Literature Review

1 Background study

With the continuous development of modern information technology, most countries in the world have entered the Internet era. Nowadays, most young people mainly apply for jobs through the Internet, and the main way for enterprises to recruit is gradually based on the Internet. Therefore, a variety of comprehensive recruitment websites have emerged on the Internet, such as India and Liepin. The enterprise publishes position information on these websites, and the candidates choose according to their needs. However, there are also disadvantages in the process of online recruitment. There will also be a release time for each published position information on the company's website. The longer the publishing time is from now, the two problems will be explained. One is that talents in this position are indeed scarce, and the other is that enterprises do not like to use Internet recruitment. Through the actual observation of several comprehensive recruitment websites, I found that they lack comprehensive display area for position information data. At present, with the continuous accumulation of comprehensive recruitment website data, there are many people want to mine useful information for analysis. D. Smith collects recruitment information about program developers from the recruitment website, and uses keyword index technology to study the demand trend of enterprise for the programming ability of job seekers, so as to provide important reference for the curriculum setting of computer major in Colleges and universities. (Smith 2014) Jia obtains the position information by retrieving the position information of data analyst in China's Lagoon network. His analysis is not only a single job but only a recruitment website, which will lead to incomplete data. (Jia 1019) In addition, most of the relevant researches at home and abroad adopt the traditional sampling method to sort out the recruitment website data by artificial statistics. (Jiang et.al 2016) The data volume is small and the efficiency is low, which is not suitable for the rapid and intelligent data collection and analysis in the big data environment.

2 Web data mining

2.1 Summary of Web data mining

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. It is the application of data mining technology in the web environment. It applies data mining technology to the web, and finds the implied, unknown, potential application value and non-trivial patterns from a large number of Web document collections and the relevant data browsing in the site. It deals with static web pages, web databases, web structures, user usage records and other information (Han 2011). Through the mining of these information, we can get the information that can't be obtained only by text retrieval. The research object of Web mining is the web centered on semi-structured and unstructured documents. There is no uniform pattern for these data. The content and representation of the data are intertwined. The data content basically has no semantic information to describe, and only relies on HTML syntax to describe the data structure. In order to analyze and process the semi-structured data, web mining must be combined with its research methods. (Li 2008)

2.2 Web data mining classification

According to the different data categories used in Web mining, we divide them into three categories. They are web structure mining, web content mining and Web usage mining. (Liu 2011).

2.2.1 Web structure mining

It refers to mining information from the organization structure and link relationship of web pages. Web structure mining is to mine the structure between web pages. At present, it mainly aims at link structure pattern According to the link relationship between web files, we can mine useful information except web content. (Li et al 2013)

2.2.2 Web content mining

Web content mining can be seen as the combination of Web Information Retrieval and information mining. It refers to the summary, classification, clustering, association analysis and trend prediction of a large number of document collections on the web. It is a process of extracting knowledge from web document content or its description, mainly divided into text information mining and multimedia information mining. The current research focuses on the use of word frequency statistics, classification algorithm, machine learning, metadata, part of HTML structure information discovery, hidden patterns between data discovery and generation of extraction rules, and the separation of concept and entity data from the page. (Li 2008)

2.2.3Web usage mining

Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of clickstream data in usage logs in order to produce the right data for mining. (Liu 2011)

2.3 The general process of Web Data Mining.

In the process of Web data mining, we can roughly summarize the main process of its work.

Firstly, we search for data from the target web resource. In this stage, the original data needed in data mining analysis will be collected. But there are usually many redundant data and incorrect data in this data set. Therefore, we need to preprocess the data next. This stage mainly includes data cleaning, data noise reduction, dimension specification, discretization and other methods. Through preprocessing, the data quality will be greatly improved, which can effectively reduce the time and cost of data mining and analysis.

