University of Leeds **School of Computing**

COMP3011, 2021-2022

Web Services and Web Data

Web Crawler

By

Jasmine Tarbard

201254554, sc18jt@leeds.ac.uk

Date: 08/05/2022

1. The Code & Index

File: crawler.py

Required libraries: time, pickle, urllib, requests, bs4, string.

There are three classes used to implement the web crawler: InvertedIndex, InvertedIndexHandler, and Crawler.

Inverted Index

The Inverted Index class has two data structures and two methods and represents the inverted index as an object.

Data structures:

- Documents: an array to store the crawled urls.
- Postings: a 2-dimensional dictionary containing pointers to the documents and the frequency of the word's occurrence in the document.

Methods:

- Search: implements the find commands functionality, accessing the index via documents and postings. The algorithm used to compute the scores of the search results is an implementation of document-at-a-time.
- Print: implements the print functionality, accessing the postings dictionary.

Inverted Index Handler

The Inverted Index Handler has no data structures and two methods and is used to handle the saving and loading of the index.

Methods:

- Save: saves a provided index object to a pickle file, index.pkl.
- Load: loads the local index from index.pkl into a Inverted Index class instance.

Crawler

The Crawler class has several variables/data structures, these are mostly supplementary to the functioning of the methods and so will not be documented here, and four methods. The Crawler class implements the crawling and parsing of the provided seed.

Data Structures/Variables:

- Root: the seed provided at the command line.
- Frontier: an array, that is utilised as a queue, containing the websites scraped links.

Methods:

- Robots: attempts to access the robots.txt file of the provided seed and, if successful, sets the delay length for the crawler.
- Parse: receives a soup object and document id and parses the text contents of the html adding each word to the index via the index reference.
- Crawl: receives a url and performs a request and scraping the page to obtain links to add to the frontier and calls parse to add the documents contents to the index.
- Run: receives a seed, calls robots to attempt to set a delay and then handles looping through the frontier until the website has been crawled. Uses the handler to save the index on completion.

2. Using the client

The client is run at the command line with the Python 3 compiler and without any arguments. It will then enter a loop to receive commands and their arguments.

Supported commands: build, load, print, find. Structure: <command> <arguments>

Build

The build command crawls a website and builds an inverted index, saving it to the local file system.

Argument 1 (required): The url/seed.

Argument 2 (optional): An option to disable politeness when web crawling, where politeness is restricting the number of web requests by the crawl-delay set in robots.txt. Defaults to true without any input. To disable politeness pass 'False'.

Example: build http://example.python-scraping.com/

Load

The load command will find and load into the program an existing index from the local file system.

Example: load

Print

The print command prints the contents of the inverted index in its completeness or for a single term.

Argument 1 (optional): Posting to be printed.

Example: print Peso

```
Command: print Peso
Peso                    57:1, 96:1, 317:1, 320:1, 330:1, 351:1, 353:1, 578:1, 603:1, 614:1, 630:1, 636:1,
637:1, 644:1, 646:1
```

Find

The find command will find and return documents containing a term or a list of terms separated by a space.

Argument 1 (required): A term to be found in the index.

Argument n (optional): Further terms that are used in conjunction with the first.

Example: find United Kingdom

```
Command: find Area Afghanistan

1. http://example.python-scraping.com/places/default/view/Afghanistan-1
Relevence Score: 2, Occurances: 'Area': 1, 'Afghanistan': 1

2. http://example.python-scraping.com/places/default/iso/AF
Relevence Score: 2, Occurances: 'Area': 1, 'Afghanistan': 1
```