Report

for eurostat
European Commission

Task 2

Data Retrieval

Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry

Quality, methodology and research Lot 1: Methodological support

Specific Contract N° 000075 ESTAT N° 2020.2044

under the Framework Contract 2018.0086

Draft v2, July 2021

GOPA
WORLDWIDE CONSULTANTS

in joint-venture with

GOPA
LUXEMBOURG

GOPA Consultants
Hindenburgring 18, 61348 Bad Homburg, Germany
Phone +49 6172 930-521 Fax: +49 6172 930-130 Email: statistics@gopa.de

Table of content

Tables

Figures

Annexes

Annex 1	Queries used in different versions for Ireland
Annex 2	Flowcharts of the scripts developed
Annex 3	Content of Ini-file for Ireland
Annex 4	Part of final output file produced

Abbreviations

API	Application Programming Interface
RPAS	Remote Piloted Aerial System
UAS	Unmanned Aircraft System
UAV	Unmanned Aerial Vehicle
URL	Uniform Resource Locator
WWW	World Wide Web

1 Introduction

This report is prepared in the frame of the specific contract on 'Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry' (specific Contract N° 000075 Ref. N° 2020.2044 under framework contract ref. 2018.0086) implemented by GOPA with the collaboration of data scientists and researchers from Statistics Netherlands and Universitat Politècnica de València.

The aim of the project is to further extend the capabilities of retrieving information on businesses from the internet, particularly for the use case of drone businesses, building on previous research on web scraping of business information in the context of official statistics, and with a view to generalising the approach to new emerging economic trends. More in detail, the main objective of the project is to develop a methodology and tools to retrieve information from the World Wide Web (www) for business based in EU countries that have their main activity in the civil drones sector. The project is a research study that will lead to experimental statistics for a selection of five countries.

This report provides a description of the data retrieval approach developed, the implementation choices made and contains an overview of the first results obtained. The findings for the countries Ireland, the Netherlands, Spain and Germany are included in this report.

Following this short introductory and a section on the composition of URLs, this report is organised as follows:

- Chapter 2 focuses on the scripts and methodology developed to collect links to websites of drone companies in a specific country for a specific language
- Chapter 3 provides an overview of the results obtained for each country and discusses the classification results for Spain.
- Chapter 4 draws first conclusions, suggests steps to improve the approach developed and indicates future work.

1.1 URL composition explained

For this report it is important to understand the structure of a Uniform Resource Locater (a URL). A URL refers to a specific location of the World Wide Web. The composition of a URL is illustrated with the example shown below:

https://www.gopa.lu/category/home-recent/#drone-industry

- The first part of the URL is the **scheme-part**; for web pages it is usually is *http://* or *https://*. This is often followed by 'www.'; this is identified as the subdomain part. A 'www' subdomain can be removed from a URL without affecting its location.
- The next part is the **second-level domain**; it is 'gopa' in the example. This part of the domain usually refers to the organisation that registered the domain name. From the project's point of view this is the most important part of a URL as it -very likely- specifically indicates websites of individual companies.
- The second-level domain is followed by the **top-domain**, which is 'lu' in this case. This abbreviation refers to web pages located in Luxembourg. There is a whole list of country specific top-domains on the web

(Wiki, 2021). Other non-country specific top-domains often used are ‘com’, for companies, and ‘org’ for organizations, etc.

- Anything referred to after the slash following the top-domain, usually, indicates a subdirectory on the server. The example URL refers to the subdirectory “/category/home-recent”. This is the location of the html-file on the gopa-server.
- Sometime other additions are included in a URL, such as “#drone-industry” in the example. The hash-sign indicates a specific location on the web page referred to. For search engines, often a question mark is included to indicate the beginning of the query of words provided.

2 Data collection methodology

In this chapter the most important decisions made during the development of the scripts used to search the web are discussed. Because the (sub)topics and decisions made affect one another, it has been challenging to find a specific sequence in which the topics should be discussed. In the end, it was decided to discuss them in the order shown below. The dependencies are indicated by referring to the other sections when discussing each topic. In addition, since the queries were also improved during the writing of this report, results of two different versions of the queries are discussed. The first version of the queries has been applied to Ireland, the Netherlands, Spain and Germany. At the moment of writing, the second version has been applied to Spain partially.

2.1 Search queries

After consulting the researchers involved in the project and experimenting with different search words and query compositions on various search engines, a decision was made on the composition of the search queries used for the two main approaches developed (see section 2.3). For each country, queries in the dominant language and queries in the English language were developed. The exception was Ireland; here only English queries were used. As an illustration, the queries used for Ireland, for both versions, are included in Annex 1. In principle, each search query is composed of three parts.

- The first part of the search query contained *the word drone and/or its acronyms.* Depending on the native language of the country searched, the word 'drone' is replaced by the native word used; e.g. drohne in Germany and dron in Spain. The acronyms used for drone are the same in each query for each country, i.e. uas, uav and rpas. In the first version of the queries the drone words are surrounded by brackets () and separated by a space and the word OR. For example: (drone OR uas OR uav OR rpas). In the second version the queries used by the first script only contains a single drone acronym.
- The second part of the query contains the word (or words) indicating either a *company*, *a type of activity* or *a synonym of member*. Examples of the first are: company, business or shop and of the second are: pilot, training, manufacture, service provider, etc. These are examples of words used when one searches for individual web sites of Drone companies. In the third case, web sites with lists of drone companies are searched for; hence words such as member, register, list, association, etc. are included. For non-English queries, all words are translated into those of the native language used. One can imagine that, by discerning even more types of activity, even more specific search queries could be used.
- The last part of the search query contains the *name of the country* searched for in the language used. For the four countries studied in English these are Ireland, Netherlands, Spain and Germany, respectively. In their native language these are: Ireland, Nederland, España and Deutschland.

Advantage of the three part construction of the queries is that they can be easily adjusted to investigate other topics. During the study, it was found that the Bing search engine was unable to deal with queries containing letters with diacritical marks (i.e. special characters); such as the ñ in the word España. When the findings for queries with and without diacritical marks were compared, on the other search engines, no obvious differences were found. Hence, it was decided to remove all diacritical marks in all the queries used. At the moment, two search approaches are considered, the

first searches for individual websites of Drone companies (implemented in Script 1) and the other searches for web sites containing lists of drone company websites (implemented in Script 2). Other search approaches have been considered but were found not to provide any additional new information (see Gopa, 2021).

2.2 Search engines

There are various ways to search the web. The most often used search engines, also known as the 'big four'[1], are Google (google.com), Bing (bing.com), DuckDuckGo (duckduckgo.com) and Yahoo (yahoo.com) (Biggs, 2021: Reliablesoft, 2021). These are also the most important search engines when trying to find companies in Europe. Code was developed to extract links from each of these search engines for the queries used (see section 2.1). In addition, Python libraries were found and included in the code that also enabled using those search engines. By opting for this approach, one can easily switch and access a search engine that cannot be reached by the first approach. This sometimes occurs because, a considerable number of, search engines want to block -any suspicious- automated access as much as possible. In addition, other search engines, namely Ask (ask.com), AOL (aol.com), Baidu (baiduinenglish.com) and Yandex (yandex.com), were also tested. AOL and Ask provided interesting results and, hence, were additionally included in the code. The search engines Baidu and Yandex were also explored and (attempted) to be implemented in code. The first is the most popular search engine in China and the other one is most popular in Russia. During this work it was, however, found that the results obtained did not really focus on the European countries studied and that access was difficult to maintain. It was, therefore, decided not to include Baidu and Yandex. As a result, a total of six different search engines could be used; e.g. Google, Bing, DuckDuckGo, Yahoo, AOL and Ask. When Google and Yahoo are used, the search page specific for the country is selected; e.g. for Spain google.es is accessed and es.search.yahoo.com is used for Yahoo. Such an option was not available for the other search engines.

One may wonder what the benefit is of using six different search engines when searching the Web? An example reveals the importance of this. In Table 1 the overlap between the domain names of the links provided by applying all 48 version 2 search queries of Script 1 for Spain is shown. In this table, the results for both the English and Spanish queries are combined (more on that below). For each search engine, the maximum number of organic search results (links/URLs) returned was collected; this was done by extracting all URLs on all pages available and ignoring the commercial links as much as possible. Depending on the search engine and how it reacted when used a lot, the IP-address from which these searches were conducted was kept the same or not[2]. It is important to notice that -in practice- much less links were obtained compared to the huge number of links reported on the first page by (some of) the search engines. For example, Google reported 3.240.000 results on the page of the first search query for Ireland, while Bing states on its first page of that particular query that 326.000 results were found. This number was actually found to be much lower in practice; often only up to 400-500 links could be extracted. This number varied considerably per search engine. Usually, Ask provided the smallest number of links (~55) and AOL and Yahoo the largest (400-550) per query. After running script 1 for all 48 queries for Spain on all six search engines in both languages, the links found were combined per search engine and deduplicated. When comparing these 'raw' (unprocessed) search results between the search engines, an unexpected,

[1] The exact composition of the top search engines varies over the years. At the moment of writing Google, Bing, Yahoo and DuckDuckGo are the most important ones from the perspective of our study.

