Web intelligence for measuring emerging economic trends: the drone industry

PIET DAAS, MARTIJN TENNEKES,
BLANCA DE MÍGUEL, MARÍA DE MIGUEL,
VIRGINIA SANTAMARINA, FLORABELA CARAUSU

2022 edition

STATISTICAL
WORKING PAPERS

eurostat

Web intelligence for measuring emerging economic trends: the drone industry

2022 edition

Manuscript completed in June 2022

Luxembourg: Publications Office of the European Union, 2022

Theme: General and regional statistics
Collection: Statistical working paper

ISBN 978-92-76-53307-8 ISSN 2315-0807 doi: 10.2785/620917 KS-TC-22-004-EN-N

Abstract

In order to reap the benefits of the data revolution and to complement traditional statistical methods and sources with innovative approaches, Eurostat recently established the Trusted Smart Statistics initiative. This also encompasses the so-called Web Intelligence Hub. In the context of the Web Intelligence Hub, Eurostat initiated experimental research towards extending the capabilities of retrieving information on European drone businesses from the internet.

The 'Web Intelligence for Drones' initiative builds on previous research into web scraping of business information in the context of official statistics. The study was triggered by the absence of consolidated data on drones, the operation of drones or market size in the EU, despite the EU having one of the most advanced regulations on unmanned aircrafts in the world. In particular, research focusses on the development of a methodology and of the tools needed to retrieve information from the World Wide Web (www) concerning businesses based in EU countries that have their main activity in the civil drones sector.

This scientific summary presents the chain-like methodology and the tools developed to identify drone companies through the www and to extract company-relevant information from their websites. The method was developed with a perspective of generalisation in mind (across countries and across economic sectors) wherever possible. It has already been fully applied to three European countries (Spain, Italy and Ireland).

Keywords: Drones, Web Intelligence, BigData, new data sources, business statistics

Authors: Piet Daas([1]), Martijn Tennekes([2]), Blanca de Míguel([3]), María de Miguel([4]),Virginia Santamarina([5]), Florabela Carausu([6])

Acknowledgments: We would particularly like to thank the coordinators of this study at Eurostat – Unit A5: Andrea Ascheri, Fernando Reis, Albrecht Wirthmann and Matyas Meszaros for their helpful comments.

Acknowledgments are extended as well to Paola Olivares (Osservatori Digital Innovation School of Management – Politecnico di Milano) for her support with the research and for her sharing sector knowledge about the Drone market in Italy.

The study was implemented by GOPA as a contractor of Eurostat for the framework contract on Methodological Support (Ref. 2018.0086) in collaboration with Statistics Netherlands together with individual researchers of the Universitat Politècnica de Valencia.

([1]) pjh.daas@cbs.nl
([2]) m.tennekes@cbs.nl
([3]) bdemigu@upvnet.upv.es
([4]) mademi@omp.upv.es
([5]) virsanca@upv.es
([6]) florabela.carausu@gopa.de

Table of contents

Tables

Figures

Abbreviations

AESA	Agencia Estatal de Seguridad Aérea (Spain)
API	Application Programming Interface
CBS	Centraal Bureau voor de Statistiek (the Netherlands)
D&B	Dun & Bradstreet
DTM	Document Team Matrix
EASA	European Union Aviation Safety Agency
EC	European Commission
ENAV	Ente Nazionale Assistenza al Volo (Italy)
ESS	European Statistical System
ESSnet	European Statistical System research network
EU	European Union
FAA	Federal Aviation Administration
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GISCO	Geographic Information System for the Commission
GOPA	Gesellschaft für Organisation, Planung und Ausbildung
IAA	Irish Aviation Authority
IP	Internet Protocol
NACE	Statistical classification of economic activities in the European Community
NLTK	Natural Language Toolkit
NUTS	Nomenclature of Territorial Units for Statistics
RPAS	Remotely Piloted Aircraft System
SESAR	Single European Sky ATM Research
SMEs	Small and Medium Enterprises
UAAI	Unmanned Aircraft Association of Ireland
UAS	Unmanned Aircraft System
UAV	Unmanned Aircraft Vehicle
URL	Uniform Resource Locator
US	United States
VPN	Virtual Private Network
WIH	Web Intelligence Hub
www	World Wide Web

Introduction

Introduction

Eurostat is the statistical office of the European Union. Its mission is to provide high quality statistics and data on Europe. In order to reap the benefits of the data revolution and to complement traditional statistical methods and sources with innovative approaches, Eurostat has recently established the Trusted Smart Statistics initiative that will enable the development of IT, methodological and quality frameworks, guidelines, tools and infrastructure that are suitable for big data processing. As part of this Centre, Eurostat in collaboration with national statistical institutes (i.e. the European Statistical System) will produce official statistics based on information that can be retrieved from the world wide web (www). In the framework of the so-called Web Intelligence Hub, Eurostat has initiated experimental research towards extending the capabilities of retrieving information on European drone businesses from the internet.

The 'Web Intelligence for Drones' initiative builds on previous research on the web scraping of business information in the context of official statistics. The study was triggered by the absence of consolidated data on drones, the operation of drones or market size in the EU, despite the EU having one of the most advanced regulations on unmanned aircrafts in the world. In particular, research targets the development of a methodology and of the tools needed to retrieve information from the www for businesses based in EU countries that have their main activity in the civil drones sector. The methodology and tools are developed with a perspective of generalisation, i.e. it targets their application to other emerging economic trends.

In a consolidated manner, this report presents the work performed to acquire web intelligence concerning the drones sector, describing the methodologies developed and the results obtained. Furthermore, it informs on the application of the methodologies developed to other emerging economic trends.

This report is structured as follows:

- Chapter 1 introduces previous research initiatives on the web scraping of business information undertaken by the European Statistical System (ESS) partners, and it sets the legal background of the drone use case, with a particular focus on the European Drones Regulation.
- Chapter 2 presents the search approaches explored for identifying drone company websites through the www.
- Chapter 3 presents the method and scripts developed for the retrieval of URLs.
- Chapter 4 presents in detail the classification model and its extension to other countries studied.
- Chapter 5 presents the data extraction method. It informs about the analyses related to the activities of drone companies in the value chain and the analyses related to their location.
- Chapter 6 discusses the generalisation of the method and tools, and it presents an automated keyword extraction method.
- Chapter 7 presents the main findings and recommendations.

1 Background

1. Background

1.1. Previous ESS research initiatives on web scraping of business information

The project builds upon previous research on the web scraping of business information in the context of official statistics. It aims to further extend the capabilities of retrieving information on businesses from the internet, in particular for the use-case of drone businesses, albeit with a view to generalising the approach to new emerging economic trends. In particular, the research the project builds upon is the work accomplished within the framework of the ESSnet Big Data I Work Package 2 (WP2) and the ESSnet Big Data II Work Package C (WPC), which focused on enterprise characteristics.

ESSnet Big Data I WP2 was set up as a pilot study to investigate whether web scraping, text mining and inference techniques can be used to collect, process, and improve general information about enterprises. Specifically, it aimed to ascertain, whether business registers can be improved by using web scraping techniques and by applying model-based approaches, so as to predict the values of certain key variables for each enterprise, together with verifying the possibility of producing statistical outputs using predicted data, in combination, or not, with other sources of data (survey or administrative data).

Based on information gained from the enterprises' websites, the specific use cases explored included estimating, whether an enterprise performs e-commerce or not, whether an enterprise manages job vacancies on its site, whether an enterprise is present in social media, its contact information as well as other profiling information (e.g. type of business, links to other enterprises, etc.). The use cases were implemented following a four-step process flow: internet access (i.e. URL retrieval and website scraping), storage (organisation of unstructured scraped data), data preparation (i.e. processing of texts), and analysis (i.e. machine learning or deterministic methods).

Two main categories of outcomes on enterprise characteristics produced by the ESSnet Big Data I WP served as a starting point for continuing the initiative within the framework of the ESSnet Big Data II WPC. These are: the methodology of URL (Uniform Resource Locator) retrieval, process and software implementation, together with the methodologies, processes and software implementations for detecting enterprises' characteristics. These outcomes proved to be suitable for implementation in all ESS countries, with some adaptations being made to local circumstances.

The ESSnet Big Data II WPC focused on the generalisation and extension of methodology and tools, in line with ESS countries' particularities. Implementing WPC involved the massive web scraping of

company websites, the collecting, processing and analysing of unstructured data, and the dissemination of national-level experimental statistics. Alongside publishing experimental statistics (data and metadata) for the participating countries, the project also developed several policy orientation and methodological guidelines, including an ESS-web scraping policy template (draft for consultation with ESS partners), the Reference Methodological Framework for producing generalised and extended methods, procedures and implementation requirements for web scraping on enterprise characteristics, the Starter Kit for National Statistical Institutes, which consists of procedures for testing and maintenance of web scraping, and the Quality Template for statistical outputs that are based on web-scraped enterprise characteristics.

As a standalone initiative at country level, but also building on the ESSnet Big Data projects, Statistics Netherlands (CBS) has developed a big data method for the web scraping of company websites with the aim of identifying small innovative businesses that are not covered by the innovation survey the institute carries out, whose population frame consists of companies with more than 10 employees. Based on the content published on the websites' main pages, the software tools developed distinguish between innovative and non-innovative businesses. Classifying companies on the basis of the text they publish on their websites proved to work well for innovation, as the survey-based results could be replicated by the text-based approach developed. A copy of (nearly) all business websites in the Netherlands was necessary to attain that goal. With small companies however, dealing with the dynamics of the link between businesses, websites and the classification of the websites of (semi-)self-employed is challenging. A model was recently developed to identify platform economy websites that are active in the Netherlands.

Developing a method to automatically link URLs to the business register, more specifically the domains of legal units, was the fruit of another standalone initiative by CBS in collaboration with the private entity Data Provider. The method covers enterprises of all sizes although the process is more challenging for recently established enterprises.

In all cases, the research projects listed above yielded successful results, and these form the basis of the 'Web Intelligence for Drones' project. The recent study did however highlight specific additional challenges that had not (yet) been addressed in the previous projects. The main statistical challenges are: (given the absence of a classification of drone businesses as per the current NACE Rev 2) the universe the study builds on the basis of the information existing drone operators' websites provide, the lack of a website for some drone businesses (in particular if these are start-ups or individual businesses), the lack of data (both at national and at EU level) that could serve as a benchmarking reference to the estimates being calculated on the basis of data retrieved from the web. The sector's dynamic and innovative features, its companies' rather small size and the existence of borderline activities (e.g. drone operator vs drone business) presented additional challenges that were explored within the scope of the project.

The *Web Intelligence for Drones* project

- Measuring emerging economic trends through web sources, together with the particular use case of the drone industry, provide an input for future developments within the European Statistical System (ESS) Trusted Smart Statistics framework and the Web Intelligence Hub (WIH).

In the context of the project, and more specifically in this report, the term Drone(s) is broadly used to refer to any Unmanned Aircraft Vehicle (UAV) or Unmanned Aircraft System, indistinctively.

A UAV, commonly known as 'drone', is an aircraft without an on-board human pilot. UAVs are a component of a UAS that includes a UAV, a ground-based controller and a communications system between the two. There is a wide range of UAVs ranging from light and simple to heavy and complex aircraft, which operate with various degrees of autonomy and a diverse set of missions.

1.2. Legal background: the Drones Regulation in Europe

Since the beginning of 2019, the European Union Aviation Safety Agency (EASA) has been working with the European Commission (EC) to provide common European rules that could improve the free circulation of drones within the European Union (EU). A key point was to develop a common and interoperable registration system that has been designed by the international public-private partnership SESAR (Single European Sky ATM Research) Joint Undertaking. Pursuant to Regulation EU 2019/947, this common framework registration became mandatory in January 2021, targeting the conversion of all national authorisations, certificates and declarations to the new EU system by July 2021. Full operability of the common registration system may be expected for 2022 ([7]).

The certification of a UAS (Unmanned Aircraft System) implies collecting information on drone manufacturers, given that, for each drone that is subject to certification, a unique digital registration number will be issued. The information collected includes the manufacturer's name, and it allows for individual identification. On the other hand, UAS operators whose operations may present a risk to safety, security, privacy, and protection of personal data, or to the environment, should also form part of the EASA register. One must note that most operations do present the risks mentioned above, as most UASs include a camera. Additional information is required for the mandatory registration of operators whose activities present a risk, and this goes beyond basic information. The additional information requested may include the activities the operators usually perform. Based on the information available at national level, the main activities of some of the national authorities are inspections/monitoring, flight tests, photography, pilots' training, and others.

