

1 How to run a scraper

1.1. Open a new cmd:

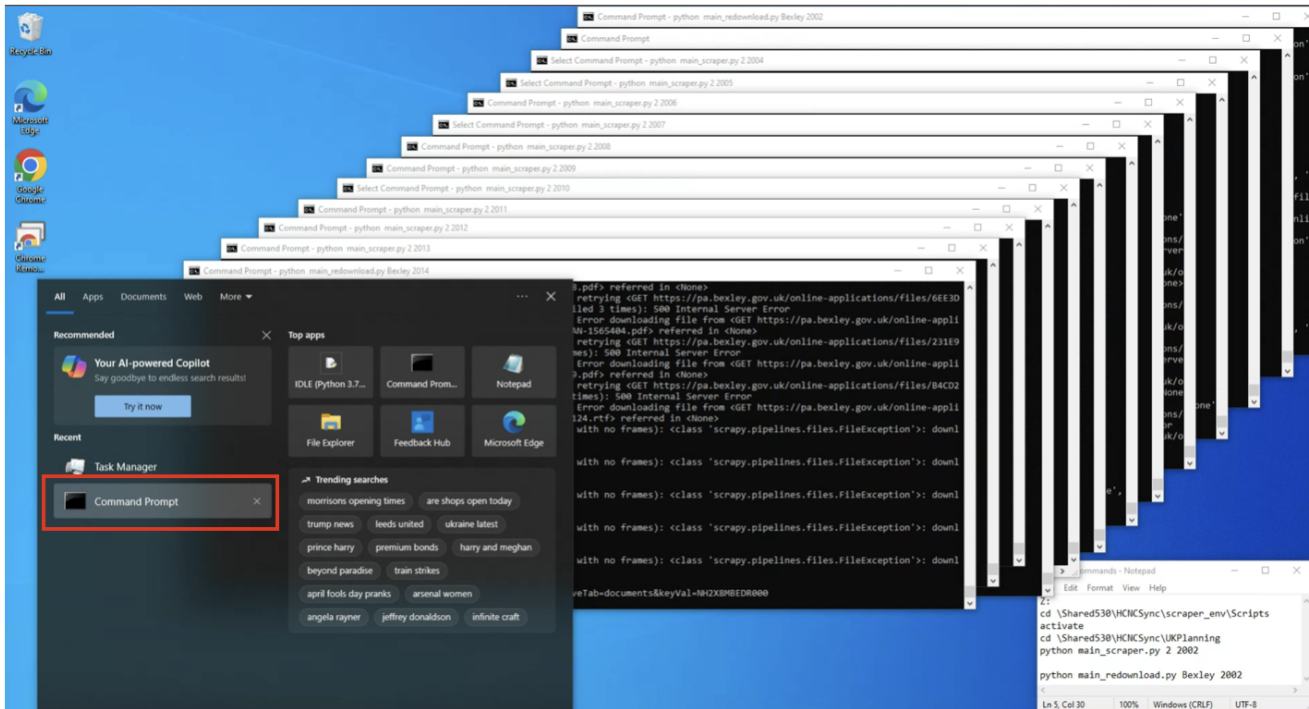


Figure 1: Open a new cmd.