[2] Preferably a VPN-connection was used with an IP-address in Brussels. However, if it was found that the search engine started blocking the queries, random VPN locations, all over the world, were used for the IP-address in between queries.

fairly low amount of overlap was observed. This was found to be partly due to differences in the subdirectory and additional part of the links and could, therefore, be improved. This was done by only focussing on the combination of secondary domain names and the top-domains returned; e.g. “dronesbarcelona.es” instead of the complete link “https://www.dronesbarcelona.es/prensa”. This had the additional advantage that this comparison focussed on the part of the link that is most logically indicative for a company website. These results of preprocessing the links are shown in Table 1 for Spain. The results obtained with version 1 of the scripts for all countries confirm similar differences for the other countries studied.

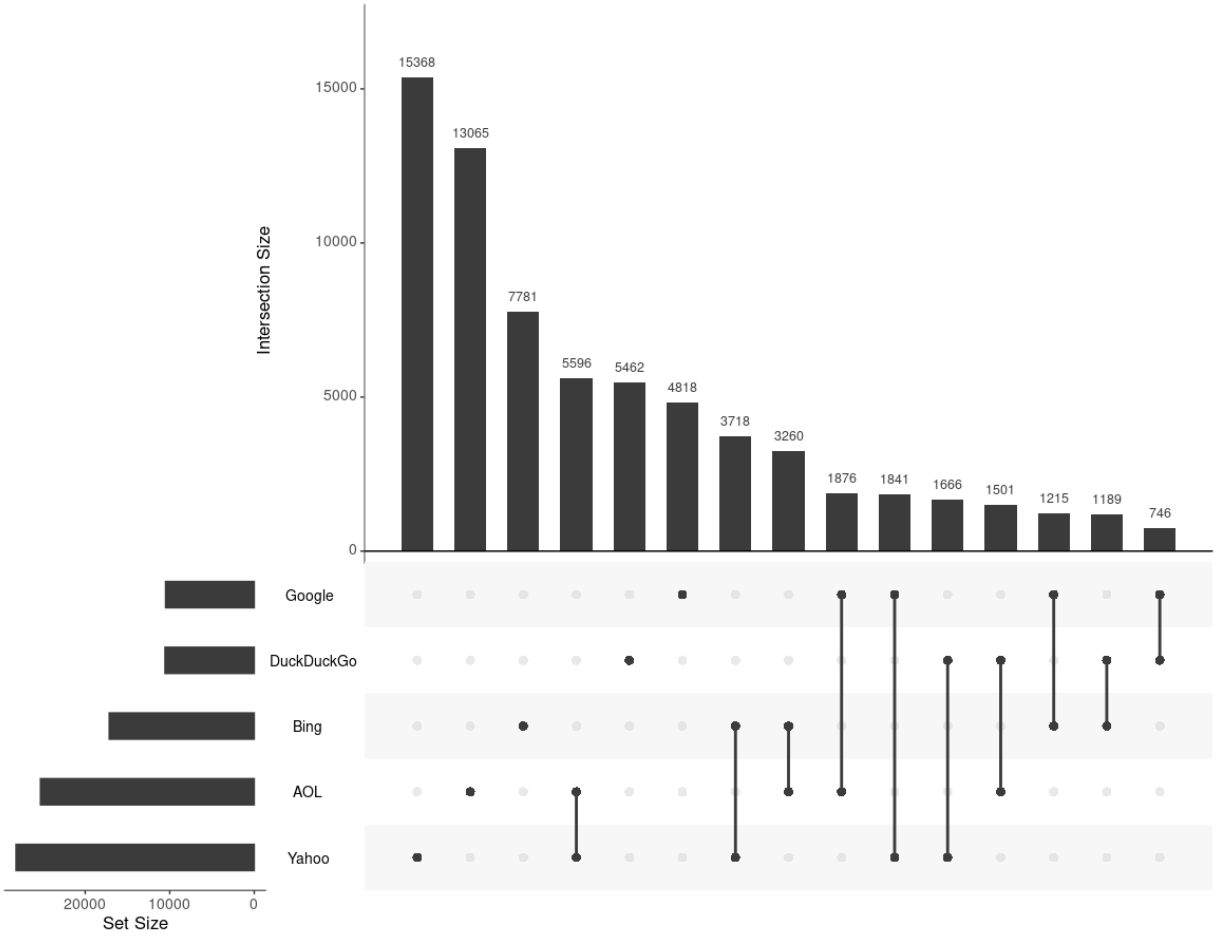
Table 1: Overlap between the secondary and top domain names of the URLs found by the six search engines for all Script 1 queries (version 2) for Spain (both English and Spanish search results are combined)

Search Engines	AOL	Ask	Bing	DuckDuckGo	Google	Yahoo
AOL	(12.055)	1.010	3.260	1.501	1.876	5.596
Ask	1.010	(1.473)	749	529	1.179	999
Bing	3.260	749	(7.032)	1.189	1.215	3.718
DuckDuckGo	1.501	529	1.189	(4.933)	746	1.666
Google	1.876	1.179	1.215	746	(3.639)	1.841
Yahoo	5.596	999	3.718	1.666	1.841	(14.369)

The results in Table 1 indicate that clearly different results are obtained for the different search engines. In the diagonal, between brackets, the number of unique results is shown for each search engine. Overall, a total of 27.712 unique combinations of secondary and top-domain names were found. From Table 1 it’s clear that Yahoo returned most unique combinations; i.e. 14.369. This was followed by AOL and Bing, with 12.055 and 7.032 unique combinations respectively. Next are DuckDuckGo, Google and Ask, with 4.933, 3.639 and 1.473 results.

The two search engines that provide the most similar findings are AOL and Yahoo; an overlap of 5.596. Next are Bing and Yahoo, with an overlap of 3.718, followed by AOL and Bing, with an overlap of 3.260. An alternative way to visualize the overlap between the various search results is by creating an upset-plot; this is an alternative for a Venn-diagram when a dataset is composed of more than three subsets. The upset plot for the data in Table 1 is shown in Figure 1. Here, the large number of findings from Yahoo, compared to the other search engines, becomes immediately apparent. However, the reader needs to realize that these results do not have to indicate that the search engine that returns the most links does not have to find the majority of the Spanish drone websites. The latter is discussed in section 3.1.

Figure 1. Upset plot of the overlap of the secondary and top domain names found by the six search engines for Spain



Very similar findings, to those shown in Figure 1 -but not identical-, are found when the importance of each search engine is based on the total number of unique results provided. Subsequently, those links are removed and the remainder is determined for the links provided by the other search engines, etc. Here, Yahoo is the most important contributor (14.369 links), followed by AOL (6.459), DuckDuckGo (2.965), Bing (2.537), Google (1.277) and Ask (270). Compared to the overall number of domain names found, this makes DuckDuckGo more important than Bing. Google and Ask provide the least number of unique links (see also section 3.1).

All in all, from the results described above it is clear that, to maximize the number of links found, **the findings of all search engines need to be combined**. Because of access issues during this work with some of the search engines (access was regularly blocked), API access to Google was bought by one of the Statistics Netherlands researchers to enable continuation of the work for the remainder of the project. Here, the service provided by Valueserp.com was used as it was found to be the cheapest. Other paid services were checked but were found not be satisfactory as many of them lacked paid access to the majority of the search engines (paid access to DuckDuckGo is a challenge) and those that provided access did not enable scraping multiple pages of the same query (via their APIl). Hence, only one alternative remained: scrape via a VPN-connection.

For most of the other search engines, the IP-address of the connection for each query needed to be changed to prevent blocking; this was particularly relevant for DuckDuckGo, Yahoo, AOL and Ask. To enable this, the VPN connection

provided by Surfshark was used (one of the Statistics Netherlands researchers already had a contract there). During this work it was found that the Bing search engine could be accessed without changing the IP-address as was also the case for the paid Google access. For the other search engines, VPN-based with IP-rotation between queries was needed to perform the first two scraping steps (see section 2.1 and Annex 1) without issues.