An important question is whether the lists of registered manufacturers and operators will be publicly available in the future. The EASA registration system differentiates between natural and legal persons; therefore, the first list could be brought under the protection of the General Data Protection Regulation (GDPR). It is assumed that the second list (i.e. registered legal persons) could be made available to other EC departments, if not made publicly available. Once the common European registration system becomes operational, these lists' future availability is important to the exploratory sector analysis and to the validation of results collected from web data sources for the specific use case of drone businesses.

1.3. Analysing the industry through company websites

An industry is composed by a group of companies that buy from suppliers of similar goods and services to offer similar products and services to their buyers (Rothaermel, 2019). Official registers organise companies according to national classifications of activities. In the EU, the NACE classification allows to identify a company with the sector in which it performs its main activity. This brings data to measure the importance of a sector and its impact in Europe. Companies usually disclose this information on their websites, where they may also reveal their main activities as well as the sectors they target.

Analysing an industry from the economic perspective indicates the number of businesses in that industry, where they are located, the impact the industry has on its location and neighbouring locations (in terms of GDP, of employment), and the impact an industry has on other sectors (Boix et al., 2015; de Miguel Molina et al., 2012).

Over the last years, academic researchers have developed methodologies to extract information from company websites that might well become a substitute for surveys, especially for the analysis of emerging sectors such as nanotechnology, graphene and green goods (Youtie et al., 2012, Shapira et al., 2014; Kinne and Axenbech, 2020). Statistical experts have also developed methods that might

([7]) See for further details: https://www.easa.europa.eu/newsroom-and-events/news/drone-guidance-extended-and-updated-support-safe-drone-operations-easa

substitute national innovation surveys as they are carried out nowadays (Daas and Van Der Doef, 2020).

Previous studies cite the advantages in using websites as a source of information, such as public availability, lower costs, complementary data, increased timeliness, and suitability to work with time-series data through the Wayback Machine (Youtie et al., 2012; Shapira et al., 2016). They do however indicate some challenges with using websites as a source of data, such as the difficulties in accessing some of them due to their design (Arora et al., 2016; Shapira et al., 2016) and the fact that the information disclosed differs, depending on the company, sector and country (Arora et al., 2013). The studies analysed also indicate that keywords might capture various meanings and this would imply searching for more information about a company (Shapira et al., 2014). Other authors find that less information is available for big companies when they use the Wayback Machine (Gök et al., 2015), which also entails a challenge when subdomains are growing due to new investors or new languages, as the company enters new markets (Arora et al., 2016).

The studies analysed mainly focused on SMEs. Three main tasks can be detected in the methods they performed: a) searching for a list of companies in a different way than through using activity codes, b) extracting the necessary keywords from websites, and c) analysing the information to classify websites based on the keywords they include. The main characteristics of previous studies exploring company websites as a source of information for the analyses of industries are listed in the following table.

Table 1: Main characteristics of previous studies exploring company websites as a source of information for industry analyses

Reference	Sector analysed	List of companies	Sample (companies)	Use information from Wayback Machine	How information is extracted	How information is analysed
Ellinger et al. (2003)	Motor carrier	Top 100 motor carriers	98	No	Content analysis	Descriptive
Libaers, Hicks and A. Porter (2010)	Various	≤ 500 employees & ≥ 15 patents	407 US SMEs	No	Search for 89 keywords on firms' websites	Factor analysis with keywords
Youtie et al. (2012)	Nanotechnologies	List available	30 US SMEs	Yes	Web scraping, content analysis	Cluster analysis
Arora et al. (2013)	Graphene	Search firms online, in papers, patents and business database	20 SMEs in 3 countries	Yes	Web scraping, searching for keywords on webpages	Cluster analysis
Shapira et al. (2014)	Green goods (green technology manufacturing companies)	Search companies (and their employees) in Fame database, with a list of keywords (from different sources)	304 UK companies	No	Search for list of keywords on companies' websites	Descriptive
Gök et al. (2015)	Green goods	Database by Shapira & Harding (2012) and Fame database	296 SMEs UK-based companies	Yes	Web crawling and text mining (237 out of 296 had a website)	Descriptive and correlations
Arora et al. (2016)	Green goods	Own database from a project	300 SMEs US-based companies	Yes	Web crawling	---
Shapira et al. (2016)	Graphene	Search companies in patents, social media, papers, business databases, reports, websites	65 SMEs worldwide (manufacturers)	No	Web crawling and web mining (content analysis)	Cluster analysis and binary logistic regression
Li et al. (2018)	Green goods manufacturers	Search in D&B database with 100 keywords	300 SMEs US	Yes	Web crawling and content analysis	Hausman–Taylor estimation
Arora et al. (2020)	Green goods	Search in D&B database with 100 keywords	223 SMEs US	Yes	Web crawling and content analysis	2SLS regression
Bruni & Bianchi (2020)	Various	List from the Community Survey on ICT usage and e-commerce	4755 Italian companies	No	Web scraping and text mining	Categorization/ classification algorithms
Daas & Van Der Doef (2020)	Various	Companies in CIS Netherlands	824,972 Dutch companies	No	Web scraping	Classification algorithms, logistic regression
Kinne & Axenbech (2020)	Various	Mannheim Enterprise Panel	2.4 million German companies	No	Web scraping (ARGUS, proprietary software) and data mining (1.15 out of 2.4 million firms have URL)	Descriptive and probit regression
Mirtsch et al. (2021)	Various	Mannheim Enterprise Panel	912,850 German companies	No	Web scraping (with ARGUS), web content mining and web structure mining	Descriptive and probit regression

Source: own elaboration based on research published in the Web of Science database

When one considers the papers analysed in Table 1, the main challenge to the exploratory sector analysis is that of obtaining a list of company websites. Most of the industry analyses based on company website information (see Table 1) made use of an already existing database. The advantage of those datasets is that they offer additional information for further analyses, such as the number of employees in each company, for example. When such a list was not available, the authors created a list of keywords and searched for them in a commercial database (Fame) ([8]) or in the Dun and Bradstreet (D&B) database. Alternative solutions available were: i) buying access to a list of companies and their websites or ii) directly searching on the www. In all studies that directly searched on the www, the search was performed manually and it only made use of a single search engine.

Developing a method to create lists of drone businesses from the information that is available on the www is explored through the search strategies described in Chapters 2 and 3.

1.4. Exploratory analyses of the drone sector

Exploratory drone sector analyses were performed at country level, in a number of Member States (Spain and Ireland). This provides a more thorough conceptual understanding of the sector and its characteristics. It also acts as an input into the development of methodologies and tools for the retrieval of information on the sector from the web. A web search strategy was developed based on the exploratory sector analyses, with the aim of identifying the universe/population of drone businesses, based on web content, and to explore the feasibility of collecting information on these. The main findings of the exploratory drone sector analyses in the two countries are provided in the following paragraphs.

Although the universe of the study is built on the existing websites of drone businesses, for the purpose of analysing the industry, a starting point taken was to identify publicly available lists of drone operators and businesses. This facilitates the identification of drone-business websites and, in particular, it furthers the exploration of the information one can extract from the websites (i.e. classification and variables).

In both countries for which drone sector exploratory analyses were carried out, the study team made use of publicly available information concerning registered drone operators or companies. The sources of information were: for Spain, AESA (Agencia Estatal de Seguridad Aérea) and, for Ireland, IAA (Irish Aviation Authority), UAAI (Unmanned Aircraft Association of Ireland) and the Drones Ireland Directory website. Several lists of drone operators and companies were compiled and analysed from these web sources.

For Spain, the AESA website yielded the following registers:

- A list of approximately 5 200 drone operators (as of December 2020), including individuals and companies, and details on the activities these were authorised to undertake;
- A list of approximately 67 companies (as of December 2020), including information on specific operational scenarios ([9]). Those 67 companies were also all listed in the list of drone operators (see first bullet point);
- A list identifying 11 Spanish drone manufacturers.

For Ireland, the three web sources analysed yielded the following registers:

- A list of 350 drone operators authorised by the IAA, including the names of professionals, companies and offices of local governments registered through the IAA;
- A list of 5 flight schools and 65 UAAI members. 42 % of the records in this list overlapped

([8]) https://fame.bvdinfo.com/
([9]) Three application scenarios are defined in the new EU Drone Regulation: Open, Specific and Certified. They depend exclusively on the flight's risk level, independently of the nature of the use of civil drones (leisure or professional).

with the IAA list (see first bullet point);

- A directory of drone professionals and companies operating in Ireland, which allowed for the identification of those entities having their headquarters in Ireland. Out of the 56 results obtained for this directory, applying the headquarter filter, 29 % of records overlapped with those in the IAA list (see first bullet point).

In Spain, one could distinguish a certain level of centralisation and consistency, in the information on drone operators and companies. All three lists were published by the national air safety authority (AESA in Spain) and the relevant records in the shorter lists were, in all cases, already recorded in the comprehensive list of all drone operators authorised in Spain. The shorter lists published by the same source (i.e. AESA) were helpful for the further characterisation of specific companies. In Ireland, the additional websites with information on drone companies and operators complemented the list published by the IAA, not only with additional information, but by also extending the list of known companies or operators.

Based on these web sources listing operators and companies that are active in the drone sector in the two countries, a sample of companies was selected. For Spain, the sample was given by the companies listed in the short list (i.e. the 67 companies authorised for specific operational scenarios and the 11 Spanish drone manufacturers). The final sample thus included 76 companies (two companies were referred to in both lists). For Ireland, the sample was selected based on the IAA list. Namely, those operators that are located in Dublin County were selected (a total of 110 operators, roughly 1/3 of the total list for the whole country).

For both countries, the URLs of selected companies were picked manually. In Spain, a website was available for each company in the sample while, in Ireland, no website could be found for 13 of the 110 companies.

In order to analyse the information extracted from the identified websites, two steps were followed: i) web observation and definition of variables, and ii) content analysis with the support of QDA Miner 5. Step 1 also led to the identification of individual companies, specifically in the case of Ireland. The detection of those, along with other exclusion criteria (i.e. companies not headquartered in Ireland) further reduced the sample of companies analysed. The final sample size in Ireland amounted to 82 companies. Furthermore, web observation was used to obtain data on a set of predefined variables (e.g. availability of information about drones on the specific website, the economic sector the company is active in, its main activity, value chain identifiers). Step 2 saw the identification of key words that are used by companies to refer to ‘drones’, the type of services or operations the companies offer, with or in relation to drones, and the economic sectors targeted by the company’s offer).

The drone industry exploratory analysis showed that many company websites do not include any information concerning drones. In the case of Ireland, only 44 % of the companies do disclose information about drones on their webpages.

The industry analysis showed that, in both countries, the sector presents a different structure, influenced, one could say, by creative industries in Ireland and by engineering in Spain. In other words, the information disclosed by companies differs between countries, which could lead to errors if the quality of the websites’ content reflect sectoral reality.

The data obtained through the industry analysis indicate that:

- The sector’s value chain can be represented, based on the following activities:

1. Services,
2. Manufacturer,
3. Design,
4. Components,
5. Software,

6. Training,
7. Consultancy,
8. Distribution,
9. Rental of equipment.

- The services drone businesses offer to other companies and customers can be identified, and the following were identified as being the most common:

1. Inspections,
2. Mapping,
3. Aerial photography,
4. Aerial video,
5. Surveying.

- The sectors drone businesses target with their offer can be identified. A classification could, for example, include:

1. Agriculture,
2. Infrastructure,
3. Architecture,
4. Real estate,
5. Construction,
6. Television, etc.

- Identifying the services offered and the sectors targeted by the drone businesses allows one to establish a link to those sectors' NACE codes and thus to classify the companies according to those codes.
- It is possible to ascertain drone client companies' specialisations as well as their geographical area / location.
- Economic variables, such as the number of employees or turnover, are not usually available on the websites.