2.4 Data preprocessing

2.4.1Data cleaning

The main purpose is to correct errors, standardize formats, and eliminate duplicate and abnormal data. Data cleaning includes filling in missing values, identifying or removing outliers. (Si et al 2018) Data cleaning can be divided into two categories: repeated data cleaning and missing values imputation. In order to improve the speed and accuracy of data mining, it is necessary to remove the duplicate records in the data set. If two or more instances represent the same entity, they are duplicate records. In order to find duplicate instances, it is usually done to compare each instance with other instances to find the same instance. For numerical attributes in an example, statistical methods can be used to detect. According to the mean value and standard deviation value of different numerical attributes, confidence intervals of different attributes can be set to identify the records corresponding to abnormal attributes, identify the duplicate records in the data set, and eliminate them. Similarity calculation is a common method in the process of repeated data cleaning. By calculating the similarity of each attribute of records, and considering the different weight values of each attribute, the similarity of records can be obtained after weighted average. If the similarity of two records exceeds a certain threshold, the two records are considered to be matched; otherwise, the two records are considered to point to different entities. (Luengo et al 2016)

2.4.2 Handling missing values

Most of the missing data is due to the wrong operation of manual input, the need for confidentiality of some information or the unreliable data source, which makes the content of the data set incomplete. When the wrong data mining model is applied to the front-end decision-making system, it will lead to serious deviation between the analysis results and the implementation decision. Missing values imputation can be regarded as incomplete data that can be ignored. It can ignore incomplete data directly by deleting attributes or instances. (Galar el at 2012) Missing values interpolation can also be a missing value interpolation algorithm based on filling technology. (Kong et al 2018) We use numerical information to fill in missing values. The simplest method is the average filling method. It takes the arithmetic mean of all complete data as the value of the missing data. The disadvantage of this method is that it may affect the original correlation between missing data and other data. If the missing values of large-scale data sets are all filled with the average method, because there are too many median values, more peak frequency distribution may mislead the mining results. (Galar el at 2012, Kong et al 2018)。Another method is to fill in the missing values by classification and clustering. Common missing value filling algorithms include EM algorithm (expectation maximization algorithm), MI algorithm (multiple imputation) and KNNI algorithm (k-nearest neighbor imputation) In the expectation maximum algorithm, the probability model is created to find the maximum likelihood estimate or the maximum posterior estimate. The success of the probability model depends on the unobservable hidden variable. (Gao et al 2011, Sotoca et al 2010)

2.4.3 Data transformation and integration

In this stage, the preprocessed data is formatted and stored in the database for subsequent data mining. According to the designed data warehouse structure, the preprocessed data is loaded into the database. After that, it will be more convenient to add, delete, modify and query the collected data, and improve the efficiency of subsequent data mining. (Si et al 2018）Data integration is to merge the heterogeneous data in the multi file or multi database environment to solve the semantic ambiguity. This part mainly involves data selection, data conflict and inconsistent data processing. (Kong et al 2018)

2.4.4 Data transformation

Data transformation is to find the characteristic representation of data. It uses dimensional transformation or conversion to reduce the number of valid variables or find invariants of data, including normalization, switching and projection operations. Data transformation is to transform the data into a form suitable for various mining patterns. According to the data mining algorithm used later, we decide which data transformation method to use. Common transformation methods include: function transformation, using mathematical functions to map each attribute value; normalizing the data, scaling the attribute value of the data, as far as possible falling into a small specific interval Standardization not only helps to implement all kinds of classification and clustering algorithms, but also avoids over dependence on measurement units, and avoids the occurrence of weight imbalance. (Guan 2015, Kong et al 2018)

2.4.5 Data reduction

It is based on the understanding of the discovery task and the content of the data itself, looking for the useful features of the expression data that depend on the discovery target, in order to reduce the data model, so as to simplify the data as much as possible and promote the more efficient big data mining on the premise of keeping the original data as possible. (Kong et al 2018）