2.3 Sequence of steps

In the first deliverable of this project (GOPA, 2021, section 5) various ways to collect data from the internet by using search engines on Drone companies in various countries were envisaged and tested. The most promising approaches were:

1. Collect URLs of individual Drone companies with search words
2. Collect URLs of web sites that provide overviews of Drone companies

These are the two starting points for searching the web. Let it be clear, that all other suggestions made in the previous report, such as using drone images (Gopa, 2021), did either not work or did not provide more relevant information. However, during the work described in this report it was found that, apart from links to web pages (html-pages), also links to PDF-documents were obtained. Often those PDF-files contained (a section with) links to web sites or contained email addresses of potential Drone companies. This indicated the need to not only extract links from web pages but also from PDF-files; both web-links and domain names from email address. Hence, it was decided that the first two steps needed to provide both a list of links to web pages and a list of links to PDF-files. The PDF-files referred to, subsequently, needed to be downloaded from the web after which any links to web domains could be extracted. Hence, the following three steps emerged:

1. Collect URLs of individual Drone companies with search words (web- and PDF-links)
2. Collect URLs of web sites that provide overviews of Drone companies (web- and PDF-links)
3. Extract URLs from the PDF-files found in steps 1 and 2

The flowchart of the scripts created for these tasks are included in Annex 2. Input of the first two scripts are the country and language used and the search queries specific to each task (see Annex 1). The second script needs additional input, these are: a list of words indicative for the country, words indicative for drones and words indicative for being a member of an organization (see Annex 3 for the Ireland example). Also, the maximum number of subpages scraped per domain and a limit to the number of search results are provided. The first is used to select and extract as much information as possible from the sites found to be indicative for providing overviews of Drone companies and the latter was used to prevent that a huge number of web pages needed to be checked and scraped. These settings can all be adjusted via the country and language specific Ini-file. An example of the Ini-file for Ireland is included in Annex 3; it contains a main section and sections with settings specific for each script. The Ini-file must be imported by each script upon execution. For the PDF-script (Script 3), the PDF-link files produced in step 1 and 2, words indicative for the country studied and words for drones are required. The output of script 3 is a list of URLs extracted from PDF-files. After script 3 has run, a total of 4 URL lists are available and used by the fourth script. The flow charts in Annex 2 make clear how each scripts works and what the input and output are. For the readers convenience, Table 2 provides an overview of the in- and output of each script, their dependencies and an indication of their runtime (for the version 2 queries for scripts 1 and 2).

Table 2: Overview of the scripts developed, input, output and more

Script	Input	Output	Script dep.	Run-time (day)
1. Find individual drone websites	- Ini-file (with country and language settings and queries) (Can be run on all or individual search engines)	- Potential links to drone websites - Links to PDF-files - Log files	None	1-2
2. Find drone overview websites	- Ini-file (with country and language settings, queries and website scrape depth) - Lists of words (in Ini-file) (Uses results of all search engines)	- Potential links to drone websites (from overview sites) - Potential links to drone websites (not from overview sites) - Links to PDF-files - Log file	None	~0.5
3. PDF-file link extraction	- Ini-file (with country and language settings) - Lists of words (in Ini-file) - PDF-link files from Scripts 1 and 2	- Potential links to drone websites (including those derived from email addresses) - Log file	Script 1 and 2	~0.5
4a. Cleaning and social media check	- Ini-file (with country and language settings) - Domain extension and country names list (in py-file) - File with country and city names - Files with potential links from Scripts 1, 2 and 3 - Website visit emulation script	- Potential links to drone websites probably in the country (enriched and cleaned) - Log file	Script 1, 2 and 3	~0.5
4b. Content and location check: Identifying drone websites	- Ini-file (with country and language settings and selection criteria) - Domain extension and country names list (in py-file) - File with country and city names - File with potential links from Scripts 4a - Website visit emulation script	- CSV-file with results for websites located in the country or with an unknown location, including drone website indication information - Log file	Script 4a	2-3

Report

In the fourth and final step, the lists of URLs produced in the previous steps are combined, deduplicated and checked. During the checking stage, it was found that a considerable number of links, sometimes up to 10%, were found to refer to social media platforms; such as Twitter, Facebook, LinkedIn, Instagram and Pinterest. Apart from links to individual messages, these can also point to company specific accounts on those platforms. The latter are interesting as these pages may contain i) information on the location of the company and could -potentially- ii) contain the URL of a company's actual web-page (the 'www-page'). Hence, a social media specific check was included with the aim to find the location of the 'account' and extract the link to its www-page (if available). This was a challenging task, as social media platforms tend to spend much effort in blocking scripted access; this is especially the case for Facebook. However, anyone with an account on such a platform can always visit it by using a browser. This provided the solution. By using an actual browser, guided by a script emulating user-activity, the pages referred to by social media links –that did not point to individual messages- could be visited and collected. Any accounts that were located in the country studied or had an unknown (undetermined) location, were additionally checked for containing a URL (to the www-page) of that account. If such a link was found and it was not already detected by scripts 1 and 2 it was added to the list of URLs. The latter proved to be the cases for around 2% of the social media links found. Because this task needed to be performed prior to the more detailed studying of the URLs found, a separate script (script 4a) was created for this task. It also included most of the checks performed to remove - as early in the process as possible - the most non-relevant links. Input for this script were the 4 URL containing files produced by scripts 1,2 and 3, lists of words indicative for the country, drones and members and a list of names of the (largest) cities in a country; both in their native language and, when applicable, the name of the city in English. The latter was done as some cities have a different name in English, examples are the Dutch city of 'Den Haag', named 'The Hague' in English, and the German city of 'München' which is called 'Munich' in English. Wikipedia was used to construct these lists and to find the English names of the largest cities (an alternative is geonames.org). More on the exact steps used in the location search is described below.

After script 4a has run, the large list of unique URLs obtained needed to be checked thoroughly. In all cases for all countries, this step took the longest to perform; although running version 2 of the queries with script 1 comes close. First, it checked if the website still existed. If that was the case, the site was visited and its location and content were investigated. Both the location (is it located in the country studied?) and content of the website (is it a website of a Drone company?) were attempted to be determined. For the location search, a stepwise approach was used, that resulted in either: i) deciding that the website was located in the country studied (Yes), ii) deciding that the website was NOT located in the country studied (No), or iii) not being able to determine its exact location (Unknown). In this process, a a sequential series of steps is applied which stops when the answer is either True or False or when the end is reached; than the answer is Unknown. The sequence of steps used is:

1. Check the URL for including the top-domain name of the country studied (e.g. .ie for Ireland) - **True**
2. Check the URL for including the top-domain of another country (e.g. .nl when studying Ireland) - **False** (requires a list of country specific top-domains)
3. Check the text of the web page for words specific for the country studied or its language (such as Ireland, Eire and Irish for Ireland), - **True**
4. Check the text of the web page for words specific for another country in the world (e.g. Netherlands when studying Ireland) - **False** (requires a list of country names)
5. Check the text of the web page for city names specific for the county studied (such as the city of Limerick in Ireland) – **True** (requires a list of city names in a country)

6. Check the web page for links to ‘contact’, ‘about us’ and ‘who are we’ pages (or similar), scrape those pages and subsequently apply steps 3,4 and 5. Up to 10 pages are visited. - **True** or **False**
7. Check the web page for links to social media accounts, visit those pages and subsequently apply steps 3,4 and 5. Up to 10 pages are visited. - **True** or **False**
8. Decide that the location of a web page cannot be determined with certainty - **Unknown**

When a webpage is identified as located in the country studied or as unknown, the text of the page originally visited (usually the main page of a domain) is studied in more detail. Here the occurrences of drone, aerial, and water synonyms and words indicative for businesses and contact pages are counted. In addition, the top 10 words of the text are extracted, after removal of stop words specific for the language, and stored. This results in a file containing the original URL visited, the actual URL visited, the result of locating the country (True, False or Unknown), the scores for the various word categories (drone, aerial, water, business, contact), the top 10 words and the action taken by the script (e.g. web page does not exist etc.). At the moment of writing, the fact that drone words or its synonyms are included on a page are used as an indication that the web page could belong to a company active in the Drone industry. Future work focusses on using the scores of the various word categories to develop a word-based classifier to identify Drone company websites. This is a challenging task because such a classifier should make use of the words and their frequency for different languages. The overall process followed by script 4a and 4b is shown in Annex 2. A part of the CSV-output file of script 4b is included in Annex 4.