The exploratory analyses performed were essential to defining the keywords that are to be used by the web search strategy (see Chapter 2), to the classification model applied to drone companies (see Chapter 4), and to data extraction (see Chapter 5). Search engines do not perform well if too many different words are used in a query. The selection of search terms should therefore be based on the preliminary analysis on a small sample. Keywords will also be needed in the classifying phase, given that those terms will be included in scripts to search for them on the websites.

2 Approaches for searching for companies websites

2. Approaches to searching for company websites

Searching for the websites of companies that are active in a specific sector and in a particular country can be accomplished in several ways. The approaches explored for this study are discussed in this section. But first, some introductory comments and general remarks are provided.

When using a search engine such as Google, Bing or DuckDuckGo, it was determined that the findings can be affected by:

- the location (IP address) of the searching user,
- the user's previous search history (cookies),
- the country extension of the search engine used (e.g. using Google.nl vs. Google.ie), and
- the User-agent of the 'browser' program used.

These effects can be reduced (although not entirely) through:

- using a VPN connection,
- using a browser that has no search history or searching the web via an anonymised (incognito) browser, and
- using a search engine that is specific to the country under study.

For the initial studies described below, a VPN connection with a Luxembourg ([10]) location and a generic Mozilla browser User-agent identifier were used to mimic, as closely as possible, the search results of a browser program used at Eurostat's premises. All queries were coded in Python Version 3.

The following search strategies were identified and investigated.

([10]) For subsequent tasks the program's VPN connection was set with a location in Belgium.

2.1. Searching for websites that provide an overview of drone companies

The first step of this approach involves choosing several drone-specific terms, a term for 'list', 'overview' or 'pdf', and the country's name. For countries, in which the main spoken language was not English, two strategies were used. The first made use of search terms in the main spoken language including the word drone, drone abbreviations, and other words specific to several activities. The second approach used the same words but all written in English. The first 50 links provided by the search engines were checked and any links referring to websites with country codes other than the country under investigation were removed. For example, when studying Irish websites (such as www.drone.ie) any link referring to a UK website (such as www.dronedirect.co.uk) was removed. Please be aware that any overview of drone companies may be included in webpages (html-files), pdf files or both. In both cases, it subsequently becomes necessary (where available) to extract the references to websites. Other than references to websites, the names of companies were also identified and extracted. This step was followed by a search for the URLs corresponding to the company names obtained.

Findings:

- 'Pros': This is a very efficient strategy from a search-engine perspective. Only one search query is needed. It can work very well on specific countries, in particular when national air safety authorities publish overview lists of drone operators (e.g. Spain).
- 'Cons': This strategy does not work for all countries. For example, no such overview website could be found for Germany and Italy, indicating that this strategy may never be the only approach to follow.

2.2. Searching for websites of individual drone companies

This approach begins with the selection of several terms that are specific to drones, as well as the country's name. For countries, in which English is not the main language, two search strategies were applied. In the first, the search terms used, except for the drone abbreviations, were all written in the country's main spoken language. In the second strategy, all words were written in English. All links found by the search engines (up to between 200 and 300) were checked, and any links referring to websites with country codes other than the country being investigated were removed.

Findings:

- 'Pros': This strategy enables one to find websites of individual companies. This may include those that are missing the from overview websites.
- 'Cons': This strategy requires considerable amounts of search requests to be performed, which calls for the necessity to regulate requests over time, so as to prevent the search engine used from blocking. At the time of drafting, it is unclear why only up to 300 websites could be found.

2.3. Finding drone websites via pictures of drones

This approach begins by searching for pictures of drones in a specific country. This is done by limiting the search query to image-type results. Next, the domain names of the links of the images found are collected. The results were not satisfactory.

Findings:

- 'Pros': Using this strategy, websites are found that contain pictures of drones.
- 'Cons': Because of the image-oriented approach, considerable amounts of websites are detected, on which drone-enthusiasts post pictures. These are however often blog-like websites and the websites of other drone fora. Compared to strategies 2.1 and 2.2, a fairly limited number of drone company websites were found.

2.4. Finding drone websites via a company's impressum page

This approach focuses on finding the 'impressum' pages of companies containing drone-specific words in a given country. This is especially interesting for Germany, for example, where many companies are legally required to have an impressum page on their website. Such pages provide legal and public information to customers, in a standardised way. Basically, a search query is performed including drone-specific words, the country's name and the word 'impressum'.

Findings:

- 'Pros': Using this strategy, websites of drone companies are found to include drone-specific words on their *impressum* webpage.
- 'Cons': Depending on the country, only a limited number of companies may have an *impressum* page. The page in question must also include a drone-specific word. Due to that, a limited number of companies are usually identified, compared to search strategies 2.1 and 2.2.

2.5. Finding drone websites using a non-search engine approach

Apart from using a search-engine based approach, one may also attempt to find drone companies in another way. For example, one can search for open access Chamber of Commerce data as well as for websites that provide overviews of the companies that are active in a specific region of a country. Using drone-specific words as search terms, one may find the names of companies active in the drone industry, when such words form part of the company name or if they are included in the short description (if provided).

Findings:

- 'Pros': Compared to the other strategies, this is – currently – the only way in which to detect drone companies without a website.
- 'Cons': The websites searched may not include URLs or they may provide a limited amount of URLs for the companies listed. This raises the additional need for a search strategy to find those companies' URLs (if available). In the ESS, this constraint it tackled through the search strategy that was already developed in the context of the ESSnet Big Data II Workpackage C – Enterprise characteristics.

In addition to the findings listed above, a number of additional observations were made during the study's inception phase.

The first is the need for extracting URLs from documents including websites/html files, pdf-files, etc. Specific approaches need to be developed for each type of file.

The second is the need to identify and extract company names from documents providing an overview of drone companies, so as to determine those companies' URLs (if available). This often resulted in finding copious amounts of names, including those of many non-drone companies, whose URLs subsequently needed to be searched for.

Thirdly, it is necessary to develop a method able to determine, whether an identified website really belongs to a company active in the drone industry. Because the search strategy is applied in multiple countries, in which different languages are spoken, it is necessary to develop a word-based set of rules to discern between drone and non-drone companies. The aim is to have a specific set of words available to use to identify a drone website in other European countries. This approach was implemented at the end of the search strategy, as it was found that, toward the beginning of the search strategy, it is much more efficient only to remove URLs that clearly do not correspond to companies active in the country under study.

Lastly, an intriguing question arises: how exactly does using a search engine determine, whether a website can be found or not? Imagine a small drone company that has only recently created a website which, so far, has not attracted any visitors. It is unlikely that such a website would be covered by any of the strategies making use of a search engine. The latter must be correct as a website's popularity is the result of the number of links to that website found by the various search engines. Because companies must register with their country's Chamber of Commerce, it is extremely likely that the type of 'young' company described may only be detected by the fifth search strategy. The same holds for companies that do not have a website.

3 URL retrieval

3. URL retrieval

This chapter provides a description of the data retrieval approach developed and the implementation choices made. It begins with a section on the composition of URLs, and further focuses on the scripts and methodology developed to collect links to websites of drone companies, in a specific country and for a specific language. Furthermore, this chapter provides an overview of the results obtained for Ireland, the Netherlands, Spain, Germany, and Italy. This section is complemented by Annex I, which provides details on how a list of drone company websites was created for Spain, as a benchmarking framework (i.e. control list) to check the results provided by the different versions of Script 1 and, further, to develop of the method used to detect and classify drone company websites (see Chapter 4).

3.1. URL composition explained

It is important to understand the structure of a URL. A URL refers to a specific location in the World Wide Web. The composition of a URL is illustrated by means of the following example:

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

- The first part of the URL is the **scheme-part**. For web pages, this is usually http:// or https://. It is often followed by 'www.', which 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**. In the example, it is 'gopa'. This part of the domain usually refers to the organisation that registered the domain name. This is the most important part of a URL, from the project's point of view, as it − very likely − specifically represents the website of an individual company.
- The second-level domain is followed by the **top-domain**. In the example, it is '.lu'. That abbreviation refers to web pages located in Luxembourg. A complete list of country-specific top-domains can be found in the web (Wikipedia, 2021). Other non-country-specific top-domains often used are '.com' for companies and '.org' for organisations, 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 where the html-file is located, on the Gopa server.
- Sometimes, 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, a question mark is often included to indicate the beginning of a query provided by the words following that symbol.

3.2. URL collection methodology

This section discusses the most important decisions made in developing the scripts used to search the web. Because topics, sub-topics and decisions made affect one another, it has been challenging to find a specific sequence, in which the topics should be discussed. It was decided to discuss them in the order shown below. When discussing each topic, dependencies are described by referring to the other sections. In addition, since queries were also improved during the course of developing the methodology, the results of the previous versions of the queries are discussed, whenever relevant.

3.2.1. SEARCH QUERIES

Two search approaches were implemented. The first searches for individual websites of drone companies (search implemented through Script 1) ([11]). The second searches for websites containing lists of drone company names and/or websites (search implemented through Script 2) ([12]). The other search approaches considered (see Chapter 2) were found not to provide additional information.

Based on the exploratory sector analyses, and after experimenting with different search words and query compositions on a number of search engines, a decision was made as to the composition of the search queries used for the two main approaches developed (see Section 3.1). For each Member State, queries were developed in the dominant language as well as in English. The exception was Ireland, where only English queries were used. In principle, each search query consists 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 under research, the word drone is replaced by the native word used, e.g. ‘Drohne’ in Germany and ‘dron’ in Spain. The acronyms used for drones are the same in each query and for each country, i.e. uas, uav and rpas. When individual URLs are searched for, each query always contains a single acronym representing drones. When sites providing overviews of drones were searched for, the words referring to drones are included in brackets and separated by a space and the word ‘OR’, for example: (drone OR uas OR uav OR rpas).
- The second part of the query contains the word (or words) indicating either a company, a type of activity, a synonym of membership or the word ‘pdf’. Examples of the first are: company, business or shop. Examples of the second are: pilot, training, manufacture, service provider, etc. These are examples of words used, when searching for individual websites of drone companies. In the third case, websites containing 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 the native language in use. One can imagine that, by discerning more types of activities, more specific search queries are used. When documents providing overviews are searched for, the word pdf is additionally included.
- The last part of the search query contains the name of the country searched for, in the language used. For the five Member States studied, in English, they are: Germany, Spain, Ireland, Italy, and Netherlands. In their native languages they are, respectively: Deutschland,

([11]) Script 1 is available in the GitHub repository at:
https://github.com/eurostat/wih_drones_companies/blob/master/Script1_IniS2.py
([12]) Script 2 is available in the GitHub repository at:
https://github.com/eurostat/wih_drones_companies/blob/master/Script2_IniS.py
See also

Table 3: Overview of scripts developed: input, output, script deployed and run-time and the dedicated GitHub repository at: https://github.com/eurostat/wih_drones_companies

España, Ireland, Italia and Nederland.

The advantage of this three-part construction of the queries is that each part can easily be adjusted to investigate other topics. During the course of 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 tilda 'ñ' in the word España. When comparing the findings for queries carried out on the other search engines, with and without diacritical marks, no obvious differences could be found. Hence, it was decided to remove all diacritical marks, from all queries.

3.2.2. SEARCH ENGINES

There are various ways of searching the web. The most often used search engines, also known as the 'big four' ([13]), are Google (google.com), Bing (bing.com), DuckDuckGo (duckduckgo.com) and Yahoo (yahoo.com) (Biggs, 2021; Reliablesoft, 2021). They are also the most important search engines, when searching for companies in Europe. Code was developed to extract links from each of those search engines by means of queries (see Section 3.2.1). In addition, Python libraries were found and included in the code, that also facilitated using those search engines. By opting for that approach, one becomes more flexible and can easily access a search engine in a different way (compared to the first approach). One should note that many of the search engines seek, as much as possible, to block any − suspected − automated access. In addition, a number of other search engines, namely Ask (ask.com), AOL (aol.com), Baidu (baiduinenglish.com), and Yandex (yandex.com), were tested. AOL and Ask provided interesting results and, hence, they were also included in the code. Regarding Baidu and Yandex, the first is the most popular search engine in China and the second is the most popular search engine in Russia. During the exploratory work it was found that the results obtained did not really focus on the 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 were selected: Google, Bing, DuckDuckGo, Yahoo, AOL, and Ask. When Google and Yahoo are used, the search page that is specific to a country is selected, e.g. for Spain, google.es and es.search.yahoo.com are accessed, respectively. Such an option was not available with the other search engines selected.

