Read me

Hanxiong Wu

wh854

1. Run environment

* Python 2.7.8
* Standard Python library

1. Input format

Program can be executed in two ways:

1. Run function “query(keywords, number)” of program module with two function parameters:

keywords: single or multiple query keywords concatenated by space;

number: the expected page number

1. Run main function with two command line parameters separated by space:

keywords: single or multiple query keywords concatenated by ‘\_’;

number: the expected page number

1. Output format

While normally crawling, program will print lot of information about current executing states. It’s generally in this format:

* Firstly print 10 urls get by Google search;
* As a new crawling level x begin, it’ll print “\*\*\*\*\*\*\*\*\*level x begin\*\*\*\*\*\*\*\*\*\*”;
* For each url visited, it’ll print:
* “Visiting ”+ current visiting url;
* Exception information, include:

"Robot access prohibited": the robots.txt forbid program visiting this url;

"Invalid robots.txt url": cannot read robots.txt by this url;

"Empty url": current visiting url is empty;

"Unable to get page source: url may be invalid": may due to various reasons;

"Unable to fetch page source": fail to fetch page source, may due to various reasons;

* Relevant score of this page;
* The number of new links get from this page;
* The number of visited urls(As the question statement, program should keep crawling until get enough pages specified by user, so the number of visited urls may larger than the number due to invalid urls).
* Program will finish at two states:
* Success: Got enough pages specified by user;
* Fail: Haven’t get enough pages but cannot get any more url. If this, program will print “Dead end”
* Once program finished, it’ll print “\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*”, then print the total number of downloads and total download size.

The pages downloaded will be stored at a directory named “pages” at the root directory of program.