2.4 Hard- and software requirements

The scripts are written in Python 3.7, with exception of the page visit browser emulation script which was written in bash. All scripts are started at the command line and must –at least- include the name of the Ini-file to be loaded; for example: python3 Script2.py ES_es.ini. In addition, Script 1 has the option to indicate a specific search engine to be used during that specific run; for example: python3 Script1.py ES_es.ini B1. The latter indicates that only the Bing search engine will be used to collect links. Progress of the script is stored in a log-file and shown on screen. The scripts have been developed on computers using Ubuntu 20.04 as an operating system and have not been tested on non-Linux systems. For the scripts to run properly, apart from Python3 (preferably Anaconda), the following (Ubuntu) libraries need to be installed: libgl1-mesa-glx, libegl1-mesa libxrandr2, libxrandr2, libxss1, libxcursor1, libxcomposite1, libasound2, libxi6, libxtst6 and libgconf-2-4. Next, the chrome browser and the corresponding chromiumdriver have to be installed; they are specifically needed to access the Java-based DuckDuckGo search engine and scrape data from some social media and website pages. Also, xdotool needs to be installed to enable the browser emulation script to run; TIP: make sure to switch the Wayland desktop to gdm3. The following python libraries need to be installed additionally: googlesearch-python, timeout-decorator, nltk (make sure to import the stopwords for each language afterwards), pdfminer.six, selenium, search-engine-parser and nest-asyncio.

The scripts have been run on an Ubuntu laptop with 16GB, an Ubuntu desktop with 32GB and a raspberryP4 with 8 GB of memory successfully. In principle, each script has an option (in the Ini-file) to run as much of the program in parallel to speed things up, but in practice it has become clear that this should NOT be done for Scripts 1 and 4. When run in parallel, Script 1 will access multiple search engines over a long period of time which has –unfortunately- always resulted in a crash of at least 1 of the search engines; when this happens during parallel scraping all links collected (including those collected by the other threads/cores) are lost. As a result, Script 1 is usually run sequentially and

because some search engines cannot be visited at a high frequency, it takes this script a considerable time to complete. Also a standard wait time of 20 seconds is used between queries, except for Bing (5 sec) and for the payed Google subscription (0 secs). This is done to prevent blocking. Script 2 and 3 can, and are routinely, run in parallel. This is possible, because Script 2 only collects up to the first 50 links provided by each search engine (for each query) and script 3 does not require search engines at all. Script 4a and 4b contain a browser emulation script that cannot be run in parallel; it otherwise messes up the assignment of the browser ids used and data will be lost. The latter combined with the huge number of links to be processed and in particularly the time it takes to determine if a web site really exists, is the reason script 4b takes long to be completed. The best way to speed up script 4b is to run it on multiple computers and provide each with a different part of the list of links produced by script 4a. Any duplicated results need to be removed afterwards. Such an approach could also be applied to speed up script 1, by dividing the queries over multiple machines, if one can assure that those machines have non-identical IP-address during the period of scraping.

3 First results for the four countries studied

3.1 Links found

In this section, an overview is provided of the number of links found when searching the web with drone specific queries for the four countries studied. In Table 3 the results of script 1-4 for each country and language (if applicable) are shown. The scripts are located on gitlab and can be found here: https://gitlab.com/pietdaas/drone. Let it be clear that the results in Table 3 are based on version 1 of the queries (see Annex 1). Results based on version 2 of the queries are in progress but have not been finalized. The number between brackets in Table 3 shows the combined results of both languages for a country.

Table 3. Number of links found by version 1of each script for each country for each language*

Scripts	Links found	Ireland	Netherlands English/Dutch	Spain English/Spanish	Germany English/German
Script 1	Web-links	2.608	2.586 / 3.464	829 / 981	786 / 885
	PDF-links	57	98 / 87	11 / 10	11 / 27
Script 2	Web-links (a)	11.108	4.602 / 10.496	4.767 / 8.034	9.154 / 92.993
	Web-links (b)	28	14 / 68	17 / 39	18 / 24
	PDF-links	714	307 / 665	465 / 671	400 / 6.286
Script 3	Web-links	1.384	1.213 / 675	371 / 449	3.364 / 4.851
Script 4a	Web-links	8.914	4.905 / 7.125 (10.793)	3.020 / 5.753 (8.295)	5.019 / 18.754 (22.587)
Script 4b	Web-links in country (combined)	2.171	1.224 / 2.753 (3.530)	634 / 2.489 (2.995)	754 / 4.741 (5.302)
	Web-links, unknown country (combined)	1.028	443 /366 (709)	306 / 459 (711)	647 / 1.055 (1.540)
	With drone words, in country (combined)	535	441 / 634 (913)	235 / 404 (569)	201 / 1.173 (1.322)
	With drone words, unknown country (combined)	120	71 / 29 (88)	54 / 31 (77)	96 / 48 (136)

* The findings for Ireland are based on an early version of the scripts and may - likely- differ somewhat when this work will be repeated with the most recent version of the scripts. Especially the country location part has improved considerably. All results shown for script 1 and 2 are based on version 1 of the queries (see Annex 1 for details).

When one looks at the output of each script, it is clear that the search strategy implemented in script 2 provides most of the links, followed by script 1. This is what is to be expected considering their focus (see section 2.1). For Germany, the numbers are especially high. When all goes well, script 2 and 3 can be run on a single day. Script 1 may take between 1 and 2 days to complete. Script 4a usually takes between 0.5 and 1 day to be completed and Script 4b takes between 1 and 2 days to be completed (see Table 2). To get an idea of the number of unique links found for the whole search procedure, one can best look at the results of Script 4a in Table 3. Here, all unique links, including those found after social media checking, are shown. These contain all the links to web pages that will be visited, scraped and their location checked. Between brackets the combined set of unique URLs for each language is shown. Comparing the results of the language specific findings in each country, clearly demonstrate the need to search the web for both languages for the countries studied; with the exception of Ireland of course.

The findings in Table 3 indicate that for the countries studied, 535, 913, 569 and 1.322 websites were found that contained at least one acronym for drone (penultimate row in Table 3). In addition, potentially, 120, 88, 77 and 136 websites could be additionally included (last row in Table 3). So likely between 535 and 655 web sites of drone companies are located in Ireland, between 913 and 1.001 in the Netherlands, between 569 and 646 in Spain and between 1.322 and 1.458 in Germany. More detailed studies need to be performed to verify these first findings; in particular, there is a need to compare the links found with those of the known drone websites in a country. The first results of such a comparison are described in the next section for Spain. All findings shown in Table 3 will be repeated for each language for each country with the updated version of the queries (see Annex A) and imporved scripts.

3.2 Drone websites findings for Spain

Script 1 version 1 results

For Spain script 1 has been used for both versions of the queries in both languages. In Annex 1 the differences between the different versions of the queries are shown for Ireland. From this, it is clear that version 2 contains much more combinations of search words and these queries are less complex; they lack the OR construction for drone acronyms. In Table 3 the number of unique links for version 1 of script 1 are shown for the combined results of the, at that point in time, five search engines used; AOL, Bing, DuckDuckGo, Google and Yahoo. A total of 829 and 981 unique links were found for Spain in the Spanish and English language, respectively. When combined, a total of 1.243 unique combinations of secondary and top domain names were obtained.

Script 1 version 2 results

For version 2 of the scripts much more detailed results are available for Spain. The combined results for the English and Spanish results are listed at the search engine level in Table 1. With this approach a total of 17.712 unique combinations of secondary and top domain names were obtained. This is a much higher number compared to the version 1 results and reveals that applying more and simpler queries are to be preferred. However, for both cases it is unknown how many of those links actually belong to a Spanish drone company. Creating such a list is the topic of the next section.