One may wonder what the benefit is of using six different search engines, when searching the Web. An example reveals the importance thereof. Table 2 shows the overlap between the domain names of the links found, when applying all 48 Version 2 search queries of Script 1 for Spain. In the table, the results for both the English and Spanish queries are combined (see further for additional details). For each search engine, the maximum number of organic search results ([14]) (links/URLs) returned was collected; this was done by extracting all URLs from all pages available, ignoring commercial links as much as possible. Depending on the search engine and how it reacted, when used intensively, the IP-address from the machine from which the searches were conducted was either kept the same or not. It is also important to note that − in practice − much fewer links were in fact obtained than the large number of links reported on the first page by (some of) the search engines. For example, Google reported 3 240 000 results on the front page of the first search query for Ireland while, on its front page, Bing stated that 326 000 results had been found for the same query. In practice, the actual number of results that could be extracted was found to be much lower; usually, between around 200-300 and 400-500 links maximum could be extracted. The number of results varied considerably, depending on the search engine used. Usually, for individual queries, Ask provided the smallest number of links (~55), while AOL and Yahoo provided the largest (400-550). After running Script 1 for all 48 Version 2 queries, for Spain, on all six search engines and in both relevant languages, the links found on each search engine were combined and deduplicated. When comparing those 'raw' (unprocessed) search results between the search engines, an unexpected, fairly low amount of overlap could be observed (Table 2). This, in part, was found to be due to differences in the non-domain name part of the links. The result could therefore be improved. This

([13]) 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.
([14]) Organic search results refer to the non-sponsored (commercial) links found.

was done by only focusing on the combination of secondary domain names and on the top-domains returned, e.g. 'dronesbarcelona.es' instead of the complete link 'https://www.dronesbarcelona.es/prensa'. This had the additional advantage that the comparison could focus on the part of the link that is most logically indicative of a company website. The results of pre-processing the links are shown in Table 2 for Spain. The results obtained using Version 2 of Script 1 for all other countries confirm similar differences.

Table 2: 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

Note: both English- and Spanish-language search results are combined

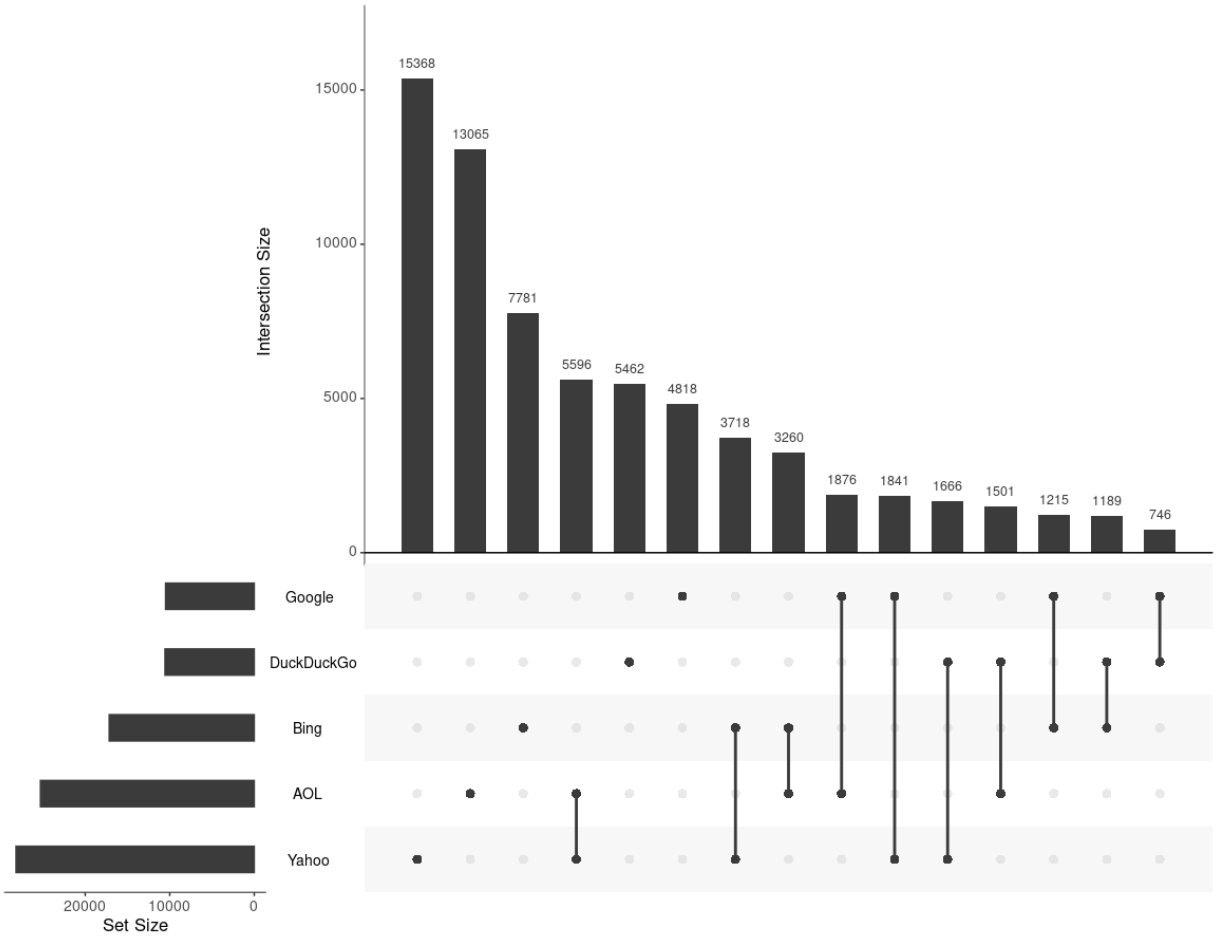
Search Engine	AOL	Ask	Bing	DuckDuckGo	Google	Yahoo
AOL	(12 055)	1 010	3 260	1 501	1 876	5 596
Ask	1 010	(1 473)	749	529	1 179	999
Bing	3 260	749	(7 032)	1 189	1 215	3 718
DuckDuckGo	1 501	529	1 189	(4 933)	746	1 666
Google	1 876	1 179	1 215	746	(3 639)	1 841
Yahoo	5 596	999	3 718	1 666	1 841	(14 369)

Source: authors' elaboration based on data processing results

The results in Table 2 indicate that clearly different results are obtained for the different search engines. In the diagonal, the number of unique results is shown for each search engine in brackets. Overall, a total of 27 712 unique combinations of secondary and top-domain names were found. From Table 2, it is clear that Yahoo returned most unique combinations, i.e. 14 369. It was followed by AOL and Bing, with 12 055 and 7 032 unique combinations respectively. Next came DuckDuckGo, Google and Ask, with 4 933, 3 639 and 1 473 results.

The two search engines providing the most similar findings are AOL and Yahoo. There is an overlap of 5 596 items between those two. 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 of visualising the overlap between the various search results is to create an upset-plot; this is an alternative to a Venn-diagram, when a dataset consists of more than three subsets. The upset plot for the data in Table 2 is shown in Figure 1. Here, the large number of findings from Yahoo, compared to those returned by the other search engines, immediately becomes apparent. The reader should however note that those results do not necessarily indicate that the search engine returning the most links is, in effect, finding the majority of the Spanish drone websites. The latter is discussed in Section 3.3.

Figure 1: Upset plot of the overlap between the secondary and top-domain names found by the six search engines for the English and Spanish results combined (Spain Script version 2)



Very similar − yet not identical − findings to those shown in Figure 1 are found, when the importance of each search engine is measured by the total number of unique results returned. Those links are moved to the final llist, and the remainder is determined for the links provided by the other search engines. Here, Yahoo is the most important contributor (14 369 unique links), followed by AOL (6 459), DuckDuckGo (2 965), Bing (2 537), Google (1 277), and Ask (270). Comparing to the overall number of domain names found, this makes DuckDuckGo a more important contributor than Bing. Google and Ask provide the lowest number of unique links.

All in all, from the results described above, it is clear that, in order to maximize the number of links found, the findings yielded by all search engines need to be combined. Because of access issues with some of the search engines during this work (access may regularly have been blocked), API access to Google was purchased, so as to enable the continuation of work for the remainder of the project. For that purpose, the service provided by Valueserp.com was used, as it was found to be the cheapest. Other paid services were looked into but they were found to be unsatisfactory, as many of them lacked paid access to the majority of the search engines (obtaining paid access to DuckDuckGo is a challenge), and those that did provide access did not enable the scraping of multiple pages of the same query (via their API). Hence, for the other search engines, only one alternative remained: scraping via a VPN-connection.

For most of the other search engines, for each query, the connection's IP-address needed to be changed to prevent blocking; this was particularly relevant, where DuckDuckGo, Yahoo, AOL, and Ask are concerned. The VPN connection provided by Surfshark was used for that purpose. While performing this work, it was found that the Bing search engine could be accessed without changing the IP-address, as was also the case for paid Google access. For the other search engines, VPN-based queries with IP-rotation between queries was necessary, in order to perform the first two steps of scraping without encountering issues (see Section 3.2.1).

3.2.3. SEQUENCE OF STEPS

Based on the findings resulting from the assessment of the approaches proposed for searching for company websites (see Chapter 2), the two most promising approaches for collecting data from the internet on drone companies in various countries by using search engines, were the two starting points for searching the web, namely:

1. Collect URLs of individual drone companies using search words.
2. Collect URLs of websites providing overviews of drone companies.

During the course of work, it was however found that, apart from links to web pages (html pages), links to PDF-documents were also being obtained. Often, those PDF-files contained (a section with) links to websites or they contained the email addresses of potential drone companies. In addition, they always contained the names of companies. This indicated the need, not only to extract links from web pages, but also from PDF-files, which may provide i) web-links, ii) domain names that can be deduced from email addresses and iii) the names of companies that could be used to find the corresponding URLs. 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 such PDF-files. The PDF-files referred to subsequently needed to be downloaded from the web, after which any links to web domains (both URL and email based) and company names could be extracted. The five following steps were therefore established:

1. Collect URLs of individual drone companies using search words (web- and PDF-links).
2. Collect URLs of websites providing an overview of drone companies (web- and PDF-links).
3. Extract URLs from the PDF-files found in Steps 1 and 2.
4. Extract company names from the PDF-files found in Steps 1 and 2.
5. Search for the URLs corresponding to company names extracted from the PDF-files in Step 4.

The flowcharts of the scripts created for these tasks are displayed in Figure 2. The inputs of the first two scripts concern the country and language used, together with the search queries specific to each task. The second script requires additional input: a list of words indicative of the country, words indicative of drones and words indicative of being a member of an organisation. They may also include the word pdf. The maximum number of sub-pages scraped per domain is also provided, as is a limit to the number of search results. The first script is used to select and extract as much information as possible from the sites found to be indicative of providing overviews of drone companies, while the second was used to prevent that a very large number of web pages needed to be checked and scraped. The settings can all be adjusted via the country- and language-specific Ini-file. The Ini-file must be imported by each script upon execution. The first PDF-script (Script 3a), fed by the PDF-link files produced in Steps 1 and 2, requires words that are indicative of the country studied and words for drones. The output of Script 3a is a list of the URLs extracted from the PDF-files found and a list of company names. After Script 3a has run, Script 3b is run to perform a search for the URLs corresponding to the names extracted. The latter script only uses the Google search engine, and it collects the top 10 or 11 search results provided for each name. In the end, a total of 5 URL lists are produced, which are then used as inputs for Script 4a. The flow charts in Figure 2 show how each script works, and what the inputs and outputs are. For the reader's convenience, Table 3 provides an overview (for the final version of queries for Scripts 1 and 2) of the in- and output of each script, their dependencies and an indication of their runtime.

