DeepWeb

Assignment 2 - DeepWeb and Content Summary

Team Member

Prakhar Srivastav (ps2894)

Running the program

On CLIC

To run on the **CLIC** machines, just run the following set of commands.

```
$ cd /home/ps2894/ps2894-proj2
$ ./run.sh
```

Locally

```
$ cd <project_folder>
$ ./run.sh
```

Bing Account Key

The bing account key is `Byygq1zI2KKyssKp8UvVe3DV/v6Aa0FEsKrE+pqDa0s`

Internal Design

The project is comprised out of 4 files out of which 2 are auxilliary and the rest two are primary to the purpose of the project. The `config.py` and `bing.py` files contain boilerplate configuration variables that are required for running the project. Two other files, `crawler.py` and `starter.py` are the main files that deal with building content summaries and classfication respectively.

The `crawler.py` file is responsible for the following -

- crawling the page (using `lynx`)
- writing the output to a cached file
- cleaning the page content by filtering out special characters
- generating a content summary for a particular database and category
- writing the results of the summary to an output file

The `starter.py` file is responsibe for the following -

- takes input from the user

- starts reading from the top of the taxonomy (i.e root.txt)
- classifies a database, then proceesds on the child category
- builds a final map of categories to documents and then passes this data to the `crawler` for use in building the content summary

Lastly, the `cache` folder stores the intermediate output of crawling content of the webpages. The `result` folder conversly stores the generated content summaries.

Other Info

For multiple-word entries, I have decided **not** to include multiple-word information in the content summaries.

File List

```
|-- README.md
|-- cache
|-- data
|   |-- computers.txt
|   |-- health.txt
|   |-- root.txt
|   |-- sports.txt
|-- results
|-- run.sh
|-- src
    |-- __init__.py
    |-- bing.py
    |-- config.py
    |-- crawler.py
    |-- starter.py
```