Report

Creating a list of drone company websites for Spain

To accurately determine how many and which Drone websites exists in Spain a list of Drone websites in that country need to be created. To construct such a list, a set of PDF-documents collected by the team that provided various overviews of Spanish Drone companies was used as a starting point. The documents used were:

- listado_proveedores_ats_ais_cns_certificadas.pdf,
- registro_autorizaciones_operac_aereas_uas.pdf,
- 20201222_Listado.Fabricantes.pdf,
- Listado entidades reconocidas.pdf
- listado_operadores.pdf
- Tabla Maestra de modelos de UAS.pdf

From all 6 PDF-files the tables were extracted. Unfortunately, only 1 document contained links to the website of drone companies; from this document a total of 3 URLs were obtained. Because of this, it was decided to extract the column containing the names of Drone companies from all files. This required a combination of automated and manual work. In the end, a total of 5.370 unique Drone company names were obtained. Next, the websites corresponding to those companies had to be found. For this task the URLsFinder approach developed in the ESSnet Big Data (2020) was used. In principle, this approach uses a search query, which in our case includes the name of company combined with 'Espana', and the links returned may indicate the potential website of that company. The payed Google search engine was used for this task and the first 10 links returned were collected for each search query. However, one very important new extension was included. Next to the first 10 links, the search page was also checked for the presence of a Google knowledge graph section; this is commonly shown in the right upper part of the web page and usually includes a map with the location of the company searched for and more info. When this section was present, it was checked for the occurrence of a button with a website link. If such a button was found, the corresponding link was extracted and added as the *first* suggested link to the top 10 results obtained. After searching all company names, a file with a 5.369 unique links was obtained. The combination of company names and links were manually checked by the team-members of the University of Valencia and the manager of the project. This resulted in a final list of 1097 unique combinations of secondary and top-domain names for Drone company websites in Spain. Remarkably, this final list contained 907 of the first links suggested by the extended URLsFinder approach[3]. A total of 190 manual corrections were required; of these 35 contained new (not yet known) domain names. This list was used to check the results provided by version 1 of Script 1 and the various search engines results of version 2 of Script 1, in both languages. The latter findings are shown in Table 4.

When the 1.243 unique combinations of secondary and top-domain names found by version1 of Script1 were compared with those included in the Spanish drone website list, only 32 links of the 1097 links were found; that 3%. For the 17.712 unique links found by the combined results of version 2 of Script1, a total of 272 links of the 1097 links were found that is nearly 25%. Clearly version 2 of the queries was able to find much more relevant links, but overall still only a quarter of the links were found by this approach. In Table 4, more detailed results are shown that give an idea of the importance of each search engine when searching for individual drone web site. By looking at the number of links to Drone websites detected by each search engine (Table 4) it becomes clear that Yahoo finds the most, followed by AOL,

[3] The importance of this finding will not be discussed here, but in the discussion section at the end of the report.

Table 4: Detection of Drone websites by comparing the secondary and top domain names of the links found by the six search engines for version 2 of the queries in Script 1 for Spain (both English and Spanish results are combined)

Search Engines	AOL	Ask	Bing	DuckDuckGo	Google	Yahoo	Total
Spanish drone websites found	187	122	120	94	150	202	272
Spanish drone websites found by all search engines	46	46	46	46	46	46	46
Spanish drone websites found by a single search engine	23	3	5	8	12	27	78
Spanish drone websites found by using one search engine after the other (order of importance)	31 (2)	3 (6)	5 (5)	9 (4)	22 (3)	202 (1)	-

Google, Ask, Bing and DuckDuckGo. From the table it also becomes clear that every search engine finds a common set of 46 links. In addition, row 3 in Table 4 reveals that each search engine finds drone website links not detected by others. Of these links, Yahoo, AOL and Google clearly contribute the most. This indicates that -at this point in time- there is no reason to reduce the number of search engines used. If a preference should be given to the order in which the search engines should be used than the relative contribution of the links provided by the search engines when run one after the other, makes clear that the order Yahoo, AOL, Google, DuckDuckGo, Bing and Ask would be most efficient. It will be interesting to compare the list of Spanish drone websites to those obtained by script 2 for that country.

First comparison of the overall drone classifier results (version 1)

Since there is a list of drone web sites for Spain we can now also check the final results obtained for Spain collected with version 1 of the Scripts (shown for Script4b in Table 3). When the domain names of all urls located in Spain (2.995) and those with an unknown location (711) are combined and compared, a total of 91 Spanish drone websites are found. This is 8.3%, a fairly low number. Of these links, 85 have been assigned to Spain. These finding confirm the need to adjust the initial queries developed and also to check if there is a way to improve the location assignment algorithm.

4 Conclusions

From the work described in this report it is clear that:

- Multiple search engines need to be used to get an as complete as possible overview of the web sites in a country
    - Apart from the big four (i.e. Yahoo, Google, Bing and DuckDuckGo) AOL and -to a lesser extent- Ask are advised to be used.
- When the search engine based approach described in this report is going to be used often or is going to be applied to a large number of (European) countries, it is suggested to gain access to as much paid search engines services as possible. This will not only speed up the runtime of the scripts (because it reduces the wait time between each request) but will also remove the change of being blocked. An alternative way to deal with the latter issues, when for instance a paid service is not available or when this does not provide the level of access needed, is using a VPN-service with a rotating IP-address.
    - Preferably a service should be used that allows access to multiple search engines and social media platforms via a single account. The latter will also speed up the steps included in Script 4a and 4b.
    - Downside of a rotating IP-address is that the results may be affected by its (faked) location. However, compared to having no access at all, this is less of a problem.
- Multiple languages need to be used for countries where English is not the dominant language
    - For countries where English is not the dominant language spoken, English queries should always be included in addition to queries in the dominant languages spoken. For Belgium, for example, it would be advised to search in Dutch, French and English.
- For script 1 it has been found that multiple, more simple, queries are to be preferred to complex and combined queries. The former provides much more links to the websites of drone companies.
    - Because of the –in general- fairly limited number of results obtained per search query (max between 400 and 500), it is wise to create as much specific search queries as possible. It is expected that this will be particularly useful when studying countries with large numbers of drone companies, such as France.
- Although the occurrence of the word drone or an acronym is definitely an indication for a website active in the drone industry, a more advanced classification method needs to be developed.
    - A language specific word-based classifier needs to be developed that enables a better identification of the web sites of drone companies in a particular country. Such an approach would also make it a lot easier to identify drone companies in country not yet included in the current study.
- Although the steps involved in determining the location, i.e. country, of a website work fairly well, there is room for improvement; this will reduce both false positives and false negative.
    - This could, for instance, be done by making use of specific structures indicative for address information; such as the zip-codes in Irish and the Netherlands.
- The first results obtained are predominantly based on the search approach implemented in script 1. These results demonstrate that up to now 25% of the links to Spanish drone websites were found in this first step.

Compared to the overall Spanish results for version 1 of the scripts (8.3%), this is already a huge improvement. However, it is still far from the 100% goal, certainly when one realises that the new step 1 search approach requires 2x48 queries (for each language) and takes more than 2 days to be completed.

- We also need to realize that the findings obtained with the URLsFinder *drone company name* based approach (described in section 3.2) produce -after a few hours- in a total of 907 (83%) of those links. To obtain this result, simply the first link for each company provided by Google is chosen prior to any manually checking performed. This observation also indicates that very specific, company names based, searches are to be preferred compared the more general drone word based approaches studied so far.

The results described in this report point to a number of important points:

1. It is challenging to find all drone websites in a country with a generic drone word search engine based approach.

2. To check the results obtained and the contribution of each step, it is the important to have a list of drone websites for (at least one of) the countries studied available. Advantage of this is that it enables an accurate checking of the results obtained and makes it possible to improve the approach developed by checking the effect of any adjustments made to the search strategies.

3. The construction of the list of drone websites for Spain reveals an alternative way to obtain links to drone websites. One could first focus on finding websites and documents containing NAMES of drone companies, and subsequently used those names to find the accompanying websites via a search engine based approach. For the latter only a single search engine can be used.

4. Another approach would be to obtain an overview of the domain names in a country. There are companies from which such lists can be obtained. The latter could be used to scrape those websites and check if they belong to a drone company.

4.1 Future work

Future work will focus on:

- Improving the search approaches for Spanish drone websites by checking the findings with the list of 1097 domain names created;
- Discuss the need to adjust the overall search strategy developed with all project members;
- Apply the new strategy to Spain first, followed by Ireland, the Netherlands and Germany;
- Improving the classification of the websites of drone companies for the countries studied. The 1097 websites of drone companies in Spain will be scraped and used in this task;
- Create a generic word-based model that can be converted and applied to languages spoken in other countries;

Report

- Check and improve the country location approach developed to reduce the number of unknowns and wrongly classified countries;
- Fine tune and debug the scripts developed (when needed);
- Apply the search-based approach to France.

References

Biggs, T. (2021). Is Bing best? Testing four search engine alternatives to Google. Located at: https://www.smh.com.au/technology/is-bing-best-testing-four-search-engine-alternatives-to-google-20210204-p56zi5.html

ESSnet Big Data (2020) Starter kit for NSIs V.1, Deliverable C4. Workpackage WPC Implementation –Enterprise Characteristics.

GOPA (2021). Deliverable D1, Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry.