Table 3: Overview of scripts developed: input, output, script deployed and run-time

Script	Input	Output	Script depen-dence	Run-time (days)
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
3a. 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 - Extract company names from PDF-files	- Potential links to drone websites (including those derived from email addresses) - Log file	Scripts 1 and 2	~ 0.5
3b. Search for URLs	- Ini-file (with country and language settings) - Names extracted from PDF-files	- Potential links to company websites (max. 10-11 per name)	Script 3a	~ 1
4a. Cleaning and social media check	- Ini-file (with country and language settings) - List of domain extensions and country names (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 in the country (enriched and cleaned) - Log file	Scripts 1, 2, 3a and 3b	~ 0.5
4b. Content and location check: identifying drone websites (by removing the obvious non-drone sites in other countries)	- Ini-file (with country and language settings and selection criteria) - List of domain extensions and country names (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 and text extracted from page - Log file	Script 4a	~ 1.5
5 Merge script	- Ini-file (country's native language) - Files produced by Script 4b for both languages	- CSV file of combined and deduplicated website links	Script 4b	0.1 (< 3 hours)
6. Classification script	- Name of merged file produced by Script 5 - Stored classification model - Optimally: translation file	- CSV file with website classification scores (drone or not)	Script 5	< 0.5
7 Information extraction script (country specific)	- Name of classification file of Script 6 - File with city names, area codes and regions, and their relation	- CSV file with information on websites classified as drone companies active in country studied (probability > 0.6)	Script 6	0.06 (~ 2 hours)

Through Scripts 4a and 4b, the lists of URLs produced in the previous steps are combined, deduplicated and checked. During the checking stage, 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 links can also point to company-specific accounts on those platforms. The latter are interesting as those pages may contain: i) information on the company's location, as well as – potentially − ii) the URL of a company's actual web-page (the 'www-page'). A social-media specific check was included with the aim of finding the 'account's' location and of extracting the link to its www-page (if available). This was a challenging task, as social media platforms tend to spend much effort on blocking scripted access; this is especially the case with 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 information could be collected. Any accounts located in the Member State researched or the location of which was unknown (undetermined), were additionally checked for content of a URL (linking to the www-page). If such a link could be found, and if it had not already been detected by Scripts 1 and 2, 3a and 3b, it was added to the list of URLs. The latter proved to be the case for around 2 % of the social media links found. Because this task needed to be performed prior to analysing the URLs found in more detail, a separate script (Script 4a) was created for this specific task. It also deduplicated links and included most of the checks performed for the removal – as early in the process as possible – of the less relevant links. The input for this script were the 5 files containing URLs, produced by Scripts 1, 2, 3a and 3b, lists of words indicative of the country, drones and membership, as well as a list of names of a country's (largest) cities. Both those lists were written in the native language and, where applicable, city names were expressed in English. The latter was done as some cities may 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 named 'Munich' in English. Wikipedia and GISCO Eurostat datasets were used to construct those lists and to find the largest cities' English names (an alternative source is geonames.org). In addition, a list containing the names and abbreviation of US states was constructed, together with a list of the world's largest cities and their respective countries. The latter contained around 1 100 city names. The exact steps followed in the location search are further described below.

Once Script 4a had been run, the large list of unique URLs obtained needed to be thoroughly checked. In all cases, for all countries, this step took the longest to perform. First, it was checked, whether the website still existed. If that was the case, the site was visited, and its location and content were investigated. It was attempted to determine both the location (is it located in the country studied or not?) and the content of the website (is it the website of a drone company?). A stepwise approach was taken for the location search, which 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).The sequence of steps applied is:

1. Check the URL for inclusion of the studied country's top-domain name (e.g. '.ie' for Ireland) – True.
2. Check the URL for inclusion of another country's top-domain (e.g. '.nl' when researching Ireland) – False (requires a list of country-specific top-domains).
3. Check the web-page's text for words specific to the country studied or its language (e.g. Ireland, Éire and Irish, for Ireland) – True.
4. Check the web page's text for words specific to another country in the world (e.g. Netherlands, when studying Ireland) – False (requires a list of country names).
5. Check the web-page's text for city names specific to the country 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 we are' 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. Check web-page text or social media account names for US-state names or abbreviations and check whether one of the largest cities in the world is included – False or True.
9. Decide that the location of a web page cannot be determined with certainty – Unknown.

Once a web page has been identified as being located in the country under research, or as being unknown, the text of the page originally visited (usually a domain's main page) is studied in more detail. After removing stop words that are specific to the language, the text's top 10 words are extracted and stored. This results in a file containing the original URL visited, the actual URL visited, the result of country location (True, False or Unknown), the top 10 words and the action taken by the script (e.g. web page does not exist, etc.). Overall, after running Scripts 1 to 4b, one obtains a list of URLs of potential drone companies ([15]).

The next step specifically applies to countries, for which searches are being performed in several languages, e.g. Spain and Italy. Script 5 was developed specifically for this task. It combines the findings of the Script 4b files produced for both languages. Here, the URLs found are compared between both files. Any URL that is not included in the native language file and that is found in the English language search is added to the first list. The script essentially deduplicates both lists with a preference for the findings that are in the studied country's native language.

Script 6 subsequently classifies the text from the website of each URL found by means of the model identifying drone websites (more details on the model's development are provided in Chapter 4). For each website, the probability of its being a drone website is returned (value between 0 and 1). A website with a probability of 0.6 or higher is considered as being a (potential) drone website in the subsequent steps. All other websites are ignored. As a single model containing English words is being used, the website text needs to be translated, prior to classification. Given that the model was developed for Spanish websites, translating their text into English was not an issue. The same holds for Irish websites, which are all written in English. A translation list was produced for the task of translating Italian websites. This list forms part of the input of the classification script. For all other countries, only websites written in English can be classified at the moment. This is the reason for which the current study is limited to Spain, Ireland and Italy.

Finally, Script 7 is used to extract information from the websites assigned a probability of 0.6 and higher by the classification model. The information extracted is used to determine, whether the URL is actually that of a drone company active in the country studied. This is the end result of the approach described above.

([15]) See for further details, Deliverable D2 – Data retrieval available here: https://ec.europa.eu/eurostat/cros/system/files/d2_dataretrieval_final.pdf

3.3. Results

This section provides an overview of the number of links found, when searching the web using drone-specific queries for the five Member States studied (Spain, Ireland, Italy, the Netherlands and Germany). Table 4 shows the total number of unique links found by applying Scripts 1-7 for each country and language (if applicable). The scripts are located on GitHub and can be found here: https://github.com/eurostat/wih_drones_companies.

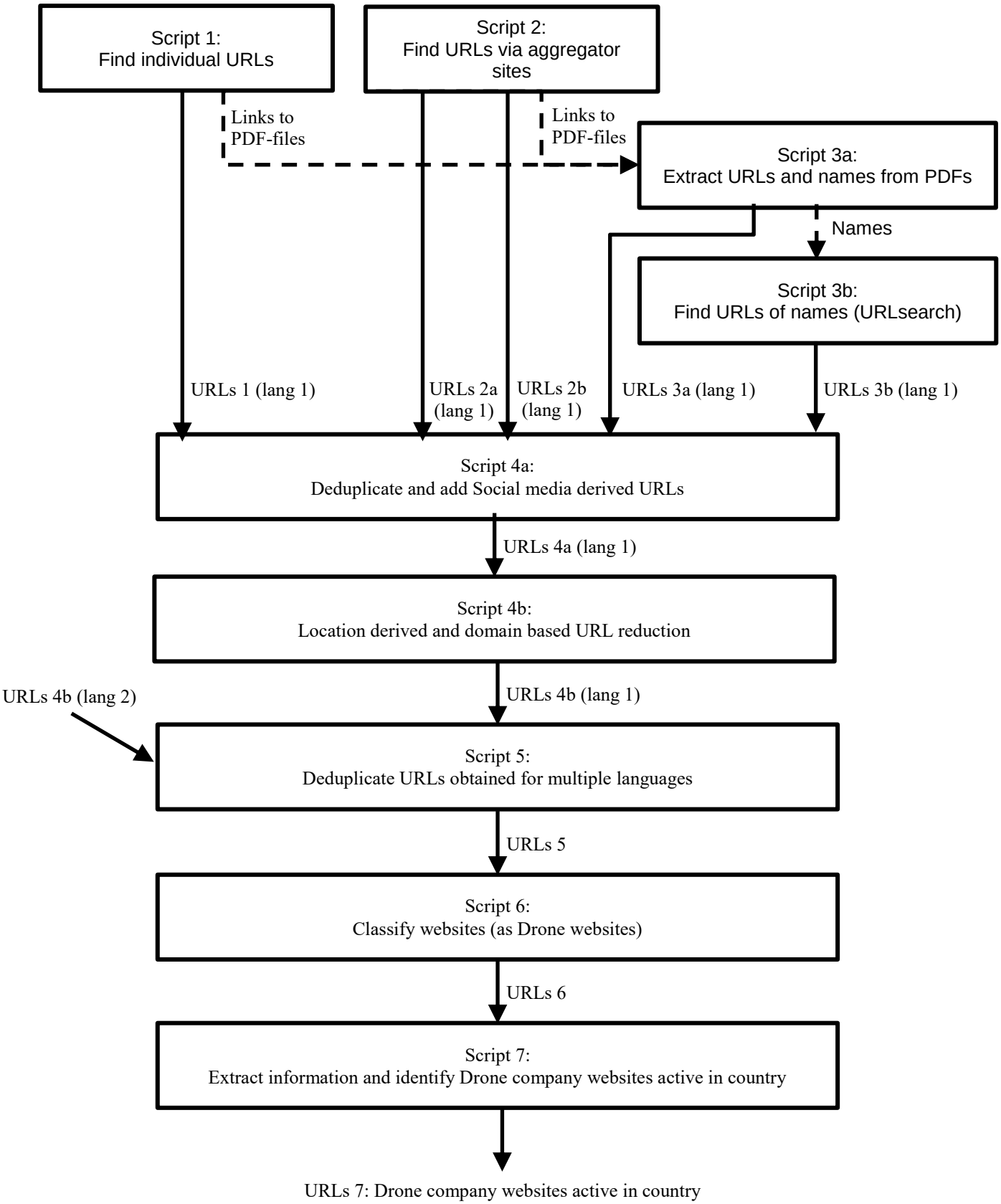
Table 4: Number of links found by the final version of each script, for each Member State studied and for each language

Script	Links found	Ireland EN	Netherlands EN/Dutch	Spain EN/Spanish	Germany EN/German	Italy EN/Italian
Script 1	Web-links	29 958	30 303 31 138	33 274 31 546	30 154 34 234	29 058 21 848
	PDF-links	878	901 815	1 027 24	1 026 1 251	768 485
Script 2	Web-links (a)	56 513	23 113 26 166	22 608 46 542	29 767 53 444	53 937 56 906
	Web-links (b)	182	141 334	134 306	162 434	105 356
	PDF-links	1974	1 533 3 584	1 861 9 541	2 217 8 798	2 421 12 281
Script 3	Web-links	6 115	7 165 4 369	5 886 1 957	11 330 17 013	7 011 3 996
	Name-based	2 816	4 768	7 065	16 463	34 185
Script 4a	Web-links	48 950	41 794 45 390	47 980 49 201	63 547 76 210	115 107 115 253
Script 4b	Web-links	14 568	43 446 47 657	46 981 47 101	68 023 80 859	112 901 112 066
Script 5	Combined	14 568	28 000 (EN: 12 907)	26 067	43 412 (EN: 13 994)	53 785
Script 6	Class. prob ≥ 0.6	681	1 069 (only EN)	1 874	770 (only EN)	1 608
Script 7	Selected	66	-	461	-	353

When one looks at the output of each script, it is clear that the search strategy implemented in Scripts 1 and 2 varies per country. The number of URLs is higher for Script 2, when a Drone overview website is available in a country. In all other cases, Script 1 finds most unique links. When all goes well, Scripts 2 and 3 can both be run in 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 complete, and Script 4b takes

between 1 and 2 days to complete (see Table 3). The number of URLs for Italy and Germany are especially high. Combining the output of Scripts 1-4b in both languages, for all countries except Ireland, reveals that there is a considerable overlap between the URLs found by the search strategies in both languages. This overlap varies between 48 and 62 %. After classification (see Script 6) only the URLs allocated a probability of 0.6 and higher are selected. This seriously reduces the remaining number of URLs, which varies between 3.0 and 8.2 % of the URLs initially provided (only findings for English websites are taken, in the case of the Netherlands and Germany). In Table 4, the end result of URLs of drone companies active in the country, the end result of Script 7, are only included for Spain, Ireland and Italy. Final numbers vary between 9.7 and 24.6 % of the number of URLs in the previous step. Clearly, many URLs found are not those of drone companies active in the country, and the broad range indicates some country-specific issues, as was confirmed during the course of development. Issues, such as extracting the address in an appropriate way and extracting VAT numbers, were solved as well as possible by writing country-specific scripts. Many parameters, regular expressions and a language-specific input were found to be required. Solving this through Ini-file settings, as was done for Scripts 1-5, required an overview that has not been achieved at this point in the project.

