US Patent Web Crawler Documentation

Introduction

This program can automatically download the information in US Patent Database and save it into .csv file.

Guide to install python environment on Windows machine

Goto Anaconda and download Python 3.7 for Windows, or simply click here.

Open Anaconda Prompt, and you can use the command line. Simple usage in command line::

```
# Go to Disk D:/
$ D:
# Go to folder /sample/
$ cd sample
# Go to parent folder
$ cd ..
```

Installation

Use `requirements.txt` to install all package dependencies.

In command line:

```
# Navigate to the folder of main.py
$ cd [TOP_FOLDER]
# Install all packages
$ pip install -r requirements.txt
```

Single mode: get one patent at a time

Single mode is to download the information of 1 patent. For example, if you want to download the information of this patent. In `input_URL.txt`, give the URL in the first line.

In command line:

```
# Run the python code and specify the mode
$ python main.py --mode single
```

And the result .csv file would be saved in `output/patent_info.csv`. Note that this will overwrite the file if there already exist `patent_info.csv`.

Many mode: get many patents at a time

Let's say you want to download all the patents information in this query page, including the next list so total of 560 patents. In `input_URL.txt`, give the URL in the first line.

In command line:

```
# Run the python code
$ python main.py --mode many
```

And the result .csv file would be saved in `output/patent_info.csv`. Note that this will overwrite the file if there already exist `patent_info.csv`.

Specifying input file or output file path

You can specify output file path and name.

In command line:

```
$ python main.py --mode [MODE] --input [INPUT_FILE_PATH] --output [OUTPUT_FILE_PATH]
```

For example

```
$ python main.py --mode many --input ./my_URL.txt --output ./output/my_result.csv
```

Checkpoint

At run time, the program would create a checkpoint file in `./output/checkpoint.pkl`. The program would automatically load the checkpoint and continue from the previously disrupted point.

Simplify command line options

Command line option can be simplify to one letter.

In command line:

```
$ python main.py --mode many --input ./my_URL.txt --output ./output/my_result
```

can also be used by:

```
$ python main.py -m many -i ./my_URL.txt -o ./output/my_result
```

CSV file format

Here a list for tag name used in .csv file:

Tag Name	Description	Example
ID	Patent number (9~11 digits)	5479556
title	Patent title	Rotation control apparatus employing a comb filter and phase error detector
date	Patent date (YYYY/M/D)	1995/12/26
inventor_name	Name of inventor	Oh
inventor_city	City of inventor	Seoul
inventor_country	Country inventor	KR
assignee_name	Name of assignee	Goldstar Co., Ltd.
assignee_city	City of assignee	Seoul
assignee_country	Country of assignee	KR
US_class	US class number	388/805
CPC_class	First four digits of CPC class number	H02P
international_class	First four digits of international class number	H02P
reference	This is a reference patent	
referenced_by	This is a patent that referenced by	

Warnings

- Remember to close output .csv file before running the programm, or it can not access the output file.
- Some time some particular patents might not have particular information, the program would give warning messages and the program would continue. if you want to suppress all warning messages, you can use:

```
$ python main.py -m many --warnings False
```

Debug

Use `--debug True` to turn on debug message.

```
$ python main.py -m many --debug True
```

Contact

Please send to tommyrpg1010@gmail.com if you have any question or need any modification of the functions.