Reliablesoft (2021). The top 10 search engines in the world (2021 update). Located at: https://www.reliablesoft.net/top-10-search-engines-in-the-world/

Wiki (2021). List of Internet top-level domains. Located at: https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains

Report

Annex 1: Queries used for Ireland

Version 1:

Script 1:

(drone OR rpas OR uav OR uas) training course ireland,
(drone OR rpas OR uav OR uas) pilot ireland,
(drone OR rpas OR uav OR uas) operator ireland,
(drone OR rpas OR uav OR uas) service provider ireland,
(drone OR rpas OR uav OR uas) manufacturer ireland,
(drone OR rpas OR uav OR uas) u-space ireland,
(drone OR rpas OR uav OR uas) components ireland,
(drone OR rpas OR uav OR uas) accessories ireland,
(drone OR rpas OR uav OR uas) software ireland

Script 2:

(drone OR rpas OR uav OR uas) members ireland
(drone OR rpas OR uav OR uas) register registration ireland
(drone OR rpas OR uav OR uas) association community ireland
(drone OR rpas OR uav OR uas) overview list ireland

Version 2:

Script 1:

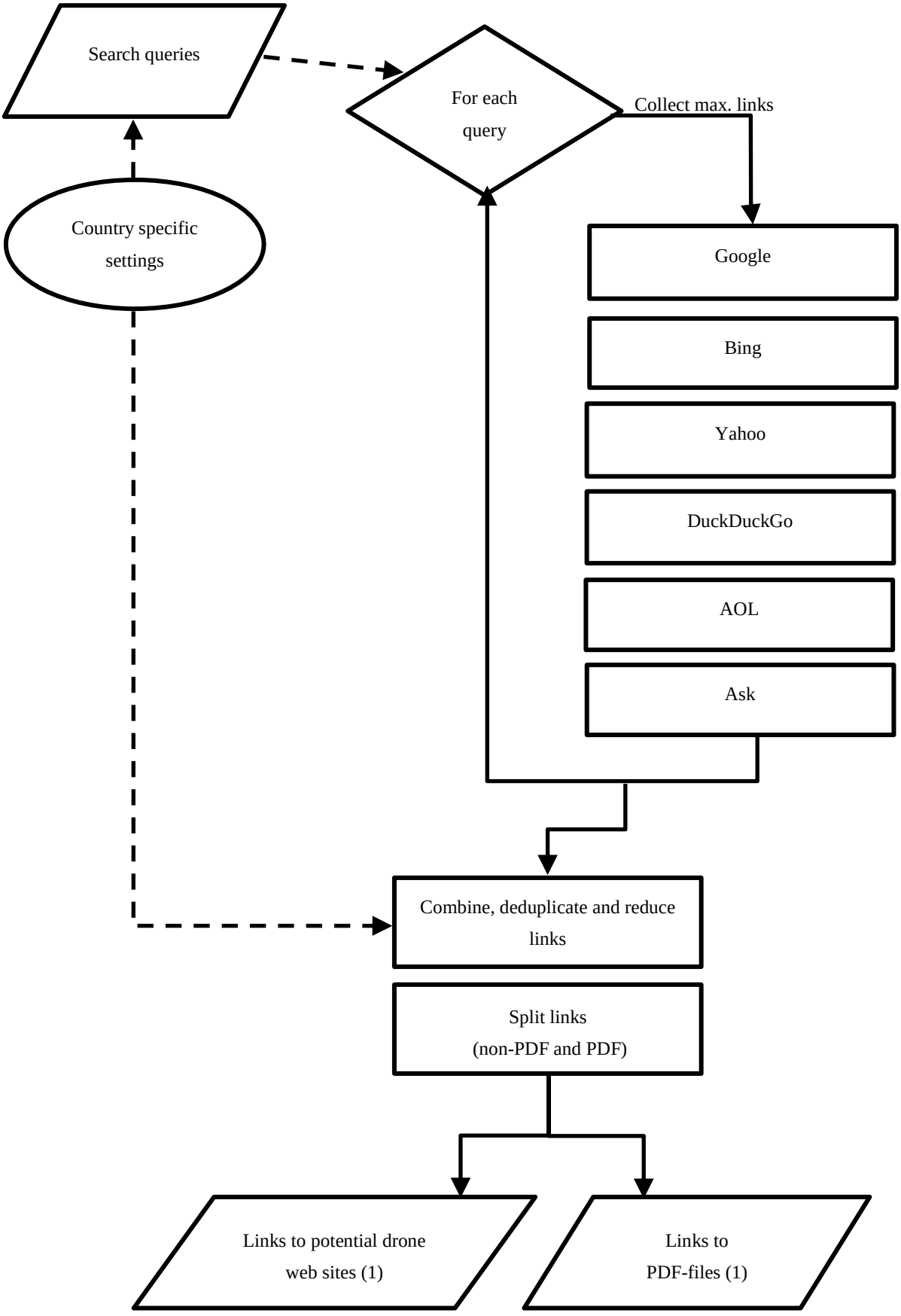
drone company ireland, rpas company ireland, uav company ireland, uas company ireland,
drone business ireland, rpas business ireland, uav business ireland, uas business ireland,
drone shop ireland, rpas shop ireland, uav shop ireland, uas shop ireland,
drone training course ireland, rpas training course ireland, uav training course ireland, uas training course ireland
drone pilot ireland, rpas pilot ireland, uav pilot ireland, uas pilot ireland
drone operator ireland, rpas operator ireland, uav operator ireland, uas operator ireland
drone service provider ireland, rpas service provider ireland, uav service provider ireland, uas service provider ireland
drone manufacturer ireland, rpas manufacturer ireland, uav manufacturer ireland, uas manufacturer ireland
drone u-space ireland, rpas u-space ireland, uav u-space ireland, uas u-space ireland
drone components ireland, rpas components ireland, uav components ireland, uas components ireland
drone accessories ireland, rpas accessories ireland, uav accessories ireland, uas accessories ireland
drone software ireland, rpas software ireland, uav software ireland, uas software ireland

Script 2:

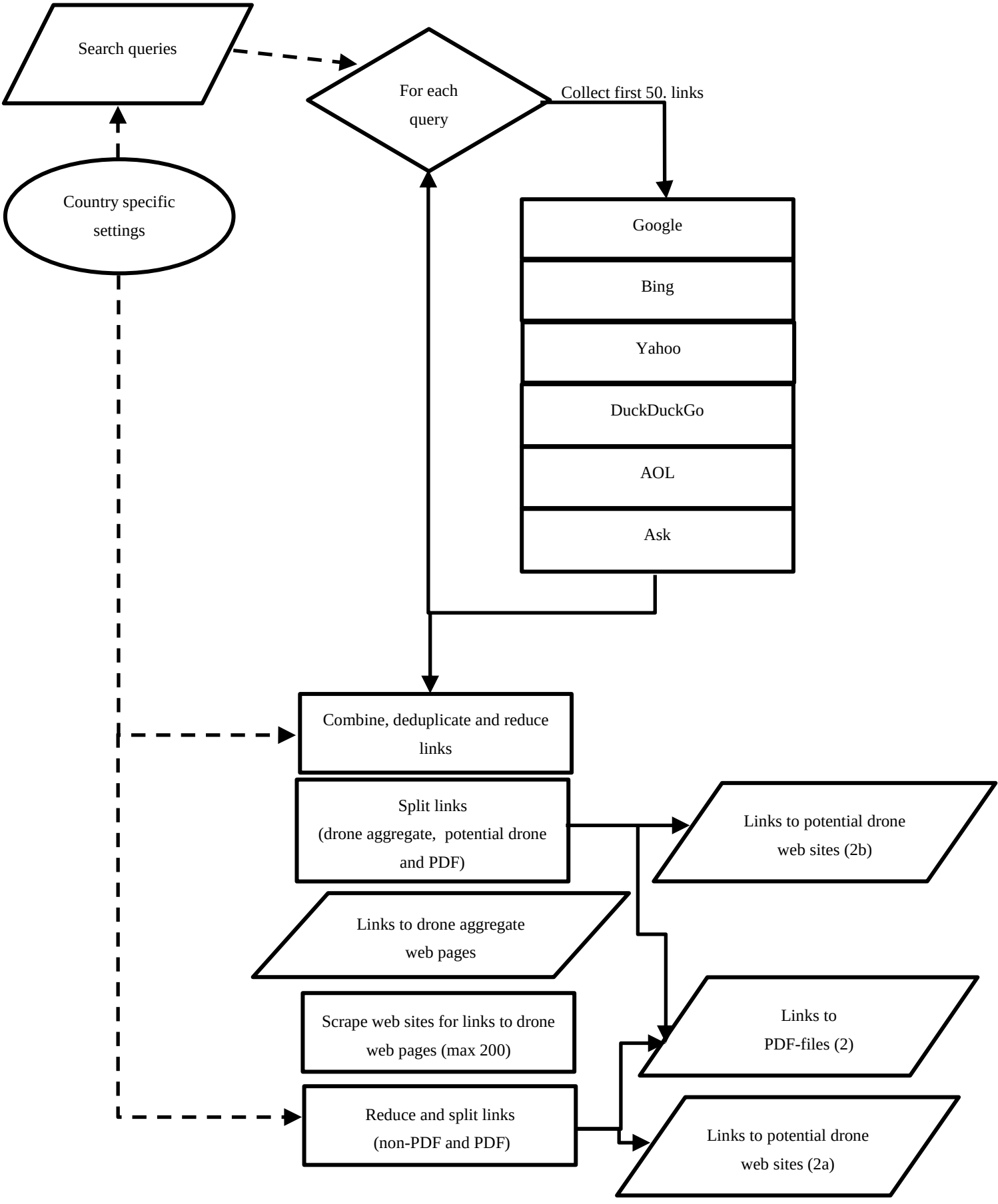
(drone OR rpas OR uav OR uas) members ireland
(drone OR rpas OR uav OR uas) register ireland
(drone OR rpas OR uav OR uas) registration ireland
(drone OR rpas OR uav OR uas) association ireland
(drone OR rpas OR uav OR uas) community ireland
(drone OR rpas OR uav OR uas) overview ireland
(drone OR rpas OR uav OR uas) list ireland