2.4.5.1 Dimensionality reduction

The technologies involved include feature selection and space transformations. The core of dimension reduction is to reduce the number of random variables or attributes. The purpose of eigenvalue selection is to obtain the attributes that can describe the key features of the problem. By removing irrelevant and redundant attributes, the machine learning process is faster and the memory consumption is less. The focus of quantity reduction is to reduce the amount of data and select a smaller data representation from the data set. (Kong et al 2018）The main numerical reduction techniques include log linear model, histogram, clustering, sampling and so on. Common algorithms include LVF (Las Vegas filter), MIFs (mutual information feature selection), MRMR (minimum redundancy maximum relevance), Relief algorithm. Space transformations is another way to reduce data dimensions. Popular algorithms include LLE (locally linear embedding), PCA (principal components analysis), etc. (Wang et al 2010, Kong et al 2018)

2.4.5.2 Instance reduction

This is now a very popular algorithm to reduce the size of data set is the instance reduction algorithm. While reducing the amount of data, it does not reduce the quality of knowledge acquisition. By removing or generating new instances, the data scale is greatly reduced. The technologies involved include instance selection and instance generation. Good instance selection algorithm can generate a minimum data set, remove noise data and redundant data, independent of the subsequent data mining algorithm. Common algorithms include CNN (condensed nearest neighbor), ENN (edited nearest neighbor), ICF (iterative case filtering), drop (decremental reduction by ordered projections), etc. Instance generation establishes various prototypes for instance generation, involving algorithms such as LVQ (learning vector quantification). (Perezortiz et al 2015, Kong et al 2018)

2.4.5.3 Discretization

Its purpose is to reduce the number of given continuous attribute values. Before discretization, we first estimate the scale of discrete data, then sort the continuous data, and then specify several split points to divide the data into multiple intervals. All continuous data falling in the same interval are mapped to the same discrete data by a unified mapping method (Prati et al 2015). According to the different identification methods of split points, discretization can be divided into top-down and bottom-up. According to whether to use classified information, it can be divided into two categories: supervised and unsupervised. At present, most discretization methods are divided into two directions: one is to discretize based on the importance of attributes, and the other is to map based on the resolution matrix. (Kong et al 2018) Common algorithms include: MDLP (minimum description length principle), CAIM (class attribute interdependency maximization), etc. (Angiulli et al 2007)

2.4.5.4 Imbalanced learning

When using supervised learning of machine learning to form data model, it is easy to produce huge priority differences in different types of data sets. Many standard classification learning algorithms often tend to ignore the priority class. (Bacardit et al 2012) Data preprocessing technology can avoid the imbalance of type distribution. The main methods are under sampling and over sampling. The former is to remove most instances as much as possible when creating a subset of the original dataset as data mining. The latter is to copy many of the same instances or create new ones during sampling. Among many sampling algorithms, the most complex and famous genetic algorithm is SMOTE (synthetic minority oversampling technique). (Kong et al 2018)

2.5 Web data mining techniques

After reading a part of literature, I found that in the method of obtaining web data information, people generally use Python for data mining, but some people will choose to use java to develop crawlers. Next, I will discuss the development of web crawler using Python and Java respectively.

3 Python-based Data analysis

3.1 Python Common Library

Python has Numpy, Pandas, Matplot lib, Scikit learn and other libraries with complete functions and unified interfaces, which can provide great convenience for data analysis. (Zhai 2018)

3.1.1 NumPy

This library, whose name means Numerical Python, actually constitutes the core of many other Python libraries that have originated from it. Indeed NumPy is the foundation library for scientific computing in Python since it provides data structures and high-performing functions that the basic package of the Python cannot provide. In fact, NumPy defines a specific data structure that is an N-dimensional array defined as ndarray. The knowledge of this library is revealed in fact essential in terms of numerical calculations since its correct use can greatly influence the performance of a computation. (Nelli 2015) Besides，Numpy also has the following features: (Zhai 2018)

1) It has functions that perform element level calculation on arrays and directly perform mathematical operations on arrays.

2)It can integrate C, C++ code into Python.

3) It can be used as a container for transferring data between algorithms.

