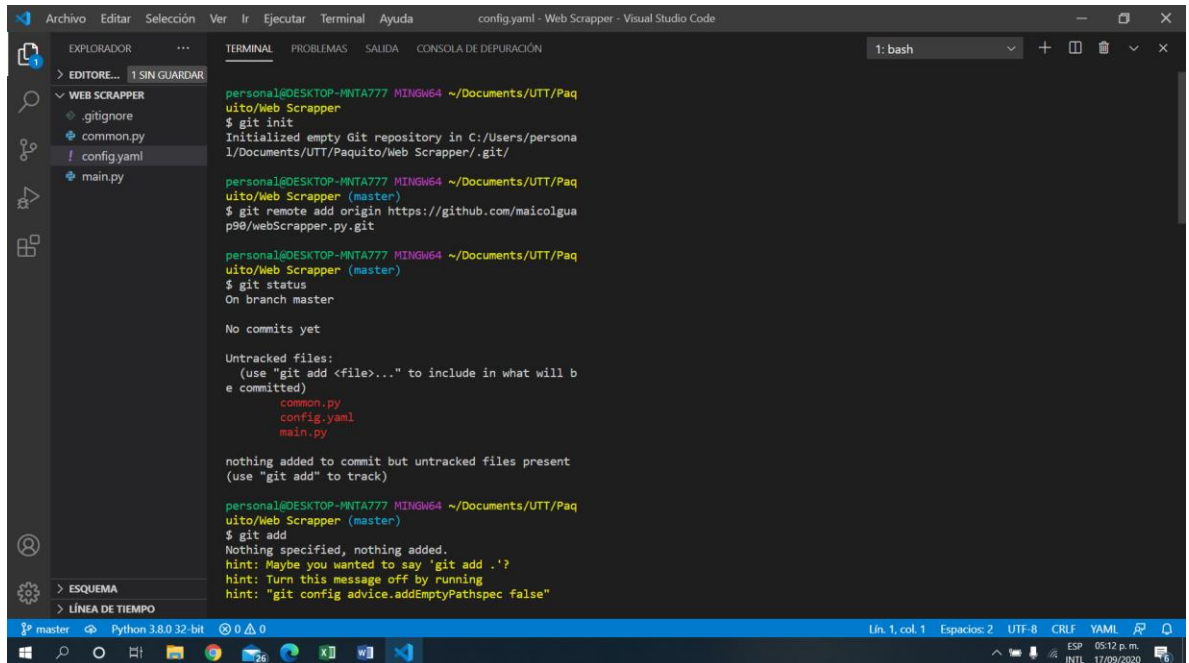


Repositorios y union del local y el remoto



The screenshot shows the Visual Studio Code interface with a terminal window open. The terminal displays the following commands and output:

```
personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper
$ git init
Initialized empty Git repository in C:/Users/persona1/Documents/UTT/Paquito/Web Scrapper/.git/

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git remote add origin https://github.com/maicolgwap90/webScrapper.py.git

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git status
On branch master

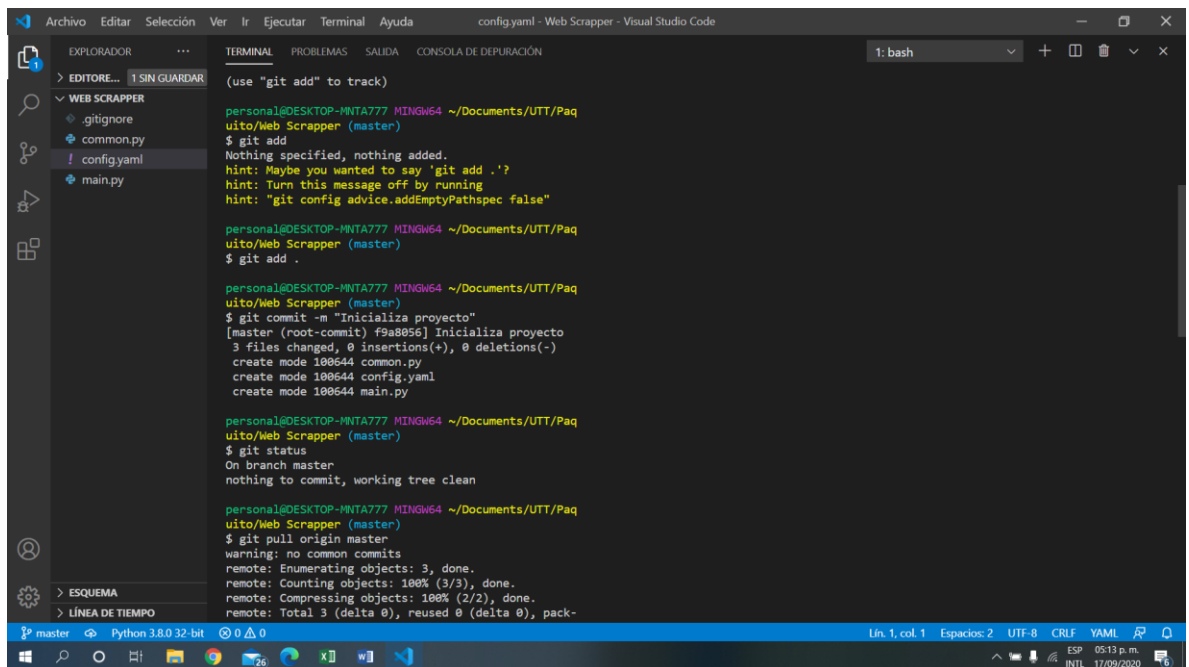
No commits yet

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    common.py
    config.yaml
    main.py

nothing added to commit but untracked files present (use "git add" to track)

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git add
Nothing specified, nothing added.
hint: Maybe you wanted to say 'git add .'?
hint: Turn this message off by running
hint: "git config advice.addEmptyPathsSpec false"
```

The Explorer sidebar on the left shows the project structure with files: .gitignore, common.py, config.yaml, and main.py. The status bar at the bottom indicates the current branch is master, the Python interpreter is 3.8.0 32-bit, and the file encoding is UTF-8.



The screenshot shows the Visual Studio Code interface with a terminal window open, continuing from the previous state. The terminal displays the following commands and output:

```
(use "git add" to track)

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git add
Nothing specified, nothing added.
hint: Maybe you wanted to say 'git add .'?
hint: Turn this message off by running
hint: "git config advice.addEmptyPathsSpec false"

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git add .

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git commit -m "Inicializa proyecto"
[master (root-commit) f9a8056] Inicializa proyecto
3 files changed, 0 insertions(+), 0 deletions(-)
create mode 100644 common.py