Annex 2: Flowcharts of the scripts developed

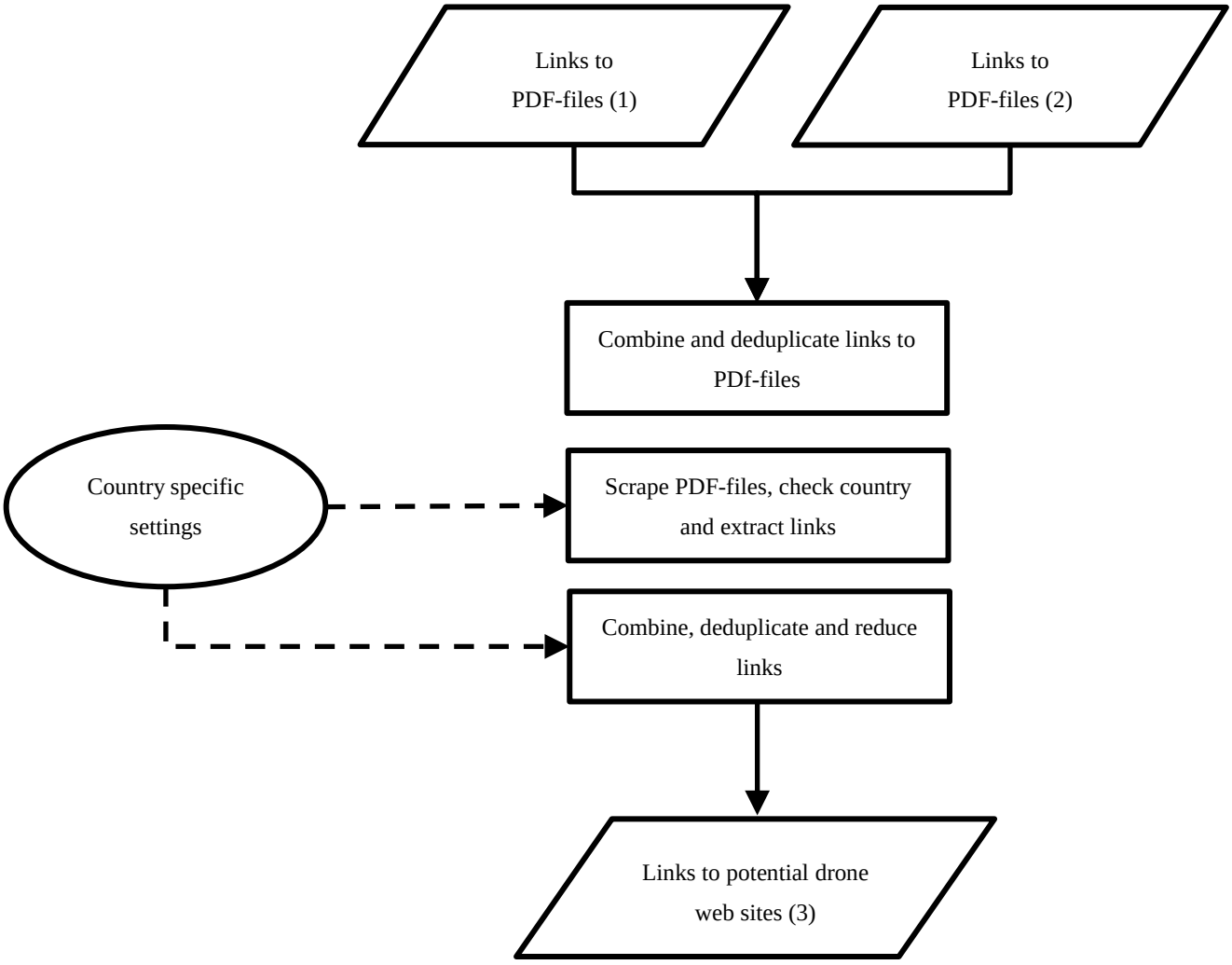
Script 1: Searching for websites of individual drone companies with multiple search engines



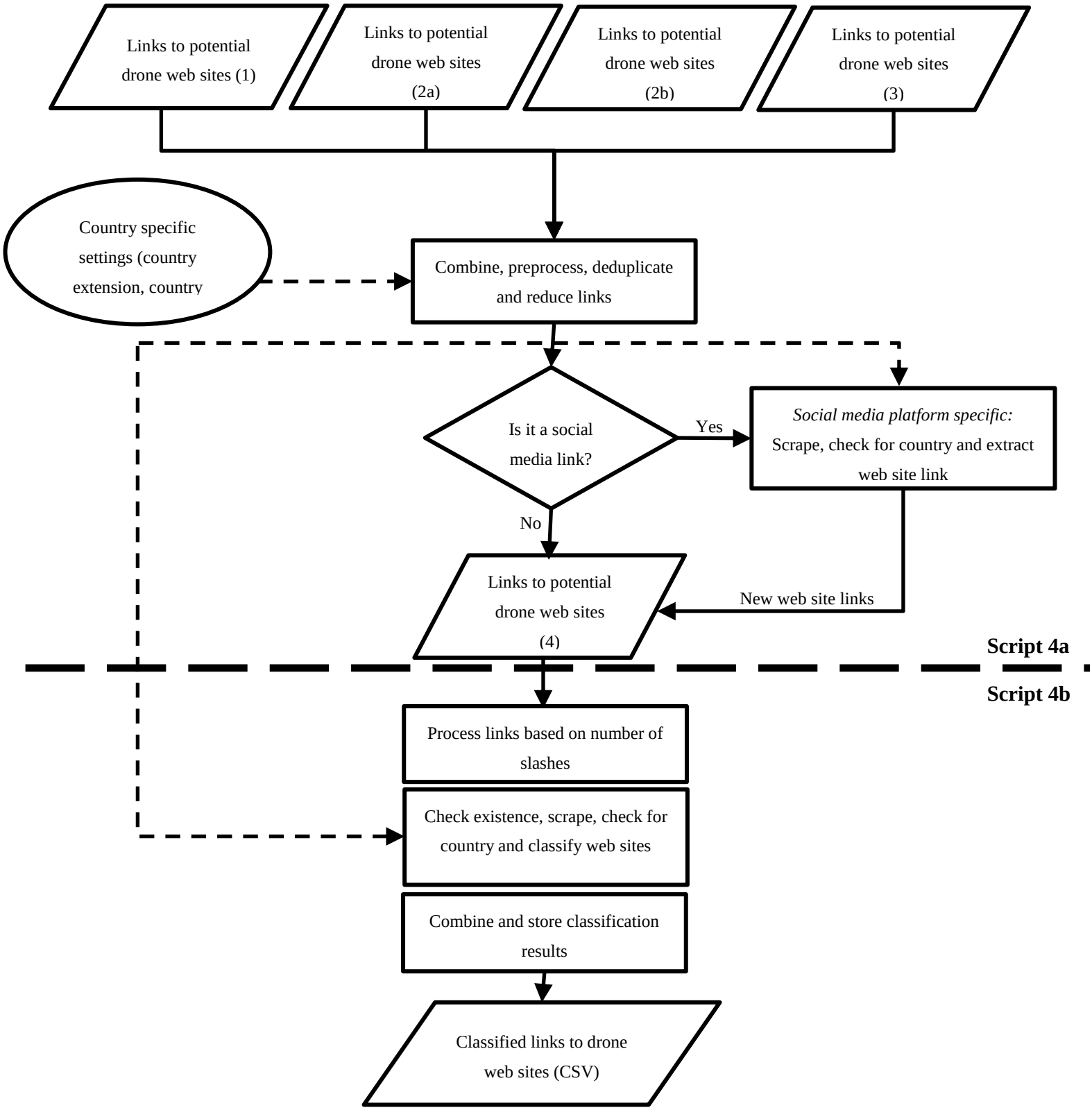
Script 2: Searching for web sites via overview sites of drone companies (with multiple search engines)



