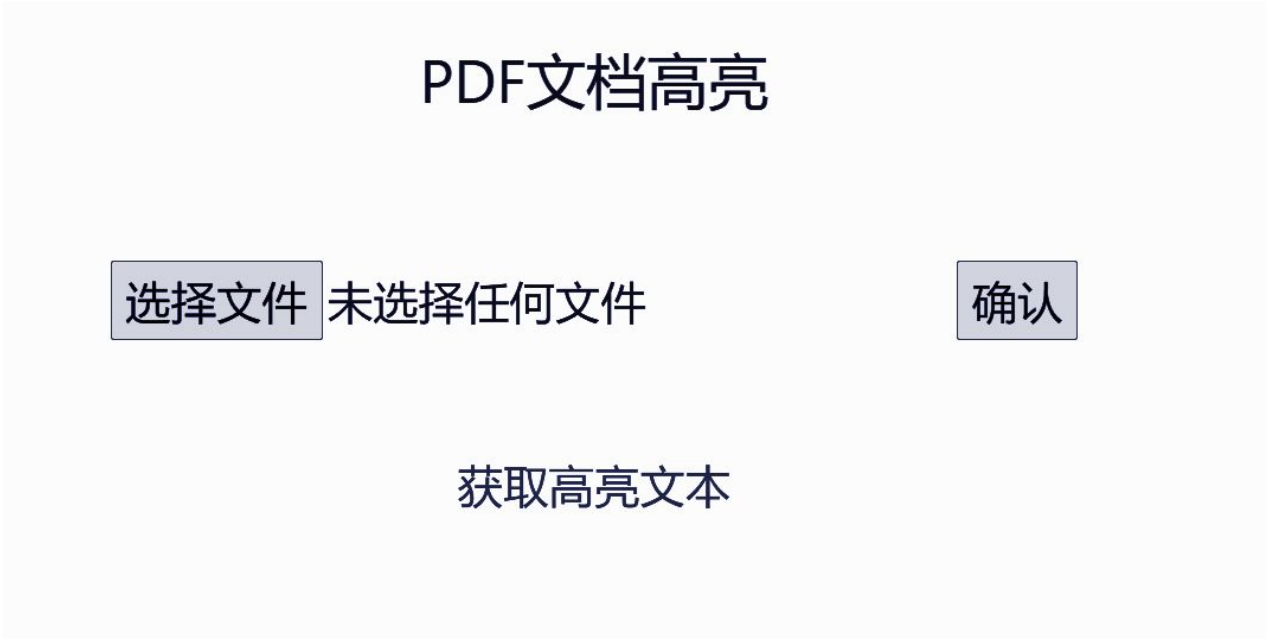
1. 把所有文章整合成文本数据；
2. 把文本分割成单个句子；
3. 将句子分割成各个词；
4. 计算得到各个词的词向量；
5. 根据TextRank计算Rankings；
6. 获取与top n的词相似的词；
7. 输出所有这些词，用于高亮显示。



文档高亮处理


转换PDF为HTML文件
采用BeautifuleSoup解析标签
向解析的文档中添加高亮样式
获取需高亮的关键词
正则匹配文档中的关键词
为所有关键词添加高亮样式

项目效果展示



项目效果展示

摘 要 使用传统的神经网络的短文本分类算法对其进行情感分类易出现定位误差等问题。为了解决对短文本情感分类时存在的定位误差，本文通过将词向量模型（Word2vec）、双向长短时记忆网络模型（BiLSTM）以及卷积神经网络（CNN）按照一定的框架进行组合，提出了 Word2vec-CNN-BiLSTM 的短文本情感分类模型。Word2vec-CNN-BiLSTM 模型采用对预处理后的文本进行向量化表示来提取文章特征向量，并在神经网络层进行双向语义捕捉实现文本的情感分类。实验结果显示 Word2vec-CNN-BiLSTM 的短文本情感分类模型有效解决了对短文本分类出现的情感分类定位误差问题。