Figure 2: Overview of the scripts developed to find Drone companies in various countries



4 Classification of drone company websites

4. Classification of drone company websites

This section describes in detail the classification model developed to detect drone company websites. Given that a list of drone company websites had already been constructed for Spain ([16]), the modelling approach initially focused on those data. The data used for the model's development were: (1) the texts of the manually identified Spanish drone websites (1 098 in total) ([17]) and (2) the texts of the websites found by the webscraping scripts developed for Spain. The data collected in the 'Data retrieval' part of the project (see Chapter 3) were used for this work, overall. This section also introduces the alternative website classification approaches explored. It discusses the model's generalisability and its extension to other European countries, as well as the pre-conditions that entails.

4.1. Drone website classification approaches

The three options considered when identifying drone websites are discussed. Given that a list of manually identified drone company websites was available and that additional websites were being searched for, modelling efforts initially focused on the data compiled for Spain.

When developing a drone website classification approach, the most important considerations are that the method:

1. is able to discern between a drone and a non-drone website as accurately as possible,
2. has a high precision and a high recall, and
3. should also be applicable to the other countries included in the study.

The data used to develop the model are the texts of the manually identified Spanish drone websites (1 098 in total) and the texts of the websites collected by the webscraping scripts developed for Spain (see Chapter 3). The texts of the websites collected by webscraping are those of the combined lists of websites obtained after applying Scripts 1, 2, 3a, 3b, 4a and 4b for Spain, while searching in both the Spanish and the English language (as described in Chapter 3). A total of 47 097 unique website URLs were obtained, as a result, of which 31 106 contained texts. It must be pointed out that the geographical location of the drone companies whose websites were found was not limited to Spain. Up to this point in the study, any information on location found on websites had been ignored

([16]) See Deliverable D2 – Data retrieval available here: https://ec.europa.eu/eurostat/cros/system/files/d2_dataretrieval_final.pdf
([17]) See Annex I for an overview of how a 'control list' for Spain was created.

given that the focus of the model's development was on detecting drone company websites and distinguishing them from the plethora of other websites found by the scripts. Location detection and assignation is further fine-tuned in the data extraction phase (Script 7).

CLASSIFICATION APPROACHES CONSIDERED

One can consider various ways of developing a method able to discern between drone and non-drone websites. From the whole realm of binary classification options available in Python (Pedregosa et al., 2011) in combination with the data available, and including the personal experiences of the Statistics Netherlands researchers involved, the most promising options were selected and tried. These are:

1. **Word-based**: Determine and count the occurrences of words expected to be specific for a drone company website. The words used in the search-engine-based part of this project are the most logical choices. Therefore, the effect of checking for the words 'drone/dron (Eng/Spa)', 'rpas', 'uas', and 'uav' were tested.
2. **PUlearning**: Develop a semi-supervised machine learning approach using a set of labelled positive and unlabelled examples. This kind of approach is known as Positive-Unlabelled learning (PUlearn, 2021) and has the advantage that *only* a labelled set of positive examples, e.g. drone websites, needs to be available. The texts in the set of positive examples are used to distinguish two groups in the unlabelled dataset. The first group are those with texts similar to the known positive group (if all goes well these are the new positive examples) and the other group is expected to be the group of non-positives (i.e. the negative cases). The *PUlearn* algorithm aims to find a way to maximally distinguish the group similar to the positive cases from the other (the non-positive/'negative') examples.
3. **Supervised ML**: Develop a machine learning approach using a set of labelled positive and negative examples. This is an example of supervised machine learning. This approach requires a set of labelled positive *and* negative examples. The differences in the texts of the positive and negative examples are used to create a model that aims to discern both groups as well as possible.

Each of those three approaches was studied and applied. The best results obtained by the trained algorithms were manually checked by experts.

4.1.1. WORD-BASED APPROACH

The data eventually collected by script 4b for Spain, after combining the Spanish- and English-language datasets and removal of duplicates, were checked for the occurrences of the words 'dron/drone', 'rpas', 'uas', and 'uav'. The raw html-file obtained after scraping each website was parsed with the Beautiful Soup 4 library followed by removal of all script and style sections. Subsequently, the visible text was extracted, ignoring any html-code and for each website text, the total number of drone word occurrences was counted. A random sample of 50 was drawn from the websites containing no drone words. A random sample of 50 was also drawn from the websites with one or more drone words. The samples of 50 non-drone and 50 drone websites were manually checked by experts to determine how many actual drone websites were included. The findings are listed in Table 5, in which measures of Accuracy, Precision and Recall are used to assess the performance of the classification approach. Accuracy is calculated as the number of websites classified correctly, out of the total number of websites found. Precision is given by the number of relevant websites identified divided by the total number of websites found. Recall is given by the number of relevant websites identified divided by the total number of existing relevant websites. The closer the indicators are to 1, the better the algorithm is performing.

Table 5: Manual classification results for the drone-specific word-classification findings

Word-based	Accuracy	Precision	Recall
All	0.75	0.83	0.62
Positives only	0.62	-	-
Negatives only	0.87	-	-

These results show that simply applying a word-search to detect the occurrence of drone words in a website's text is not a particularly accurate method (only 75 % accuracy). The low recall measured for this approach indicates that many websites will be missed when using such an approach. The higher accuracy in detecting negative cases indicates that the method works well in identifying non-drone websites but, even then, 13 % are being missed. More advanced methods clearly need to be applied in order to reach a more optimal approach.

4.1.2. PULEARNING APPROACH

As a list of known drone websites was available for Spain and the combined results of Script 4b were all unlabelled, it was interesting to study a semi-supervised machine learning method specifically developed for the combination of such datasets; the method is known as Positive and Unlabelled Learning. Such an approach aims to find the features that are specific to the set of positive examples. Based on those features, it attempts, as well as possible, to separate the group having those features (the 'positives') from the other (the 'negative') group, in the unlabelled data. Apart from the two datasets and the PUlearn algorithm, a machine learning classification method able to produce probability estimates for each class is required as an input (PUlearning, 2021). Both Logistic Regression and Support Vector Machine classification methods are able for the calculation of probabilities. Both methods were therefore tested using various hyper-parameter settings, and their findings were compared. In the hyper-parameter settings, it was found that a minimum document frequency of 100 and a minimum character length of 3 for the words included worked best.

So as to enable the algorithm to classify websites, the texts extracted needed some additional processing. First, the text's language was determined using the langdetect library. Given that most pages were either written in Spanish or in English, only those two languages were discerned, i.e. any non-Spanish page was considered to be written in English. Subsequent processing was performed by converting all words extracted to lower case. All punctuation marks and numbers were removed. This was followed by the removal of all words of less than a specific number of characters; less than 3 worked best. Next, depending on the language detected, all words included in the Natural Language Toolkit (NLTK) stop-words list for that language were removed. The processed texts were subsequently converted to a Document Term Matrix (DTM) which enabled the development of a model by using the well-known representation in the form of frequency-annotated bag-of-words (Aggarwal 2016, Ch. 13). The DTM was composed of rows, one for each website, and columns that contained the individual words occurring in the entire collection of texts. For each word in the processed webpage text, the log of the term frequency-inverse document frequency (tf-idf) + 1 was used as weight in the DTM. This value is generally considered a good way to identify words that characterise the topics in a text (Aggarwal 2016, Gentzkow et al. 2019, Daas and Van der Doef 2020). The webpage language was included in the DTM as a binary feature (0: Spanish, 1: English) as were the occurrence of particular drone words in the text and in the URL of the webpage visited. These features are identified as Feature_language and dronF, rpasF, uasF, uavF and dronU, rpasU, uasU and uavU, respectively.

Various approaches and settings were tested, revealing that the results of the PUlearn Classic Elkanto classifier with Logistic Regression as the base model produced the most promising results. Obtaining an exact indication of the quality of the classification results was however challenging for Pulearn, given that the standard way of determining the quality of classification findings requires the availability of a confusion matrix in which the True Positives, True Negatives, False Positives and

False Negatives are available (Aggarwal 2016). As only the data of the positive cases were available, reliably calculating the scores for True Negatives and False Negatives did not seem possible. However, if one assumes the majority of the unlabelled cases to be negative, i.e. it consists of non-drone websites - not a strange assumption - the negative case based measures can be determined approximately. Based on this assumption, an overall accuracy of 86 % was obtained for a 20 % random test set derived from a combination of 1 038 positive cases and 4 000 unlabelled cases, using Logistics Regression (L1-norm) and the Classic Elkanoto classifier in PUlearn.

Interestingly, it was discovered that the overall accuracy of results could be increased by 1 % when the text of Spanish websites was translated into English, prior to modelling, bringing the model's overall accuracy to 87 %. The translation was done using Apertium (2021) software and the Spanish-English Translator installed (no free online alternative could be found, that was able to translate such large amounts of text routinely at high speed). The model developed was subsequently applied to the unlabelled dataset. A random sample of 50 items were drawn from the negative cases and 50 from the positive cases. Those samples were manually checked by experts. Findings are shown in Table 6.

From the results, it becomes clear that - after manual checking - the overall accuracy of the PUlearning approach is very close to random guessing (52 %). Apparently the features selected by the PUlearn algorithm for the positive cases do not correspond well to those used by experts. The low precision and recall displayed by the results also reveal that the model is definitely not well suited to detecting drone websites. The only positive finding is the fact that 88 % of the negative cases are being correctly classified. The latter suggests that a subset of the features typical to non-drone websites is being correctly identified by the algorithm. Seen that that is clearly not the case for the features of the positive items, work on this approach was stopped.

Table 6: Manual classification results for the trained PUlearning based model

PUlearning	Accuracy	Precision	Recall
All	0.52	0.16	0.57
Positives only	0.16	-	-
Negative only	0.88	-	-

4.1.3. SUPERVISED MACHINE LEARNING APPROACH

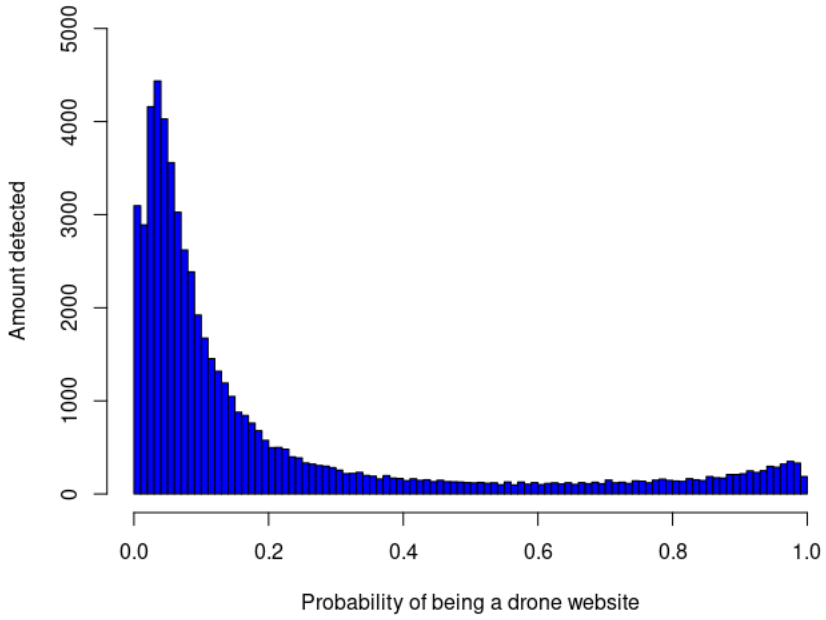
Next, it was decided to manually classify a random sample of unlabelled cases to obtain a set of negative drone websites. Since somewhat over 1 000 positive cases were already available, it was decided to randomly manually classify around 3 000 websites from the unlabelled dataset. Interestingly, after classifying 3 059 websites, a total of 2 699 non-drone and 360 drone websites had been identified. This finding suggests that around 11.8 % of the unlabelled websites found for Spain could be those of drone websites. As will be seen below, this is an overestimation. All positive and negative classified examples were processed and combined to form a DTM, as described in the previous section. Following on, the most important classification algorithms available in sklearn were tested (see Daas and Van der Doef 2020 for an overview). It was found that, to classify drone websites, a Logistic Regression with an L2-norm produced the highest accuracy on the 20 % test set ([18]). When the entire process was repeated 1 000 times, an overall accuracy of 87 % ± 1 pp was obtained for the combination of Spanish- and English-language websites. Once Spanish websites had been translated to English, and processed in an identical way, the overall accuracy increased to 88 % ± 1 pp, an increase similar to the one observed with PUlearn. The most important hyperparameter settings used in these steps were: a minimum document frequency of 100, a

([18]) Slightly lower results were obtained for a Logistic Regression with an L1-norm (acc. 86 %), Gradient Boosting (acc. 86 %) and Support Vector Machine (rbf-kernel; acc. 85 %). Of all other algorithms tested the Random Forrest classifier performed best (acc. 83 %).

maximum document frequency of 2 000, and a minimum character length of 3 for the words included.