Script 3: Extract links to web sites from PDF-files



Script 4: Visit web site links and check if they are in the country and of a drone company



Report

Annex 3: Content of Ini-file for Ireland

```
## INIT SETTING ################################################################
##Settings for all scripts for searching and processing Ireland in English language

[MAIN]
country = ie
lang = en
googleKey = <#KEY_HIDDEN#>

[SETTINGS1]
queries = drone company ireland,rpas company ireland,uav company ireland,uas company ireland,drone business ireland,rpas business ireland,uav business ireland,uas business ireland,drone shop ireland,rpas shop ireland,uav shop ireland,uas shop ireland,drone training course ireland,rpas training course ireland,uav training course ireland,uas training course ireland,drone pilot ireland,rpas pilot ireland,uav pilot ireland,uas pilot ireland,drone operator ireland,rpas operator ireland,uav operator ireland,uas operator ireland,drone service provider ireland,rpas service provider ireland,uav service provider ireland,uas service provider ireland,drone manufacturer ireland,rpas manufacturer ireland,uav manufacturer ireland,uas manufacturer ireland,drone u-space ireland,rpas u-space ireland,uav u-space ireland,uas u-space ireland,drone components ireland,rpas components ireland,uav components ireland,uas components ireland,drone accessories ireland,rpas accessories ireland,uav accessories ireland,uas accessories ireland,drone software ireland,rpas software ireland,uav software ireland,uas software ireland
payedGoogle = True
runParallel = False
limit = 0

[SETTINGS2]
queries = (drone OR rpas OR uav OR uas) members ireland,(drone OR rpas OR uav OR uas) register ireland,(drone OR rpas OR uav OR uas) registration ireland,(drone OR rpas OR uav OR uas) association ireland,(drone OR rpas OR uav OR uas) community ireland,(drone OR rpas OR uav OR uas) overview ireland,(drone OR rpas OR uav OR uas) list ireland
countryW1 = .ie,ireland,irish
countryW2 = ireland, eire ,irish
drone_words = drone,rpas,uav,uas,unmanned,aerial
mem_words = member,register,registration,list,overview,association,community
payedGoogle = True
runParallel = True
limit = 50
maxDomain = 200
mailInclude = True

[SETTINGS3]
countryW2 = ireland, eire ,irish
drone_words = drone,rpas,uav,uas,unmanned,aerial
runParallel = True
mailInclude = True

[SETTINGS4]
language = english
countryW1 = .ie,ireland,irish
countryW2 = ireland, eire ,irish
drone_words = drone,rpas,uav,uas,unmanned,aerial
mem_words = member,register,registration,list,overview,association,community
cont_words = contact,about,who
droneW = drone,rpas,uav,uas,unmanned
aerialW = fly,aerial,air
waterW = underwater,uuv
busW = business,company,operator,enterprise
contactW = contact,about
runParallel = False
cityNameFile = IEcity_names.csv
countryNames = Ireland,Eire
```

Annex 4: Part of the output of Script 4b

org_url	url_visited	In country	Score*	Top10_words	Action
http://3go.es	http://3go.es	TRUE	0 0 0 0 3	(18, 'de') (10, 'la') (10, '3go') (4, 'webcam') (4, 'para') (4, 'con') (4, 'componentes') (4, 'caja') (3, '\xa0(lunes') (3, 'viernes')	located in country studied
http://49thsquadron.com	https://49thsquadron.com	NA	1 65 0 0 1	(17, 'rc') (15, 'fly-in') (14, 'military') (14, 'jumbo') (13, 'squadron') (13, 'club') (12, 'flying') (11, 'july') (11, 'jamboree') (11, '49th')	Location unknown
http://aboutcookies.org	https://aboutcookies.org	NA	0 0 0 0 4	(7, 'cookies') (5, 'site') (5, 'cookie') (4, 'web') (3, 'website') (3, 'use') (3, 'pinsent') (3, 'help') (3, 'cookies,') (2, 'uses')	Location unknown
http://aeropuertodemalaga-costadelsol.com	https://aeropuertodemalaga-costadelsol.com	TRUE	0 0 0 0 2	(52, 'de') (33, 'aeropuerto') (20, 'del') (19, 'mÃ¡laga') (15, 'el') (13, 'en') (8, 'la') (7, 'sol') (6, 'mÃ¡laga-costa') (5, 'coche')	located in country studied
http://africanaerospace.aero/going-into-battle-against-the-deadly-hoardsafrican	https://africanaerospace.aero/going-into-battle-against-the-deadly-hoardsafrican	NA	0 0 0 0 0	(3, 'found') (1, 'web') (1, 'server.') (1, 'server') (1, 'requested') (1, 'document') (1, 'africanaerospace.aero') (1, '404')	Location unknown
http://aguadrone.com	http://aguadrone.com	TRUE	18 1 0 0 3	(13, 'pod') (9, 'sonar') (9, 'fish') (8, 'release') (7, 'drone') (7, 'aguadrone') (6, 'water') (6, 'payload') (6, 'data') (5, 'wireless')	located in country studied
http://aiprox.com	http://aiprox.com	TRUE	23 8 0 0 5	(22, 'aiprox') (18, 'system') (17, 'drone') (10, 'charging') (9, 'extreme') (8, 'landing') (8, 'analysis') (7, 'information') (6, 'systems') (6, 'sensors,')	located in country studied
http://air-pros.com/uas.php	https://air-pros.com	TRUE	8 69 0 4 8	(46, 'insurance') (34, 'aircraft') (23, 'aviation') (12, 'flight') (12, 'air') (11, 'resources') (10, 'pilots') (10, 'get') (9, 'quote') (8, 'view')	located in country studied
http://airspace.co	http://airspace.co	NA	10 16 0 3 4	(11, 'airspace') (5, 'drones') (4, 'solutions') (4, 'galaxy') (3, 'use') (3, 'new') (3, 'drone') (2, 'within') (2, 'take') (2, 'systems,')	Location unknown
http://al-top.com	https://al-top.com	TRUE	5 2 0 6 5	(45, 'de') (23, 'trimble') (16, 'para') (16, 'cookies') (14, 'topografÃa') (14, 'sistemas') (14, 'dji') (13, 'software') (11, 'servicio') (11, 'lÃ¡ser')	located in country studied
http://anacgabon.org	http://anacgabon.org	NA	0		no text could be extracted
http://anacm-comores.com	https://anacm-comores.com	NA	0 4 0 0 1	(28, 'de') (15, 'et') (12, 'la') (7, 'les') (6, 'comores') (5, 'des') (5, 'civile') (4, 'tÃ©lÃ©phone') (4, 'mÃ©tÃ©orologie') (4, 'mohamed')	Location unknown
http://asociaciondepilotos.com	http://asociaciondepilotos.com/es	NA	0 0 0 0 1	(15, 'de') (8, 'la') (4, 'en') (4, 'aviadoras') (4, 'aep') (3, 'programa') (3, 'el') (3, 'curso') (2, 'una') (2, 'seminarios')	Location unknown
http://auvsi.org	https://auvsi.org	NA	0		url or domain already checked or excluded
http://aviation.und.edu	https://aero.und.edu/aviation	TRUE	9 16 0 1 5	(36, 'aviation') (15, 'aerospace') (13, 'und') (12, 'university') (9, 'north') (7, 'search') (7, 'sciences') (7, 'dakota') (6, 'systems') (6, 'students')	located in country studied

*Here counts of the occurrences of drone, aerial, and water synonyms and words indicative for businesses and contact pages are shown when texts extracted from a webpage an it is either located in the country studied or has an unknown location.