create mode 100644 config.yaml
create mode 100644 main.py

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git status
On branch master
nothing to commit, working tree clean

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git pull origin master
warning: no common commits
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (2/2), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-
```

The Explorer sidebar on the left shows the project structure with files: .gitignore, common.py, config.yaml, and main.py. The status bar at the bottom indicates the current branch is master, the Python interpreter is 3.8.0 32-bit, and the file encoding is UTF-8.

The screenshot shows the Visual Studio Code interface with a terminal window open. The terminal displays the following commands and output:

```
personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git pull origin master
warning: no common commits
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (2/2), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (3/3), 1.52 KiB | 103.00 KiB /s, done.
From https://github.com/maicolguap90/webScrapper.py
* branch      master      -> FETCH_HEAD
* [new branch] master      -> origin/master
fatal: refusing to merge unrelated histories

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git pull origin master --allow-unrelated-histories
error: did you mean '--allow-unrelated-histories' (with two dashes)?

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git pull origin master --allow-unrelated-histories
From https://github.com/maicolguap90/webScrapper.py
* branch      master      -> FETCH_HEAD
Merge made by the 'recursive' strategy.
.gitignore | 129 ++++++
1 file changed, 129 insertions(+)
create mode 100644 .gitignore

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git push origin master
```

The Explorer sidebar on the left shows the project structure: WEB SCRAPPER, .gitignore, common.py, config.yaml, and main.py. The status bar at the bottom indicates the current branch is master, Python 3.8.0 32-bit, and the file encoding is UTF-8.

