Web Clawer 说明文件

一、项目概述

本代码是一个用于网页数据爬取和文档处理的工具集，包含了网页数据爬取、查询重写、文档格式处理等多个功能模块。通过这些模块的协同工作，可以实现从网页上抓取相关信息，并对这些信息进行处理和转换。

二、使用流程说明

1.使用 web\_clawer/solver.py爬取网页数据并保存为 JSON 文件。

2.使用 web\_clawer/Deepclean.py 对 JSON 文件进行深度清洗。

3.使用 web\_clawer/JsontoTXT.py 将 JSON 文件转换为 TXT 文件，保存为知识库。

三、功能模块说明

1. Web 数据爬取模块（WebScraper）

功能：该模块负责从网页上爬取数据，支持按深度爬取任意网站，并进行智能过滤和优先级排序。可以根据关键词过滤无关内容，提高爬取效率和数据质量。

核心类：WebScraper

关键方法：

scrape\_url(url)：爬取单个 URL 的内容，支持任意网站。

collect\_web\_data\_with\_depth(start\_urls, depth)：按深度爬取任意网站，使用优先级队列进行智能过滤。

2. 查询重写模块（LLMQueryRewriter）

功能：该模块提供了多种查询重写方式，包括 Query2doc 风格重写、语义等价改写、意图细化与子问题分解、概念扩展与相关实体联想等，用于扩展查询，提高查询的准确性。

核心类：LLMQueryRewriter

关键方法：

rewrite\_query2doc\_style(original\_query, num\_keywords)：生成伪文档并提取关键词进行扩展。

rewrite\_paraphrase(original\_query, num\_paraphrases)：进行语义等价改写 / 释义。

rewrite\_intent\_refinement(original\_query, num\_subqueries)：进行意图细化与子问题分解。

rewrite\_concept\_expansion(original\_query, num\_concepts)：进行概念扩展与相关实体联想。

3. 文档处理模块（DocumentProcessor）

功能：该模块负责处理上传的文档文件，支持多种文件格式，包括 .txt、.md、.pdf、.html 等，并将文件内容提取为文本。

核心类：DocumentProcessor

关键方法：

process\_file(file\_path)：处理上传的文档文件，返回文件信息和内容。

4. TXT 转 PDF 模块（TxtToPdfConverter）

功能：该模块负责将 TXT 文件转换为 PDF 文件，支持中文显示，并可以批量转换。

核心类：TxtToPdfConverter

关键方法：

convert(txt\_file, pdf\_file, title)：将单个 TXT 文件转换为 PDF 文件。

batch\_convert(input\_path, output\_dir, \*\*kwargs)：批量转换 TXT 文件为 PDF 文件。

5. 配置文件（config.py）

功能：主要用于存储项目的各种配置信息。config.py 中包含了系统配置、数据源配置、文本处理配置、文件配置、输出文件配置以及日志配置等；专注于宋朝历史数据相关的配置，如数据源、爬虫设置、文本处理参数、历史领域特定配置等，同时还提供了 User - Agent 池和 Cookie 池。