1.2. Copy the first 5 lines of commands from [HCNCSync/cmd commands.txt](#):

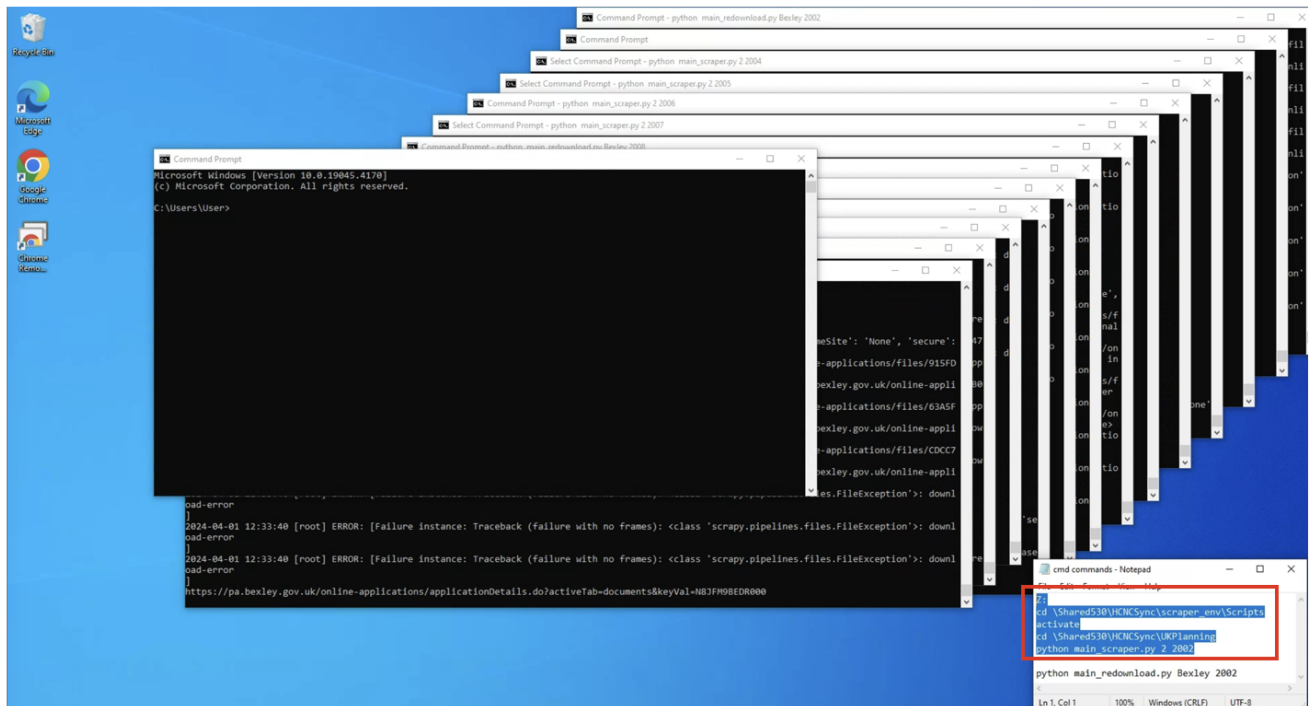
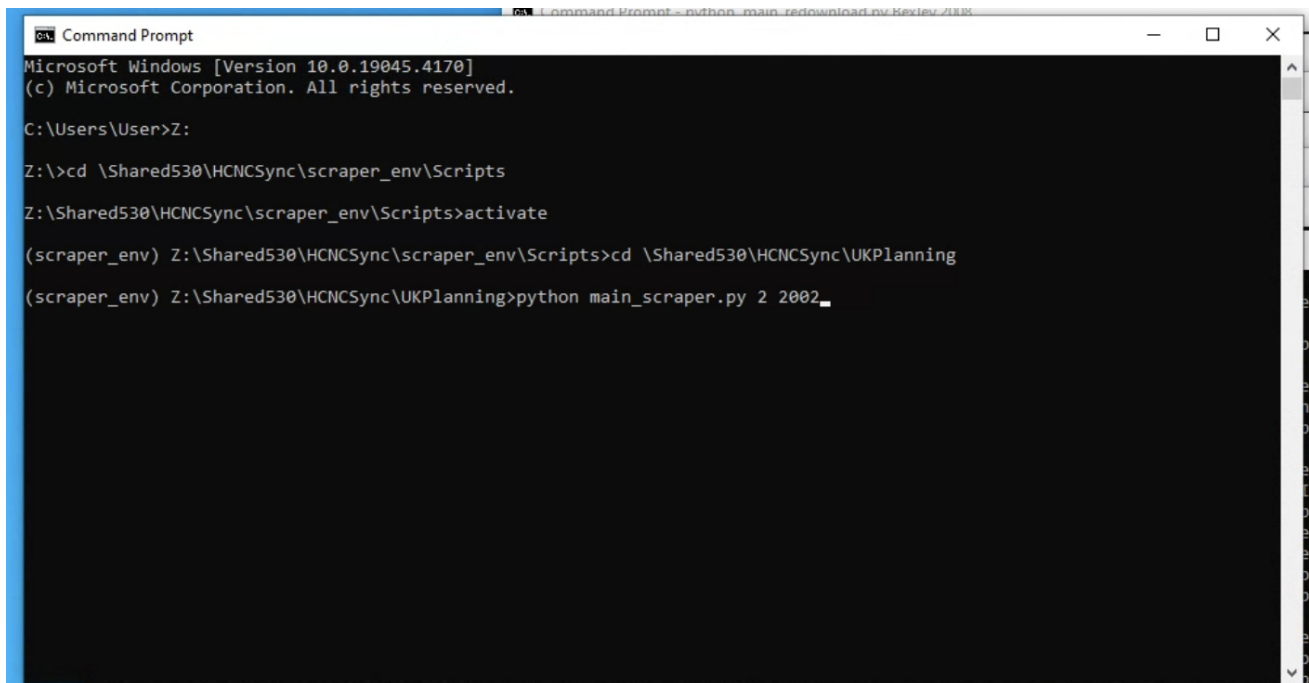