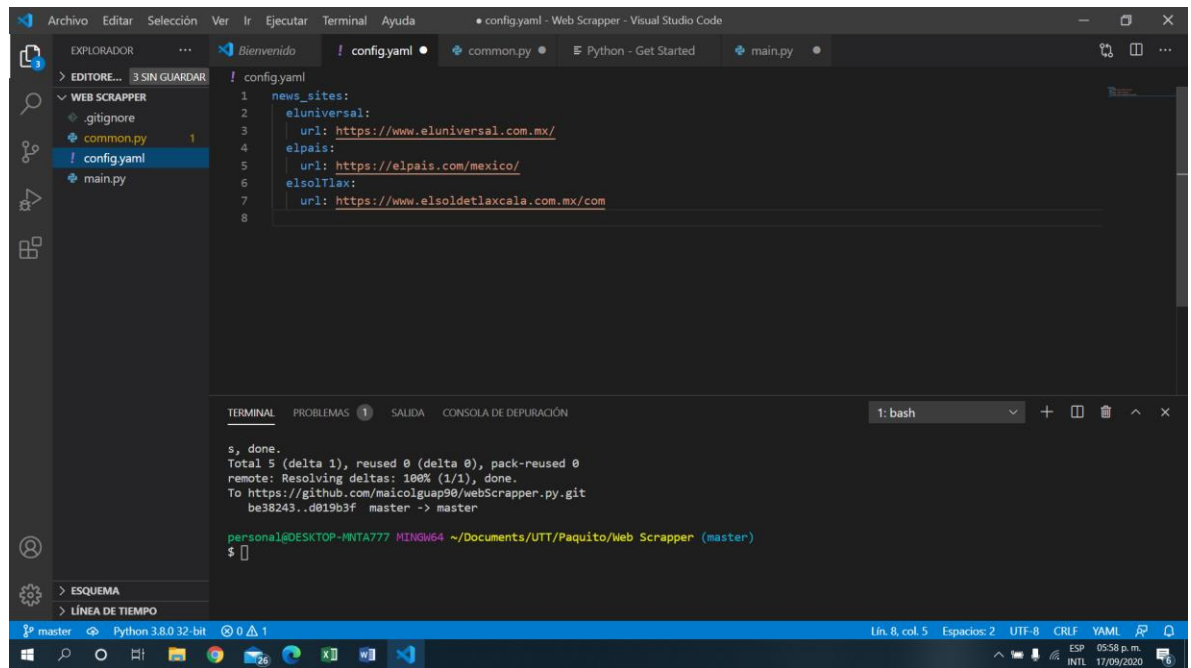
This block contains two screenshots. The left screenshot shows the Visual Studio Code terminal with the following commands and output:

```
personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git push origin master
Enumerating objects: 6, done.
Counting objects: 100% (6/6), done.
Delta compression using up to 4 threads
Writing objects: 100% (5/5), 514 bytes | 128.00 KiB/s, Total 5 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), done.
To https://github.com/maicolguap90/webScrapper.py.git
be38243..d019b3f master -> master

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$
```

The right screenshot shows a web browser displaying the GitHub repository page for [maicolguap90 / webScrapper.py](https://github.com/maicolguap90/webScrapper.py). The page shows the repository name, a green "Crear nueva rama" button, and a list of files: .gitignore, common.py, config.yaml, and main.py. The footer of the page includes copyright information for GitHub, Inc. and links to various GitHub resources.