3.1.2 Pandas

This package provides complex data structures and functions specifically designed to make the work on them easy, fast, and effective. This package is the core for the data analysis with Python. The fundamental concept of this package is the DataFrame, a two-dimensional tabular data structure with row and column labels.Pandas combines the high performance properties of the NumPy library to apply them to the manipulation of data in spreadsheets or in relational databases (SQL database). In fact, using sophisticated indexing it will be easy to carry out many operations on this kind of data structures, such as reshaping, slicing, aggregations, and the selection of subsets. (Nelli 2015) In addition, Pandas can also be used for data preprocessing, such as data consolidation, data cleaning, data standardization and data conversion. (Zhai 2018)

3.1.3 Matplot-lib

This package is the Python library that is currently most popular for producing plots and other data visualizations in 2D. Since the data analysis requires visualization tools, this is the library that best suits the purpose. (Nelli 2015) Matplotlib consists of four parts: (Li 2018)

(1) The basic figure type of Matplotlib;

(2) Adjust the style and color of figure;

(3) Add notes to the drawing (including coordinate axis range, length width ratio or coordinate axis, etc.);

(4) Other complex figures.

3.1.4 Scikit-learn

Scikit-learn is a simple and effective data mining and analysis tool. It is based on Numpy, SciPy and Matplotlib, and encapsulates some common algorithms. Its main modules are data preprocessing, model selection, classification, clustering, data reduction and regression and other machine learning algorithms, which can help users to quickly build models in the process of data analysis, and the model interface is unified, which is very convenient to use. (Zhai 2018)

3.1.5 WordCloud

The system can import Wordcloud to generate the specified word cloud, and remove keywords without substantive statistical significance in the continuous optimization process. (Guo 2018)

3.2Crawler search method

3.2.1 Breadth-first search

Breadth-first traversal is a search strategy widely used by crawlers. Its process is to have a URL queue first, pop up the URL in the queue, then extract the sub URL in the pop-up links, and put them back in the original URL queue to wait for pop-up. The URL that have been searched will be put into a table similar to the collection. Each time a new pop-up URL is processed, a judgment will be made first to see if there is a URL in the searched table. If there is one, skip it and carry out the next operation. The advantage of this operation is that it reduces the repetition rate without repeated crawling. The disadvantage is that the information update will not be timely after the page update, and each judgment will consume more resources and time. (Wang et al 2019)