When the (translated) words included as features in the model were inspected, however, it was found that some obviously Spanish words still occurred in the model, for example ‘dron’, ‘españa’ and ‘europa’. To ensure that all words included in the model were in English, any Spanish words observed in the model were replaced by the English translation (e.g. ‘drone’, ‘spain’ and ‘europe’) in the original texts. Next, the texts were processed and the model was retrained. These steps were repeated until Spanish words no longer featured in the model. That model was, subsequently, applied to the remainder (unseen part) of the unlabelled dataset. Since the findings yielded were the best result obtained so far, a number of additional checks were performed. The first was to classify all websites collected for Spain with the model, after the language translation and processing of the text extracted. The resulting probability distribution for those 26 067 websites is shown in the histogram in Figure 3. A total of 2 139 unique websites are classified as being drone websites (8.2 %).The large peak observed for probabilities below 0.1 indicates that a large number of obvious non-drone websites are included in the dataset. This confirms that drone websites are the minority class. A fairly broad, slowly increasing number of websites can be observed in the range of probabilities above 0.7. This may indicate that various types of drone websites do exist but they are difficult to discern. More details of the model findings are described in the following section.

Figure 3: Histogram of the model probabilities of being a drone website for Spanish websites



A random sample of 50 items was drawn from the classified websites for non-drone websites. Because a Logistic-Regression based model is capable of estimating the probability of a website being a drone website, 50 positive cases were additionally sampled in 5 particular probability ranges. 10 websites were randomly drawn in each range: between 0.5 and 0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9, and 0.9–1. The samples were combined with the sampled negative cases and manually checked by experts. The findings are shown in Tables 7 and 8. Table 7 shows the overall findings for the model and Table 8 shows the findings for the 5 probability ranges.

Table 7: Manual classification results for the trained Logistic Regression model

Log. Reg. L2	Accuracy	Precision	Recall
All	0.85	0.76	0.93
Positives only	0.76	-	-
Negatives only	0.94	-	-

Table 7 indicates an accuracy of 85 % for the model, the highest result obtained so far after manual checking. It is only slightly less than the 88 % found for the test set. The accuracy of classifying negative cases is found to be especially high, at 94 %. The model also displays a very high recall, with an indicator value of 93 %, which suggests that a high percentage of drone websites are being detected. The model developed is clearly suited to detecting Spanish drone websites.

Table 8 highlights some of the range-specific findings. The second column shows the levels of accuracy estimated for the probability range identified in first column. This part of Table 8 reveals that 90 % of the classification results of websites with probabilities of 0.7 and higher could be confirmed by manual classification through experts. This clearly confirms that the model is indeed identifying drone websites. The particular cut-off value used is shown in the third column. The fifth, sixth and seventh columns report the accuracy, precision and recall of specific probability ranges. These findings indicate that precision increases from probability 0.5 onwards, and this is associated with a decrease in recall. The best trade-off between precision and recall is observed at a cut-off value of 0.6, indicated by with a * in column four. The results at probability 0.6 and higher provide the best combination of accuracy, precision and recall for the model developed.

Table 8: Manual classification results for the trained Logistic Regression model at various probability ranges

Log. Reg. L2	Accuracy [ )	Cut-off value	Best	Accuracy* (range >=)	Precision* (range >=)	Recall* (range >=)
Prob. [0.5-0.6)	0.5	Prob. >= 0.5		0.85	0.76	0.93
Prob. [0.6-0.7)	0.6	Prob. >= 0.6	*	0.85	0.83	0.81
Prob. [0.7-0.8)	0.9	Prob. >= 0.7		0.83	0.9	0.66
Prob. [0.8-0.9)	0.9	Prob. >= 0.8		0.75	0.9	0.44
Prob. [0.9-1]	0.9	Prob. >= 0.9		0.67	0.9	0.22

All in all, it is clear that the supervised logistic regression model is able to accurately identify Spanish drone websites. The next section describes applying the model to websites of other countries.

4.2. Generalising the model

4.2.1. THE SPANISH MODEL

The model developed for Spanish drone websites was originally based on those sites' non-translated texts. However, early on, it was discovered that translating Spanish texts to English improved the model's accuracy by 1 %. This translation step is, in principle, the first step in generalisation, as it makes the model applicable to all websites written in English. Any Spanish words included in the model were removed by translating those words into their English analogue and retraining the model on the updated data. This was the second generalisation step, as it assures that Spanish words were no longer present in the model. Another important decision was made during the course of developing model: it was decided not to stem the words, i.e. map the different morphological variants

of the remaining words to their base form. This could potentially increase the model's accuracy (Daas and Van der Doef 2020), but it was not done because of the foreseen general application in mind.

It would be great for this project, if the model developed for Spain could, somehow, by applied to also detect drone websites in other countries. In order to make this possible, the features in the model were studied in more detail. When one looks at the model at this stage in the process, it contains around 1 600 features, on average. These are not only words but also include all the 9 features added to the DTM (see Section 4.1.2). The language feature has a negative weight associated to it, indicating that a website that is written in English is negatively associated with its being a drone website. Remarkably, this is also the case for the rpasF feature (occurrence of 'rpas' in the text) although the latter only bears a small negative weight. All other features added are positively associated with a drone website, as one would expect ([19]). These features are added after DTM creation, which can easily be done for a specific country. The challenge with generalising the model is that of dealing with the words included.

4.2.2. CREATING A (MORE) GENERIC MODEL

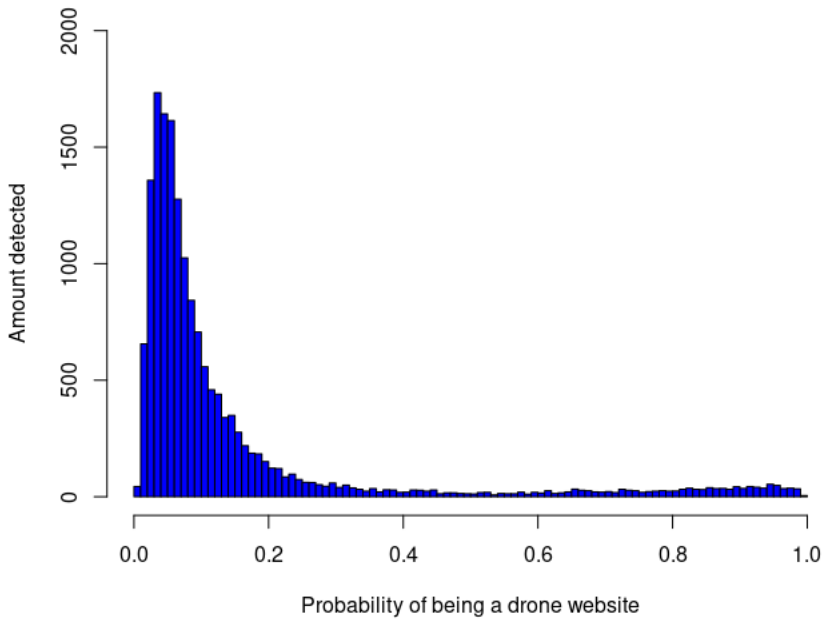
Apart from a large number of expected words, such as 'drone' and 'drones', it was found that the model also included (lowercase) words indicating a considerable number of locations and countries, such as Madrid, Barcelona, Murcia, Spain, Italy, etc. In addition, months and days of the week were also found to be included as features. The latter had also been observed by the author in previous modelling studies (Daas and Van der Doef, 2020). It was subsequently decided to remove all location-specific words from the texts (of the Spanish websites used) other than references to continents, the months of the year and days of the week. After removing these words, a new model was created and tested. It was found to be as accurate as the previous model (88 % on average, after 1 000 repeats) ([20]), while containing none of the location features, months of the year and days of the week. The final model contained 1 568 features, of which 1 559 were words. As a result of these cleaning steps, it is expected for the model to become more generically applicable, certainly to classifying English-language websites of other countries. This model's top 20 positive and negative features are included in Annex III. In order to test the model's generalisability, it was applied to the list of websites found for Ireland.

4.2.2.1. Applying the model to Irish websites

The dataset obtained for Ireland is entirely composed of websites that are written in English. The cleaned model was applied to the processed text of the websites found for this country, and the probabilities were determined (in accordance with the model). Figure 4 shows a histogram of all 14 586 classified Irish websites. A total of 785 unique websites were classified as being Drone websites (5.4 %). The histogram quite closely resembles the distribution found for the Spanish websites (Figure 4).

([19]) In Annex III, the top 20 positive and negative features included in the final version of the model are shown.
([20]) Let it be clear that the new model was found to be 0.1 % less accurate, which is a very small decrease.

Figure 4: Histogram of the model probabilities of being a drone website for Irish websites



The histogram reveals a narrow yet high peak of non-drone websites, to the left of the figure and a broad and much flatter area of the drone websites to the right.

When the occurrence of drone-specific words in the text and in the URL were determined and compared for websites with probabilities below 0.5 and equal to or above 0.5, it was found that 6 % of the first group and 99 % of the second group included one or more of these words. These findings suggest that the model is indeed differentiating Irish non-drone from Irish drone websites.

Random samples were drawn from the list of Irish websites, of 50 non-drone (probability < 0.5) and 50 drone websites (probability >= 0.5). Again, for the 50 positive cases, samples of 10 websites were drawn according to 5 particular probability ranges, from 0.5 to 0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9 and 0.9–1. The samples were manually checked by experts and the findings are shown in Tables 9 and 10.

Table 9 shows the model to have an accuracy of 86 %, which nicely confirms the observation that the model is able to correctly detect Irish drone websites. Next, the 100 % accuracy of classifying negative cases is worth mentioning. The model also has a perfect recall, a value of 100 %, which indicates that all drone websites are correctly detected. Clearly, the model developed is also suited to detecting Irish drone websites.

Table 9: Manual classification results for the classification of Irish websites

Log. Reg. L2	Accuracy	Precision	Recall
All	0.86	0.72	1.0
Positives only	0.72	-	-
Negatives only	1.0	-	-

Table 10 details the range-specific findings for the model. It reveals that 90 % of the classification results for websites with probabilities of 0.8 and higher are confirmed by the manual classification of

experts. This again clearly confirms that the model is indeed identifying drone websites. The fifth, sixth and seventh column show the accuracy, precision and recall for the data, in the different probability ranges. The cut-off value applied is shown in the third column and findings indicate that precision increases from 0.5 onwards, while recall decreases. The best trade-off between precision and recall is observed at a cut-off value of 0.6, which is indicated by with a * in the fourth column. The best combination of accuracy, precision and recall for the model developed is also found around 0.6, although the accuracy observed for cut-offs 0.5 and 0.7 is similar.