Continuamos con el proyecto en el código



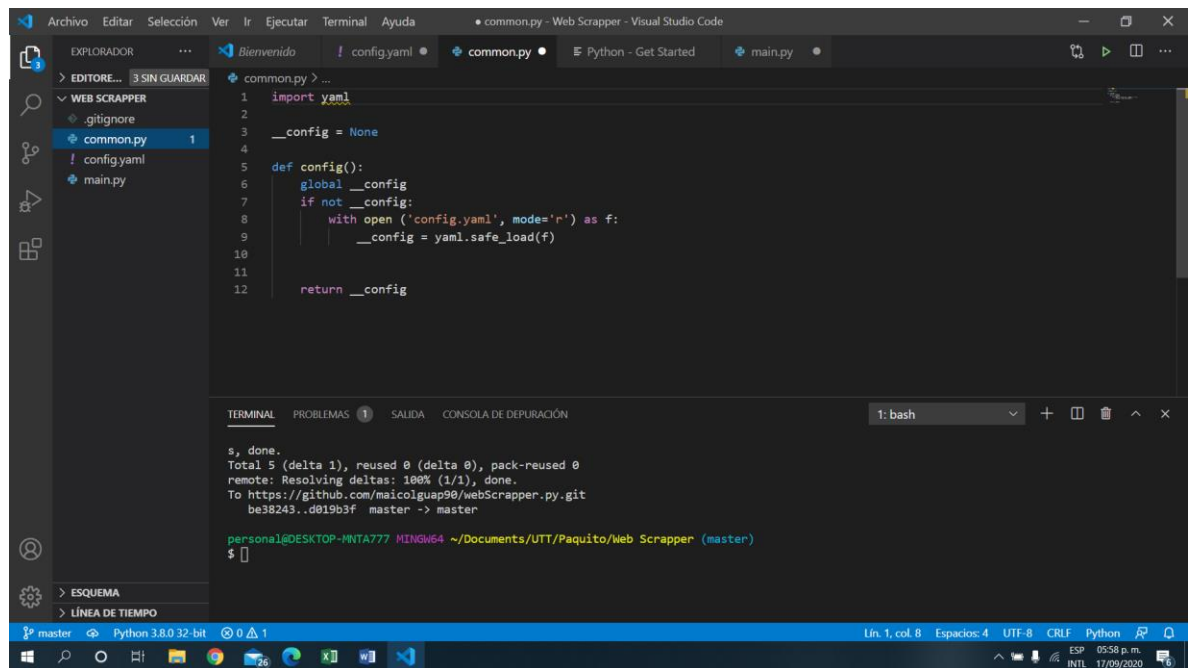
The screenshot shows the Visual Studio Code interface with the `config.yaml` file open in the editor. The file contains a list of news sites with their URLs. The terminal window shows the output of a git command, indicating a successful push to the master branch.

```
config.yaml
1 news_sites:
2   eluniversal:
3     url: https://www.eluniversal.com.mx/
4   elpais:
5     url: https://elpais.com/mexico/
6   elsoldflax:
7     url: https://www.elsoldetlaxcala.com.mx/com
8

TERMINAL
1: bash

s, done.
Total 5 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), done.
To https://github.com/maicolguap90/webScraper.py.git
be38243..d019b3f master -> master

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scraper (master)
$
```



The screenshot shows the Visual Studio Code interface with the `common.py` file open in the editor. The file contains a Python script that loads the configuration from the `config.yaml` file. The terminal window shows the same output as the previous screenshot, indicating a successful git push.

```
common.py
1 import yaml
2
3 __config = None
4
5 def config():
6     global __config
7     if not __config:
8         with open('config.yaml', mode='r') as f:
9             __config = yaml.safe_load(f)
10
11     return __config
12

TERMINAL
1: bash

s, done.
Total 5 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), done.
To https://github.com/maicolguap90/webScraper.py.git
be38243..d019b3f master -> master

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scraper (master)
$
```