关键词 神经网络；情感分类；词向量；短文本

中图法分类号 TP181　TP391.1　DOI:10.16707/j.cnki.fjpc.2020.01.003

摘　要：由于现场环境及测量技术的制约，工业测量过程中会出现关键参数数据缺失或失真现象。为了解决这一问题，将机器学习应用于参数测量过程中，实现对缺失数据集的补全和失真数据的修复，进而实现对缺失数据集的深层次挖掘和应用。以温度为例，作为燃烧过程中的重要参数，温度信息直接反映了燃烧状态，完整准确的温度测量信息对燃烧设备运行控制有着深远的意义。针对经验数据不完整的情况，使用机器学习中的主成分分析（PCA）算法迭代恢复缺失数据并进行特征提取，结合少量温度测量数据，实时重建燃烧温度分布。通过理论分析，确定重建算法的可行性。将重建算法应用于燃气燃烧仿真模型温度分布重建工作。结果表明，该方法能够准确地补全温度信息，快速实现温度分布重建，为解决工业参数测量中的数据缺失及数据失真问题提供一种新思路。

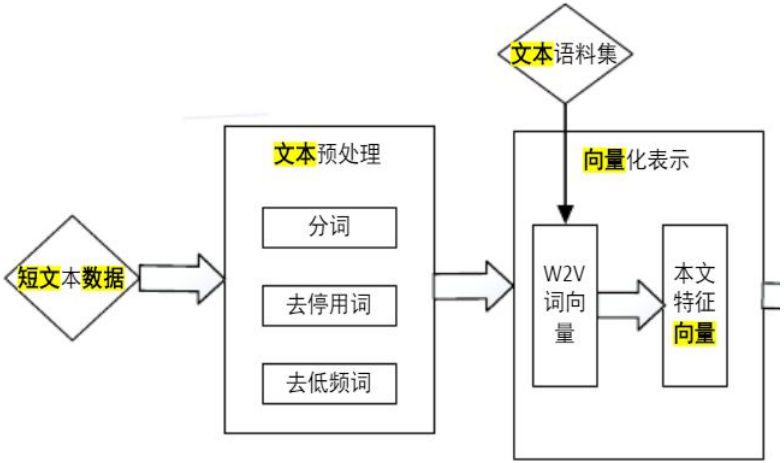
关键词：温度分布；机器学习；主成分分析；缺失数据；重建算法；最小二乘法；数值模拟；数据补全

项目效果展示

的问题，使得对文章进行情感分类产生较大的误差。通过引入 BiLSTM 捕捉较长距离的双向依赖关系[11]。利用 BiLSTM 编码从后到前的信息使对文本的情感分类的准确度得到提高[12]。使用 Word2Vec 将文章进行分析的同时使用 BiLSTM 捕捉较长距离的双向依赖关系。利用 BiLSTM 编码从后到前的信息，在更细粒度的分类时，对于褒义、中性的贬义的三分类任务更加注意情感词、程度词、否定词之间的交互，进行统计从而分析出文本的情感倾向。

本文提出一种基于 Word2vec 的 BiLSTM-CNN 混合神经网络模型，通过采用 BiLSTM 与 CNN 构建混合神经网络模型，并利用该模型进行短文本的情感分类。所构建的组合模型主要由以下几个模块

本数据
本文实
步
行训练
数据
步
CNN
经过池
步
数据分
中立
分和预



项目评价

优点：

1. 页面简单清晰，实现对PDF文档的上传，转换和本地存储；
2. 完成了对PDF文档的html转换以及文本的提取；
3. 采用TextRank提取并实现了对文本关键信息的高亮。

缺点：

1. 文本关键信息的提取可继续优化，可尝试其他模型
2. 文本内容提取只针对中文，对于中英文的文档，则不会标注英文内容。

谢谢