Figure 2: Copy the commands.

1.3. Paste commands to the opened cmd:



```
Command Prompt
Microsoft Windows [Version 10.0.19045.4170]
(c) Microsoft Corporation. All rights reserved.

C:\Users\User>Z:

Z:\>cd \Shared530\HCNCSync\scraper_env\Scripts

Z:\Shared530\HCNCSync\scraper_env\Scripts>activate

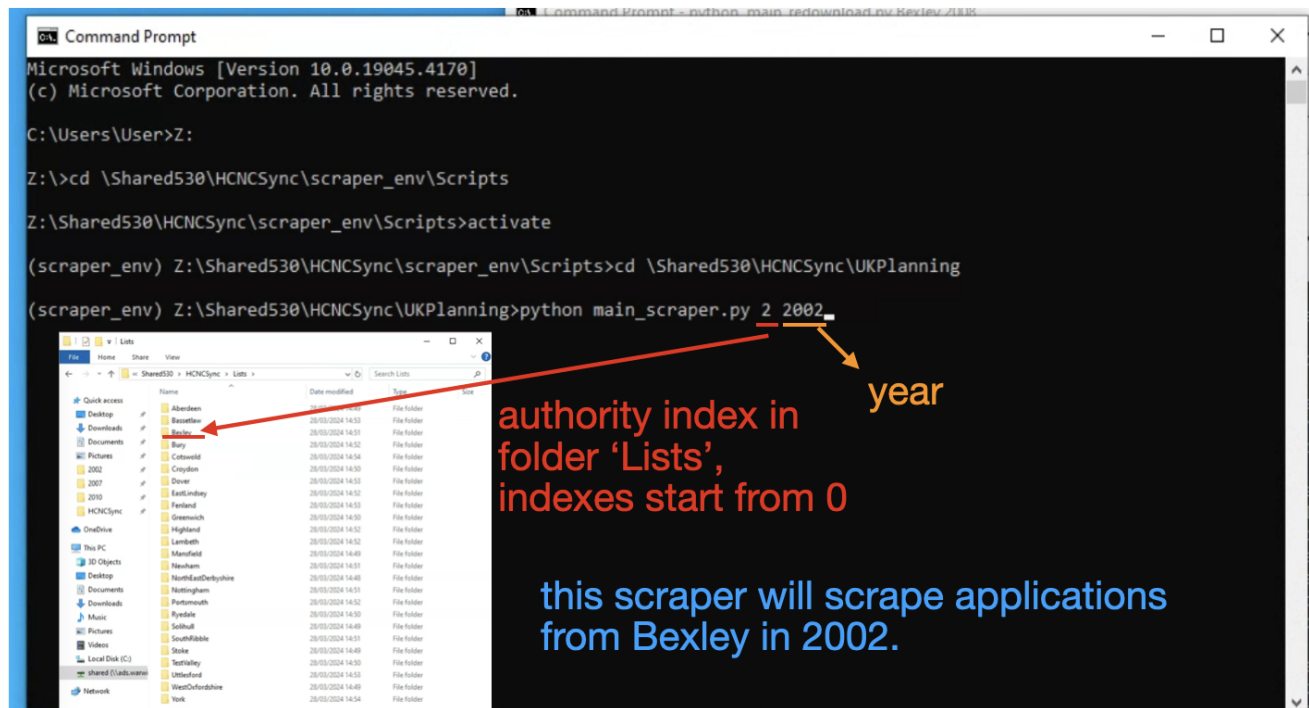
(scraper_env) Z:\Shared530\HCNCSync\scraper_env\Scripts>cd \Shared530\HCNCSync\UKPlanning

(scraper_env) Z:\Shared530\HCNCSync\UKPlanning>python main_scraper.py 2 2002_
```

