**Evaluating Capabilities of current online news scraping software**

**Section 1: Motivation**

Two of the most effective open-source news scraping softwares available are [Newspaper3k by Lucas Ou-Yang](https://github.com/codelucas/newspaper/) and [news-please by Felix Hamborg](https://github.com/fhamborg/news-please/). I will attempt to compare these two available options in terms of features, accuracy and speed, and determine the suitability of these scraping libraries in handling:

* Websites of different formats
* Websites in different languages (English, French, Spanish, Portuguese, Russian, Chinese, Indonesian, Korean, Swahili, Arabic)
* Scraping of websites at a high rate (for real-time news coverage)
* Scraping of live news
* Subscription sites
* News articles with multiple pages (e.g. chinadaily)
* Interactive formats (e.g. BMJ)
* Articles which involve clicking to read more (e.g. CNN Arabic)

**Section 2: Comparison**

**Features:**

Shown below is a table comparing the current available features of Newspaper3k and News-please:

|  |  |  |
| --- | --- | --- |
| Feature: | Newspaper3k | news-please |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Capabilities/Speed of article scraping:**

I tested both languages using news websites I had previously gathered in 10 different languages, and evaluated whether they were accurate, and how quickly the libraries could perform the scraping. These are the results:

**Capabilities/Speed of full website scraping:**

I performed a similar test on the ability of each scraper to perform a scrape on a full news website from the homepage. These are the results:

**Appendix: Country codes**

A table to convert the 2/3 letter country codes used in the tables to a country/organisation:

|  |  |
| --- | --- |
| **Country Code** | **Country** |
| **ALG** | Algeria |
| **ARG** | Argentina |
| **AUS** | Australia |
| **BEL** | Belarus |
| **BRA** | Brazil |
| **CAN** | Canada |
| **CBV** | Cabo Verde (Cape Verde) |
| **CHI** | China |
| **COL** | Colombia |
| **DRC** | Democratic Republic of the Congo |
| **EGY** | Egypt |
| **ESP** | Spain |
| **FRA** | France |
| **GBS** | Guinea Bissau |
| **HK** | Hong Kong |
| **IDN** | Indonesia |
| **IRE** | Republic of Ireland |
| **IRQ** | Iraq |
| **ISR** | Israel |
| **JAM** | Jamaica |
| **KEN** | Kenya |
| **KOR** | South Korea |
| **KSA** | Saudi Arabia |
| **LEB** | Lebanon |
| **MEX** | Mexico |
| **MOR** | Morocco |
| **MOZ** | Mozambique |
| **NZ** | New Zealand |
| **POR** | Portugal |
| **QAT** | Qatar |
| **RUS** | Russia |
| **SNG** | Singapore |
| **SUD** | Sudan |
| **SUI** | Switzerland |
| **T&T** | Trinidad and Tobago |
| **TAI** | Taiwan |
| **TNZ** | Tanzania |
| **TUN** | Tunisia |
| **UAE** | United Arab Emirates |
| **UGA** | Uganda |
| **UK** | United Kingdom |
| **UKR** | Ukraine |
| **UN** | United Nations |
| **USA** | United States of America |
| **VEN** | Venezuela |