The screenshot shows the Visual Studio Code interface with a Python file named `main.py` open. The code is a web scraper that uses `argparse` for command-line arguments and `logging` for output. It reads a configuration file (`config.yaml`) to get the URL of the news site to scrape. The script is currently running in a terminal window.

```
1 import argparse
2 import logging
3 logging.basicConfig(level=logging.INFO)
4
5 from common import config
6
7 logger = logging.getLogger(__name__)
8
9 def _news_scraper(news_site_uid):
10     host = config()['news_sites'][news_site_uid]['url']
11     logging.info(f'Beginning scraper for {host}')
12
13 if __name__ == "__main__":
14     parser = argparse.ArgumentParser()
15
16     news_sites_choices = list(config()['news_sites'].keys())
17     parser.add_argument('news_sites',
18                         help='The news site that you want to scrape',
19                         type=str,
20                         choices=news_sites_choices)
21
22     args = parser.parse_args()
23     _news_scraper(args.news_sites)
24
25
```

The screenshot shows a Windows command prompt window with the following commands and output:

```
C:\Windows\system32\cmd.exe - conda install pandas - conda install pyyaml

The following packages will be downloaded:

package | build | size
-----|-----|-----
pyyaml-5.3.1 | py38he774522_1 | 156 KB
yaml-0.2.5 | he774522_0 | 62 KB
-----|-----|-----
Total: | | 218 KB

The following NEW packages will be INSTALLED:

pyyaml | pkgs/main/win-64::pyyaml-5.3.1-py38he774522_1
yaml | pkgs/main/win-64::yaml-0.2.5-he774522_0

Proceed ([y]/n)? y

Downloading and Extracting Packages
yaml-0.2.5 | 62 KB | ##### 100%
pyyaml-5.3.1 | 156 KB | ##### 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(Higuera) C:\Users\personal\Documents\UITP\Paquito\Web Scraper>python main.py
(Higuera) C:\Users\personal\Documents\UITP\Paquito\Web Scraper>python main.py
usage: main.py [-h] (eluniversal,elpais,elsolflax)
main.py: error: the following arguments are required: news_sites

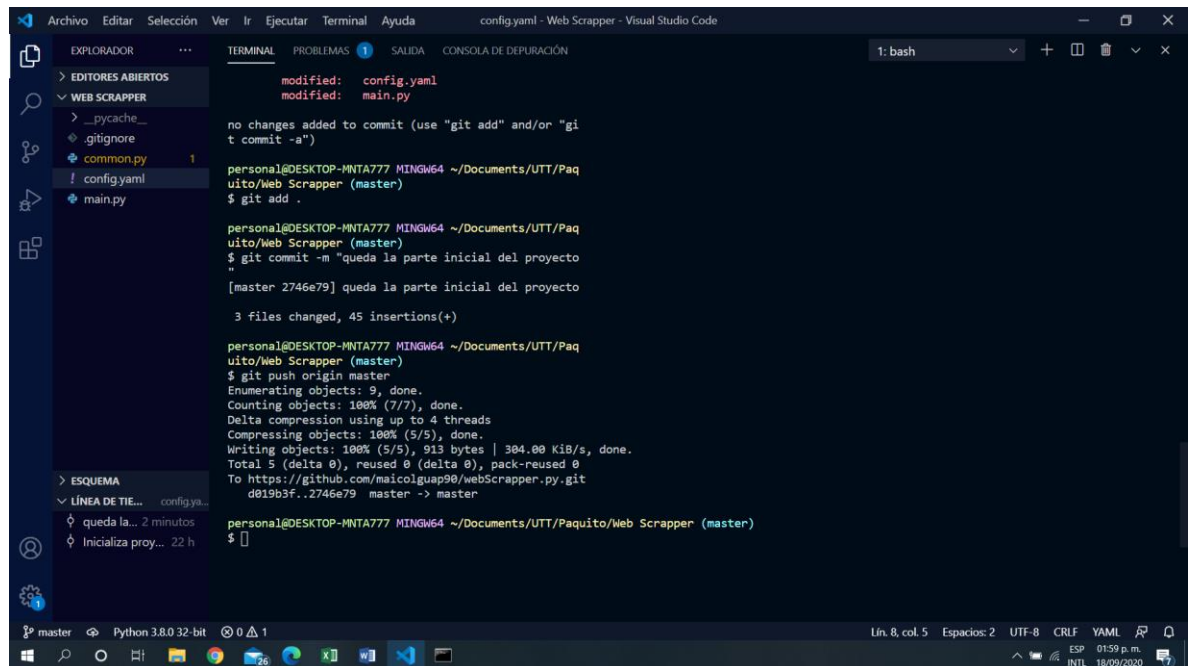
(Higuera) C:\Users\personal\Documents\UITP\Paquito\Web Scraper>python main.py --help
usage: main.py [-h] (eluniversal,elpais,elsolflax)

positional arguments:
  (eluniversal,elpais,elsolflax)
                        The news site that you want to scrape

optional arguments:
  -h, --help            show this help message and exit

(Higuera) C:\Users\personal\Documents\UITP\Paquito\Web Scraper>python main.py eluniversal
INFO:root:Beginning scraper for https://www.eluniversal.com.mx/

(Higuera) C:\Users\personal\Documents\UITP\Paquito\Web Scraper>
```