Figure 3: Paste the commands.

1.4. Modify the last command to run a specific scraper:

Please check [HCNCSync/Lists](#) to ensure the index and year of a specific authority is available before running a scraper. If the authority is not available, download it from Dropbox folder: [HCNCSync/Xunzhao/Application lists \(424 LAs\)](#).



```
Command Prompt
Microsoft Windows [Version 10.0.19045.4170]
(c) Microsoft Corporation. All rights reserved.

C:\Users\User>Z:

Z:\>cd \Shared530\HCNCSync\scraper_env\Scripts

Z:\Shared530\HCNCSync\scraper_env\Scripts>activate

(scraper_env) Z:\Shared530\HCNCSync\scraper_env\Scripts>cd \Shared530\HCNCSync\UKPlanning

(scraper_env) Z:\Shared530\HCNCSync\UKPlanning>python main_scraper.py 2 2002_
```

authority index in folder 'Lists', indexes start from 0

year

this scraper will scrape applications from Bexley in 2002.

Figure 4: Modify the last command to run a specific scraper.

1.5. The scraped data will be stored in [ScrapedApplications/authority_name/year/0.results](#), while the scraped documents will be stored in separate folders: [ScrapedApplications/authority_name/year/application_id](#).

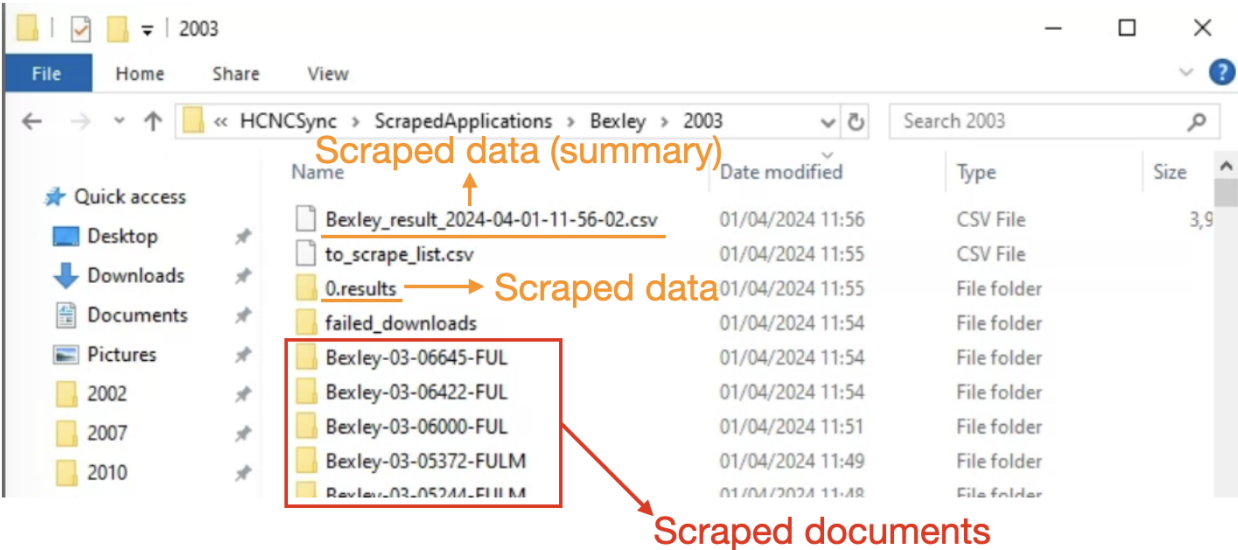