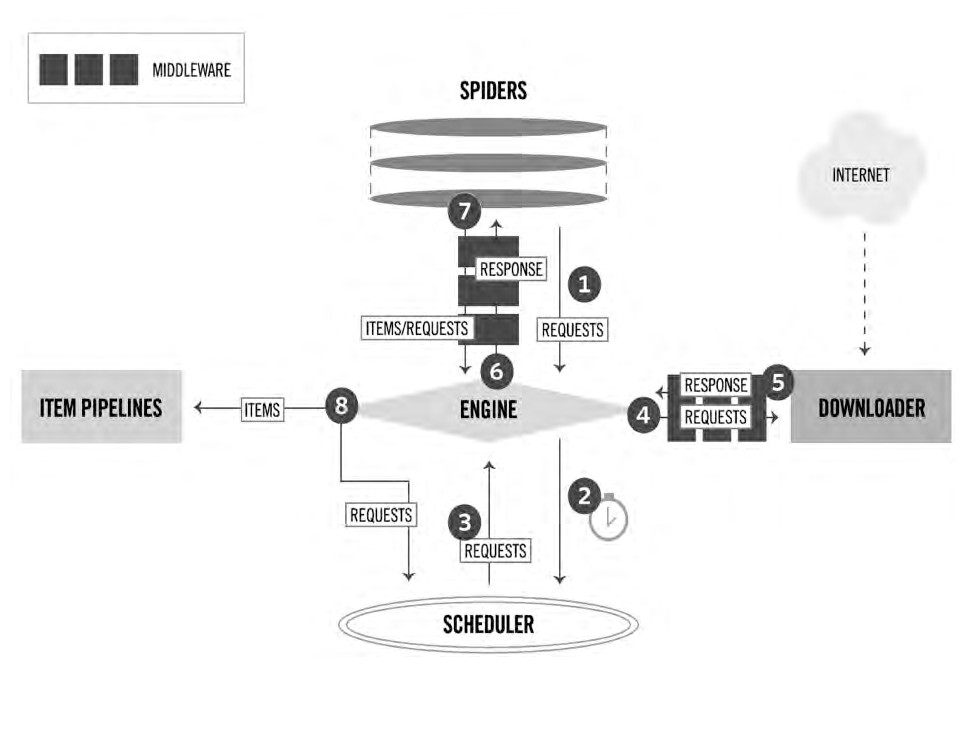
3.2.2 Depth-first traversal algorithm

Depth-first traversal is often used in early crawler development. The general process is as follows: Firstly, find the first hyperlink URL from an HTML page. Then extract the URL, and then extract the first URL within the URL. It is always a single chain mode to dig deep until there is no next URL, and then return to the first HTML interface, and perform the same operation from the second URL. That is to say, there will be a complete single chain search before the next capture. When there is no next step for all links, the search ends. (Wang et al 2019) The advantage of this approach is that it can find all the deep URL, but it also has disadvantages, because once the search starts, it may fall into a permanent deep search and cannot jump out. (Zhao et al 2017, Wang et al 2019)

3.2.3 Design and implementation of Crawler Based on scratch framework

Scrapy is a fast and high-level application framework written in Python for crawling website data and extracting structural data. At present, it is the most famous and widely used framework among all crawler frameworks. Scrapy is now capable of interacting with APIs in order to extract data. There are many reasons that contribute to the use of Scrapy as the WebCrawler. (Shi et al 2016)

Instead of designing a crawler framework from scratch, we can learn how to use Python's scratch framework simply and efficiently. (Liu 2018) Developers only need to develop a few specific modules to write a stable and efficient web crawler. (Sun et al 2019) Its detailed framework and operation process are shown in the figure. (Li et al 2017, Wang et al 2019)



3.2.3.1The running process of the Scrapy Crawler

The Engine gets URLs to scrape from the Spider and schedules them in the Scheduler, as Requests. Then, the Engine asks the Scheduler for URLs to crawl, as Requests, and send them to the Downloader, passing through the Downloader Middleware. Once the page finishes downloading, the Downloader generates a Response and sends it to the Engine. The generated response contains the copy of the static HTML of the web page. The information extracted is stored in Items that are data holders of the framework. Then through the use of the Item Pipeline, the Items or data can be saved in suitable Formats including but not limited to CSV or SQL database. (Bassam 2016, Shi et al 2016)

3.2.3.2Modules of the scrapy framework

1) Scrapy Engine: It is responsible for the transfer of regulatory data between modules in the system and calling corresponding functions to respond to specific events.

2) Scheduler : It is responsible for the unified management of all URL resources to be crawled. For example, insert the URL resources submitted by the receiving scrapy engine into the request queue. Then, the URL is taken from the queue and sent to the scrapy engine in response to the URL request from the scrapy engine.

3) Downloader Middlewares : It is responsible for delivering the URL request sent by the scrapy engine to the downloader module and the HTTP response sent by the downloader module to the scrapy engine.

4) Downloader : It is responsible for downloading the data on the web page and finally sending it to the crawler module through the scrapy engine.

5) Spider : It analyzes the data obtained from the downloader module, and then extracts the item or relevant URL resources.

6) Spider Middlewares : It is responsible for the input and output of the crawler module.