Visual Studio Code interface showing the terminal output of git commands. The Explorer sidebar on the left shows the file structure: WEB SCRAPPER, .pycache_, .gitignore, common.py, config.yaml, and main.py. The terminal window displays the following commands and output:

```
modified: config.yaml
modified: main.py

no changes added to commit (use "git add" and/or "git commit -a")

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git add .

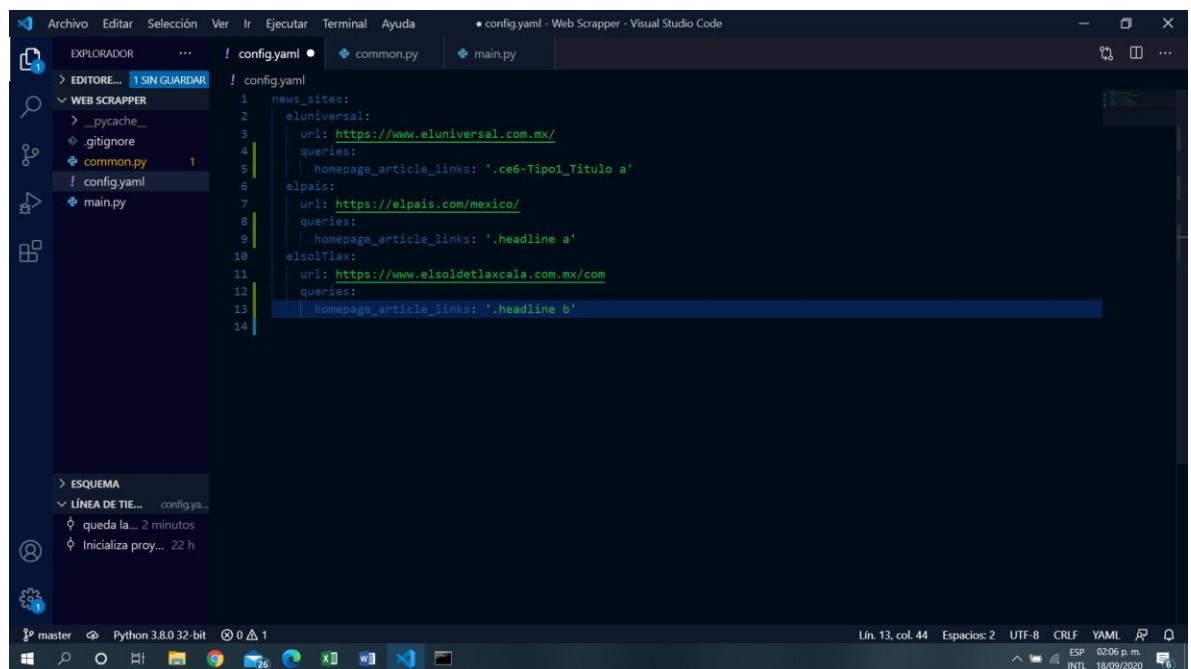
personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git commit -m "queda la parte inicial del proyecto"
[master 2746e79] queda la parte inicial del proyecto

3 files changed, 45 insertions(+)

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$ git push origin master
Enumerating objects: 9, done.
Counting objects: 100% (7/7), done.
Delta compression using up to 4 threads
Compressing objects: 100% (5/5), done.
Writing objects: 100% (5/5), 913 bytes | 304.00 KiB/s, done.
Total 5 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/maicolguap90/webScraper.py.git
d019b3f..2746e79 master -> master

personal@DESKTOP-MNTA777 MINGW64 ~/Documents/UTT/Paquito/Web Scrapper (master)
$
```