Figure 5: The scraped data and documents.

2 How to re-scrape and re-download

2.1 Re-scrape applications

2.1.1. We need to re-scrape applications for any of the following reasons:

- #Failed# Some applications are unavailable for the moment (i.e. webpage maintenance, poor network connection);
- #Failed# The scraper is crashed due to bugs;
- #Uncompleted# The scraper is interrupted by users.

The indexes of failed applications and uncompleted applications are stored in [ScrapedApplications/authority_name/year/to_scrape_list.csv](#). The list is exemplified in Fig. 6.

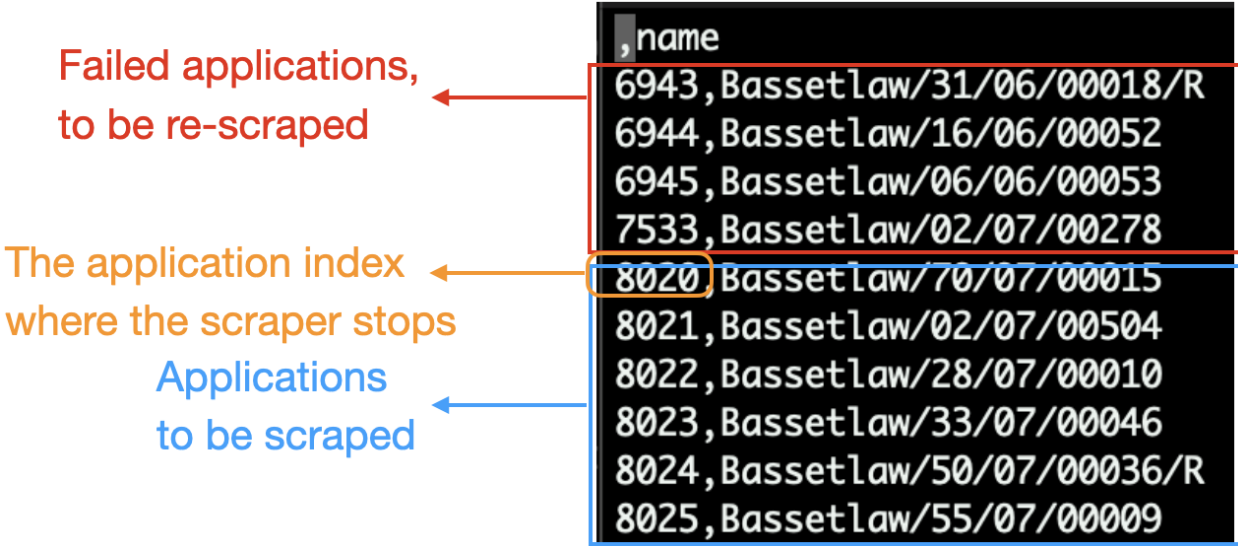


Figure 6: The content of `to_scrape_list.csv`

2.1.2. To re-scrape applications, run the command below (see Section 1.4 for explanations):

```
python main_scraper.py authority_index year
```

The scraper will load [to_scrape_list.csv](#) and scrape all applications in the list until the list is empty.

2.2 Re-download documents

2.2.1. Similarly, some documents could be unavailable when scraping. If the download of a document was failed, the id of corresponding application would be stored in [ScrapedApplications/authority_name/year/failed_downloads](#), as shown in Fig. 7.