7) Item Pipeline : Process items extracted and sent by the crawler module through data cleaning, data validation, data persistence and other operations. (Cattell 2011, Li 2017)

3.2.4Strategy of anti-crawler Technology

Nowadays, many websites prohibit crawlers from crawling data. Websites use headers, user behavior, website directory, data loading and other ways to anti crawl, so as to increase the difficulty of crawling. Thus, there are several strategies:

3.2.4.1 Set download\_delay parameter

If the download waiting time is too long, the task of large-scale data grabbing in a short time will not be completed, and too short will increase the probability of being prohibited from crawling data. So we set DOWNLOAD\_DELAY = 2 in settings.py (Li 2017)

3.2.4.2 Disable cookies，

This prevents crawler behavior from being detected by sites that use cookies to identify crawler tracks. So we need to set COOKIES\_ENABLES = False. (Li 2017)

3.2.4.3 User-agent

User agent also refers to browser, including hardware platform, system software, application software and user's personal software preferences. (Chen 2016) Every browser and regular web crawler has a fixed user agent. Camouflage user agent can judge the category of the website visitors by violating rules. For camouflage browser and famous crawler, camouflage browser is more recommended. Compared with the crawler, the browser has no fixed IP and can be anyone, while the crawler has a fixed IP. Camouflage browser can improve multiple user agents. Each time a request is sent, a user agent can be randomly selected to set the code according to the specific needs. (Liu 2019) We need to set DOWNLOADER\_MIDDLEWARES = { ' scrapy.contrib.downloadermiddleware.useragent.UserAgentMiddleware':None,' HouseInfoSpider.spiders.rotate\_useragent.RotateUserAgentMiddleware' : 400，} in settings.py. (Li 2017)

3.3 Screening technology of Python based web crawler

A web crawler is a web robot. It is an important part of search engine. It is a program that can automatically extract the content of specific pages on the Internet. As for data mining, the first step is to determine the location of data storage, which can be successfully extracted when data is found. When browsing the web page, all kinds of elements on the web page are composed of data, which needs to be mined out from the web page. The workflow is summarized as follows: (Wang et.al 2019)

(1) Grabbing the page code through URL;

(2) Obtaining the useful data or URL on the page through regular matching;

(3) Processing the acquired data or entering the next round of grabbing cycle through the acquired new URL.

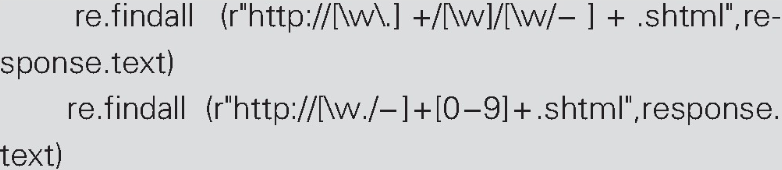
There are four ways to filter web crawlers in Python.

3.3.1 Regular expression grabs website data

Regular expressions, also known as normal representation, are often used to retrieve the text that conforms to a certain pattern. It first sets some special words and character combinations, and filters the expression through the combined "rule string", so as to obtain or match the specific content we want. It has the advantages of flexibility, logicality and functionality. It can quickly find the required information from the string through the expression. (Fu 2019). Specifying a Regex in a selector allows the matching the extracted data or string to the Regex logic and, thus filtering out the precise data. (Farooq et al 2016)

In Python, we can use the built-in RE module to use regular expressions. The most common find method we use in Python is the findall(). When we use the findall (), we can simply get a list of all the matching patterns.

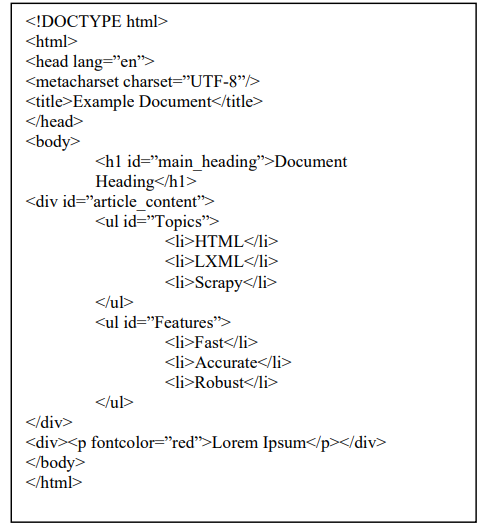
Regular expressions with the same meaning and different writing methods are as follows: (Wang et.al 2019)