The status bar at the bottom indicates the current branch is master, Python 3.8.0 32-bit, and the file encoding is UTF-8.



Visual Studio Code interface showing the config.yaml file in the editor. The Explorer sidebar on the left shows the file structure: WEB SCRAPPER, .pycache_, .gitignore, common.py, config.yaml, and main.py. The editor window displays the following YAML configuration:

```
1 news_sites:
2   eluniversal:
3     url: https://www.eluniversal.com.mx/
4     queries:
5       homepage_article_links: '.ce6-Tipo1_Titulo a'
6   elpais:
7     url: https://elpais.com/mexico/
8     queries:
9       homepage_article_links: '.headline a'
10  elsolflax:
11    url: https://www.elsoldetlaxcala.com.mx/com
12    queries:
13      homepage_article_links: '.headline b'
14
```

The status bar at the bottom indicates the current branch is master, Python 3.8.0 32-bit, and the file encoding is UTF-8.

```
Archivo  Editar  Selección  Ver  Ir  Ejecutar  Terminal  Ayuda  • news_page_objects.py - Web Scraper - Visual Studio Code

! config.yaml  • news_page_objects.py  • common.py  • main.py

news_page_objects.py > ...
4 from common import config
5
6 class HomePage:
7
8     def __init__(self, news_site_uid, url):
9         self._config = config()['news_site']['news_site_uid']
10        self._queries = self._config['queries']
11        self._html = None
12
13        self._visit(url)
14
15
16    def _visit(self, url):
17        response = requests.get(url)
18
19        response.raise_for_status()
20
21        self._html = bs4.BeautifulSoup(response.text, 'html.parser')
22
23    @property
24    def article_links(self):
25        link_list = []
26        for link in self._select(self._queries['homepage_article_links']):
27            if link and link.has_attr('href'):
28                link_list.append(link)
29
30        return set(link['href'] for link in link_list)
31
32    def _select(self, query_string):
33        return self._html.select(query_string)
```

```
Archivo  Editar  Selección  Ver  Ir  Ejecutar  Terminal  Ayuda  • main.py - Web Scraper - Visual Studio Code

! config.yaml  • news_page_objects.py  • common.py  • main.py

main.py > _news_scraper
1 import argparse
2 import logging
3 logging.basicConfig(level=logging.INFO)
4
5 import news_page_objects as news
6 from common import config
7
8 logger = logging.getLogger(__name__)
9
10 def _news_scraper(news_site_uid):
11     host = config()['news_sites'][news_site_uid]['url']
12
13     logging.info(f'Beginning scraper for {host}')
14     homepage = news.HomePage(news_site_uid, host)
15
16     for link in homepage.article_links:
17         print(link)
18
19 if __name__ == "__main__":
20     parser = argparse.ArgumentParser()
21
22     news_sites_choices = list(config()['news_sites'].keys())
23     parser.add_argument('news_sites',
24                         help='The news site that you want to scrape',
25                         type=str,
26                         choices=news_sites_choices)
27
28     args = parser.parse_args()
29     _news_scraper(args.news_sites)
30
```