Web Crawler Corpus

important Files:

> a6.py

Main Libraries Used:

- NLTK (sentence/word tokenizers, stopwords)
- BeautifulSoup4
- SKLearn (TfidfVectorizer)

What

A program to crawl anime websites and generate a knowledge base of important terms and example usage.

The program takes in a list of websites as an argument and scrapes them for data and additional links. Any additional links found within the page are added to a queue to be scraped as well. (Up to `maxSiteLinks` and `maxSearchDepth`)

The program analyzes the website data and determines the most important terms within the page. The important terms are then added to a knowledge base where any sentences containing the term are listed.

The knowledge base is output as 2 json files. The "auto" file contains every term that was automatically found whereas the "user" file contains specific terms specified within the program.

How to Run

1. Ensure Python 3.X is installed
2. Ensure all dependencies installed

Windows/Mac/Linux:

```
python3 a6_kar180005.py <link1> ...
```

Input

The program takes in a list of websites to crawl from.

The main supported websites are:

- reddit.com
- myanimelist.net

The following sites may also be visited during the crawling process:

- aniilist.co
- anime-planet.com
- animenewsnetwork.com
- crunchyroll.com
- vrv.co
- funimation.com
- youtube.com
- youtu.be
- wikipedia.org

Output

The program outputs data to the following 3 folders within the working directory:
Notes:

- If the directories don't exist the program will automatically create them
- If the directories already exist and contain data, the program will override and delete the data
- An error may occur if the directories are open/in use by another program, make sure the directories are free when using the program

/raw_files/

Contains files with raw website data with no processing done to it

/tokenized_files/

Contains files with processed and sentence tokenized website data

/knowledge_base/

Main output files of the program.

AutoKB.json contains a list of terms that the program viewed as important and examples of their usage within the websites.

UserKB.json contains a subset of AutoKB with only the terms that have been specified at in the `desiredTerms` variable.

Calculating Important Terms

Statistics for important terms are calculated using term frequency-inverse document frequency. (tf-idf). The value for a given word is calculated in proportion to the number of times the word appears in the sentence vs the number of sentences that contain the word.

As the program is running, the console displays the top 40 important terms found within the website as well as its relative value.

More info regarding tf-idf can be found at: https://en.wikipedia.org/wiki/Tf–idf

Cleaning the Data

The data retrieved from the websites is full of unnecessary data and lacks any sort of formatting. The cleaning process removes any unnecessary elements that have certain tags/classes/ids. These were manually specified after looking at the website structure.

Additionally, only certain websites are accepted and any urls containing "ignored" keywords are left out. Some examples of ignored keywords are "login" or "user" which usually lead to undesirable pages.

Sample Run

Terminal:

```
python3 a6_kar180005.py https://myanimelist.net/anime/44511/Chainsaw_Man https://www.reddit.com/r/anime/search/?c
```

Terminal Output:

```
httpsmyanimelistnetanime44511ChainsawMan:
{'anime': 9.0, 'japanese': 9.0, 'chainsaw': 6.0, 'denji': 5.0, 'devil': 5.0, 'man': 5.0, 'official': 5.0, 'theme'

httpswwwredditcomranimesearchqchainsaw20man:
{'man': 8.0, 'days': 6.0, 'byhttps': 5.0, 'chainsaw': 5.0, 'new': 4.0, 'agoofficial': 3.0, 'anime': 3.0, 'agonews

...
```

Folder Output:

Note: The following are small snippets of the actual output

`/raw_files/httpsmyanimelistnetanime44511ChainsawMan`

```
Chainsaw Man Edit Notify me when it starts! Add to My List * Your list is public by default. Status: Eps Seen: /
```

`/tokenized_files/httpsmyanimelistnetanime44511ChainsawMan`

```
Add to My List * Your list is public by default.
Status: Eps Seen: / ?
Your Score: Add Detailed Info Add to Favorites Alternative Titles Japanese: チェンソーマン English: Chainsaw Man M
Premiered: Fall 2022 Broadcast: Wednesdays at 00:00 (JST) Producers: Shueisha ...
```

`/knowledge_base/userKB.json`

```
{
  ...
  "chainsaw": [
    "agoCHAINSAW MAN CHAINSAW MAN CHAINSAW MAN CHAINSAW MAN CHAINSAW MAN CHAINSAW MAN CHAINSAW MAN CHAINSAW MAN C
    "Then like imagine him in a horror film roaring as he runs across the room to kill some murderer, primal grav
    "Now able to transform parts of his body into chainsaws, a revived Denji uses his new abilities to quickly ar
    ...
  ],
  "denji": [
    "It was discovered by Denji that Makima doesn\u2019t remember one face from another, but that she instead per
    "Through this, he was able to kill her, knowing that she only ever saw Chainsaw Man, and never once cared for
    "However, in an unexpected turn of events, Pochita merges with Denji's dead body and grants him the powers of
    "It's implied that Denji ate her.",
    ...
  ],
  ...
}
```

Sample Dialogue:

```
What is Chainsaw Man about?
```

> Denji has a simple dream—to live a happy and peaceful life, spending time with a girl he likes. This is a far cry from reality, however, as Denji is forced by the yakuza into killing devils in order to pay off his crushing debts. Using his pet devil Pochita as a weapon, he is ready to do anything for a bit of cash. Unfortunately, he has outlived his usefulness and is murdered by a devil in contract with the yakuza. However, in an unexpected turn of events, Pochita merges with Denji's dead body and grants him the powers of a chainsaw devil.

```
Who is Denji?
```

> Denji (デンジ Denji?) is the protagonist of the Chainsaw Man manga series. As a young boy, he inherits his father's debts from the Yakuza. After meeting Pochita, he becomes a Devil Hunter for the Yakuza in an attempt to clear his debt.