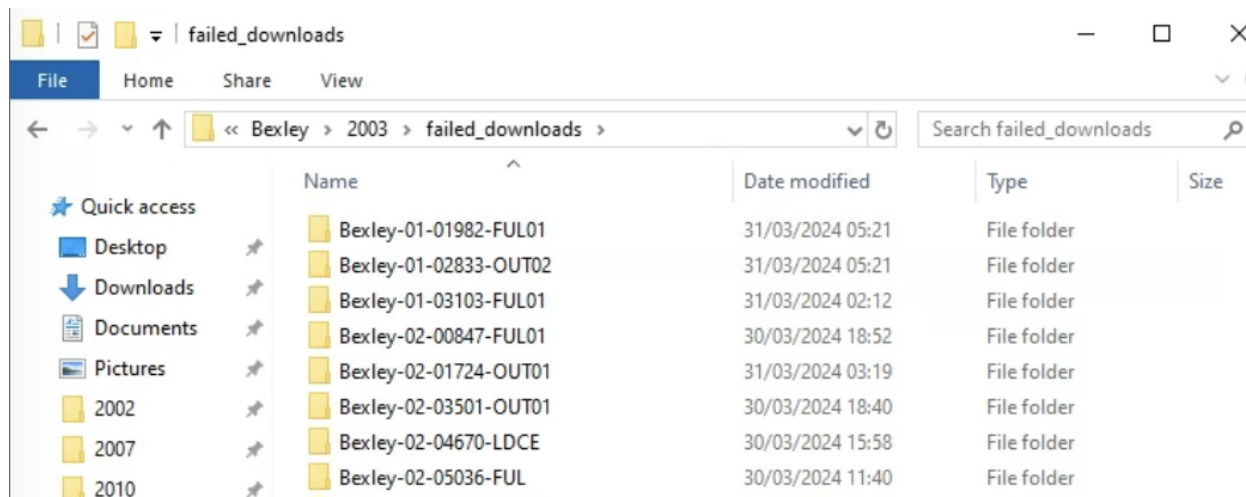


Figure 7: The failed_downloads folder

2.2.2. To re-download documents, run the command below:

```
python main_redownload.py authority_name year
```

An example is available in [HCNCSync/cmd commands.txt](#) (see Fig. 8):

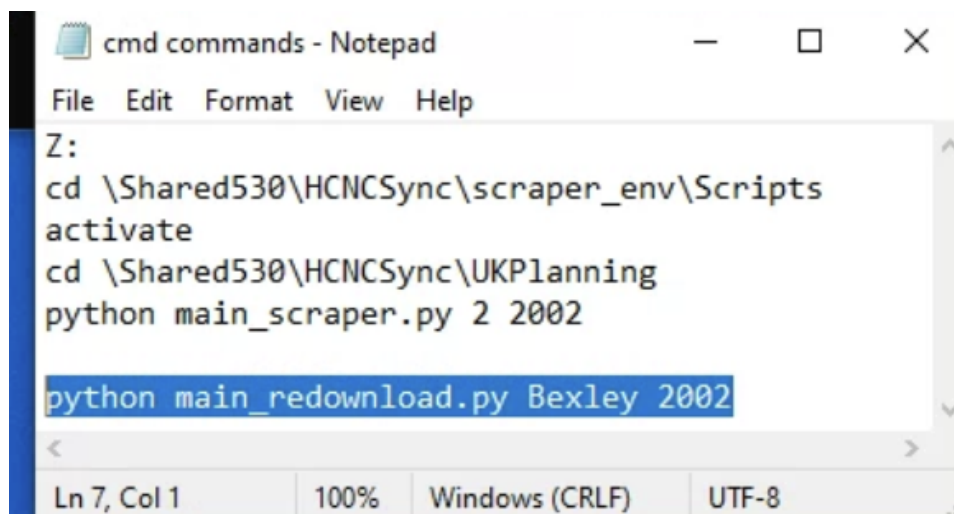


Figure 8: Re-download documents for applications from Bexley in 2002.