3.3.2 XPath

XPath is a language for finding information in XML documents, which is used to navigate through elements and attributes in XML documents. Using XPath can easily locate the interested nodes in HTML documents. Lxml library is the third-party library of python, which supports the standard XPath specification. A language that can navigate and extract tags from XML documents. In XPath, there are five basic types of nodes: document (root) node, category, text, attribute, element. In order to distinguish between the root node and the category, the XPath language stipulates that special symbols should be added before the name. The etree package needs to be imported from the Lxml library before use. (Wang et.al 2019) Besides navigating through the HTML tag hierarchy, XPath selectors can also look at the content residing within the HTML tags, making them extremely useful when crawling a heavy content-based website. (Farooq et al 2016)

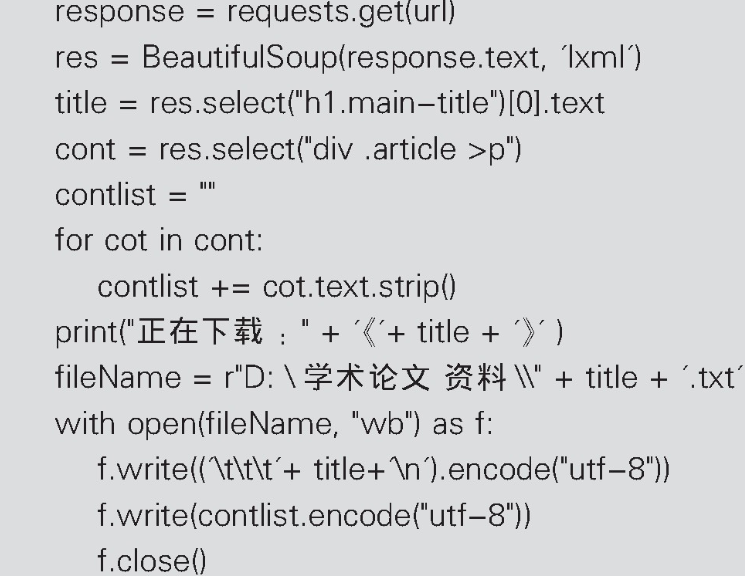
The code is as follows: (Farooq et al 2016)



3.3.3 Beautiful Soup

Beautiful soup does not have its own parser. It passes in a parameter when building an object to specify the parser. Among these parsers supported, you can use your favorite one

(Wang et.al 2019)



3.3.4 Urllib library in Python

Urllib is the library used for network requests in Python standard library The library has four modules: urllib.request, urllib.error, urllib.parse and urllib.robotparser。 其Urllib.request and urllib.error are two libraries that are frequently used in crawlers. The urllib.request module that people need to use to simulate a browser to send an HTTP request. The function of urllib.request is not only to initiate the request, but also to get the return result of the request. (Chi 2019)

3.4 Data storage

The crawled data can be stored locally or in the database.

3.4.1 Local storage

JSON files can be directly created in pipeline to write data, but the readability of JSON files is poor. Therefore, we can save the JSON file as a readable excel file after further processing. The extracted data table is a two-dimensional table structure. The Pandas library is the main data processing tool in Python. With the help of Dataframe, we can store two-dimensional table structure. First, build a dictionary, then a two-dimensional dictionary priceall, save it as a Dataframe, then build an index of Dataframe with a list, and finally save it as an excel file using the to\_excel() method of dataframe (Li et al 2018)

3.4.2 Database storage

The data model defined in scrapy establishes tables in the database, By using pymysql, a third-party module of python, to log in to the database, and executes SQL statements to insert data into the database. In order to achieve incremental crawling, a verification field needs to be created in the data table. (Li et al 2018)