Table 10: Manual classification results for the classification of Irish websites at various probability ranges

Log. Reg. L2	Accuracy [ )	Cut-off value	Best	Accuracy* (range >=)	Precision* (range >=)	Recall* (range >=)
Prob. [0.5-0.6)	0.5	Prob. >= 0.5		0.86	0.72	1.0
Prob. [0.6-0.7)	0.6	Prob. >= 0.6	*	0.84	0.75	0.83
Prob. [0.7-0.8)	0.7	Prob. >= 0.7		0.84	0.83	0.69
Prob. [0.8-0.9)	0.9	Prob. >= 0.8		0.80	0.9	0.50
Prob. [0.9-1]	0.9	Prob. >= 0.9		0.72	0.9	0.25

4.2.2.2. Applying the model to Italian websites

The next challenging task was to attempt to apply the model, in some way or another, to the websites collected for a county, in which the native language was not English. Italian websites were the most logical choice as, for this country, apart from the data collected by the scripts, a list of identified drone websites was obtained from a collaboration with the Drone Observatory at Politecnico di Milano. The list contained 730 URLs, which included a total of 686 unique domain names. To enable the classification of Italian websites with the drone classification model developed, three steps needed to be performed. The first was creating a list of Italian words analogue to the original Spanish words that were translated into English, which were subsequently used as features in the trained model. The second step was that of translating every occurrence of the Italian words included in that list - and nothing else - in the texts extracted from the Italian websites that were to be classified. The last step was validating the findings obtained, to check whether the 'translation' approach worked.

Step 1: Creating a translation list

In this step all Spanish words that were included in the model - after translation into English - needed to be identified and subsequently to be translated into Italian. This was done by creating a list of words translated from English to Italian, by sequentially applying the following steps:

1. All features added during the model's creation, such as the language feature, and the features indicating drone acronyms in the text and in the URL, were removed from the list of features.
2. All remaining features (all English words) in the model were subsequently translated using Apertium with the English-Italian translator installed. Here, the '-u' option was not used, so any non-translated (unknown) words were preceded by an asterisk symbol (*). Apart from such words, some words were also found to be preceded by a hashtag (#), indicating a less sure translation.
3. The hashtag was removed from all translated words, implicitly accepting all 'less sure' translations.
4. All words preceded by an asterisk were (manually) translated using Google translate with English as the first and Italian as the second language. The translated words obtained were included in the list, with the following additional considerations:

a) If translating a word resulted in an Italian word with a feminine and masculine variant, both variants were included in the translation list. For example, 'one' translates into 'una' (female) and 'uno' (male).

b) The translation suggestions made by Google for the following words were ignored: 'web', 'cookie', 'cookies', and 'log'. These words were simply not translated and used as-is.

c) The translation of the words 'fly' and 'unmanned' were adjusted. For these words 'volare' and 'senza pilota' were used rather than the suggested words 'mosca' and 'senza equipaggio'.

5. All translated words were converted to lowercase and the list was sorted, in decreasing order, according to the number of spaces in the translated word(s) and the length of the translated word(s). Such ordering prevents any replacements errors when converting Italian words to English.

The list was subsequently stored as a CSV-file.

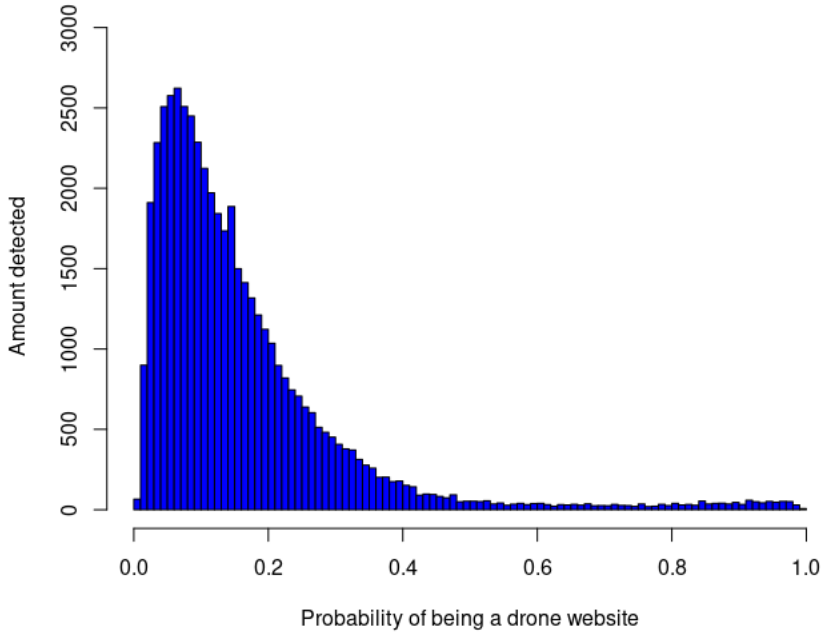
Step 2: Translating the Italian texts

The texts extracted from Italian websites were translated using the translation list created. This was done by checking the occurrence of the Italian words, in their sorted order, in the lowercase texts of the websites included, and replacing them with their English analogues. To prevent that only part of the words were translated, the occurrence of each word was checked with a leading and a lagging space added. To ensure that the first and the last word were not accidentally missed, a leading and a lagging space were added to the text prior to translating and was later removed. Because some words sometimes occur in sequence, each replacement was performed twice; this is a result of adding leading and lagging spaces. After all words in the list had been replaced in each website's text, the classification model was applied.

Step 3: Classifying findings

Figure 5 shows a histogram of all 53 781 classified Italian websites found. A total of 2 052 unique websites were classified as Drone websites (3.8 %). The histogram differs in two important aspects from the distributions shown in Figures 3 and 4, in particular in the lower probability range. The first is a much broader distribution in the non-drone range, between probability values of 0 and 0.5. This suggests a less homogeneous group of non-drone websites. The second difference is a small spike around a value of 0.15, which hints towards the relatively high occurrence of a particular group of non-drone websites. Fortunately, the profile in the range of drone websites, i.e. probability values between 0.5 and 1, very much resembles those in Figures 3 and 4, indicating that the distribution of probabilities of Italian drone websites is similar to those found in Spain and Ireland.

Figure 5: Histogram of the model probabilities of being a drone website for Italian websites



The model was subsequently applied to the 730 Italian drone websites included in the list provided by Politecnico di Milano. The websites were scraped and the text was processed and translated as described above. A total of 621 websites were classified as drone websites by the model. This corresponds to 85 % of those websites, indicating that 85 % of them are correctly classified, a very positive finding. A random sample of 50 non-drone (probability < 0.5) and 50 drone websites (probability >= 0.5) was then drawn from the list of classified Italian websites. Again, for the 50 positive cases, samples of 10 websites were drawn in 5 particular probability ranges, i.e. from 0.5 to 0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, and 0.9-1. The samples were manually checked by experts and the findings are shown in Tables 11 and 12.

Table 11: Manual classification results for the classification of Italian websites

Log. Reg. L2	Accuracy	Precision	Recall
All	0.82	0.67	0.97
Positives only	0.67	-	-
Negatives only	0.98	-	-

Table 11 shows the model's estimated accuracy to be 82 %, which additionally supports the observation that (following translation) the model is able to correctly detect Italian drone websites. In addition, the 98 % accuracy in classifying negative cases is worth mentioning. The model is clearly shown to correctly identify non-drone websites. Also worth mentioning is the very high recall, a value of 97 %, indicating that nearly all drone websites are being correctly detected by the model. The model and the translation approach developed clearly produce interesting findings.

Table 12 shows the model's range-specific findings. 90 % (or more) of the classification results of websites with probabilities of 0.8 and higher are confirmed by the manual classification of experts. This again clearly confirms that the model and translation approach are able to identify Italian drone websites. The fifth, sixth and seventh column show the accuracy, precision and recall for the data, beyond specific cut-off probabilities, which are shown in the third column. These findings indicate

that precision increases from 0.5 onwards, and that this is associated with a decrease in recall. The best trade-off between precision and recall is observed somewhere between cut-off values of 0.5 or 0.7. It is difficult to choose the optimal cut-of value. But when one compares the accuracy, precision and recall at levels 0.5, 0.6 and 0.7, the safest choice seems to be a value of 0.6. This is indicated by a * in column four.

Table 12: Manual classification results for the classification of Italian websites at various probability ranges

Log. Reg. L2	Accuracy [ )	Cut-off value	Best	Accuracy* (range >=)	Precision* (range >=)	Recall* (range >=)
Prob. [0.5-0.6)	0.4	Prob. >= 0.5		0.82	0.67	0.97
Prob. [0.6-0.7)	0.33	Prob. >= 0.6	*	0.84	0.74	0.85
Prob. [0.7-0.8)	0.7	Prob. >= 0.7		0.87	0.87	0.77
Prob. [0.8-0.9)	0.9	Prob. >= 0.8		0.83	0.95	0.56
Prob. [0.9-1]	1.0	Prob. >= 0.9		0.75	1.0	0.29

The other important check performed was to see if and how classification results are affected by any adjustments made to the translation list (described in Step 1). When adjustments were made to the translation list, with the intention of improving the translation, it was found that the classification findings of the drone websites included in the Politecnico di Milano list were negatively affected. This usually resulted in a drop in accuracy to about 75 %; no improvement was recorded. This underlines, what a critical task the creation of a well performing translation list is.

5 Data extraction

5. Data extraction

This chapter describes the scripts and results obtained from extracting specific information from the classified Drone websites.

5.1. Information extracted from potential Drone websites

It is reminded that only websites with a prediction probability of 0.6 and higher were studied. From the work described in the previous chapter, it became clear that 0.6 was the optimal cut-off point when identifying (potential) Spanish, Irish and Italian Drone company websites. The extraction of information targeted a number of variables, following the work carried out in the ESSnet Big Data 2017 research project. With the exception of the website's URL, which was already available, these variables are:

1. Name
2. Short description
3. Contact address
4. Region in country (of location)
5. Email address
6. Phone number
7. VAT number
8. Activities
9. Social media presence (multiple platforms)
10. E-commerce activity
11. Job advertisement presence
12. Start date of website.

The company or organisation, whose information is displayed on the website, is hereafter referred to as the website owner. During the course of work, it became clear that four important additional variables needed to be determined. They are the answers to the following questions:

13. Is the website really about drones?
14. Is the owner of the websites active in the country studied?
15. Is it a company website?

Although answering these questions may seem obvious, it was still essential to determine their outcomes, as not all websites found were clearly those of drone companies active in the country under study. This point had already been raised as a result of the search query applied in Chapter 3. Also, even after applying the drone classification model, websites may remain that are not about drones or that are not those of companies. In the end, it was decided that information needed to be extracted from the webpage to obtain a total of 16 variables. The procedure's implementation is described below.

The general approach was to always start with the following steps:

- First, the content of the main page returned by the URL provided was scraped and the HTML-code obtained was parsed using Beautiful soup. The latter seriously reduces any errors in badly written HTML-pages.
- When no HTML-code was obtained, it was checked, whether the URL re-directed to another domain and a subsequent attempt was made to scrape that domain. This approach resulted in scraping 98 % of all the URLs provided (98 % for Spain, 97 % for Ireland and 98 % for Italy).
- Next, any links within the same domain name referring to contact pages, e.g. links containing words such as 'contact', 'about us' and 'impressum' (in English or in the country's native language) were searched for. If found, those pages were also scraped and parsed using Beautiful soup.
- The combinations of these 'soups' and that of the main page were made the starting input for each of the functions described below.

The most important reason for following this approach was to reduce any repeated scraping of the same pages within each domain. Because many specific regular expressions were used and many specific decisions needed to be made to extract the required information from web pages, it was decided to create a specific script for each country. Individual scripts were therefore created for Spain, Ireland and Italy.

5.1.1. NAME OF THE WEBSITE OWNER

The name of the website owner was obtained from the main page. This step first looked for a meta-tag in the HTML-code with the 'og:title' attribute. Although this tag is not visible on the page shown in a browser, many webpages contain it as many search engines used it. If this tag was found the content was extracted. If the tag was not found, the title of the main page was used. If the title was absent or empty, the content of the first header (<h1>) was used. A total of 98 % of the websites provided this information (98 % for Spain, 96 % for Ireland and 98 % for Italy).

5.1.2. SHORT DESCRIPTION OF THE WEBSITE

The descriptive text for the website was obtained from the main page. This step first looked for the occurrence of meta-tags in the HTML-code. Those tags were then checked for the inclusion of the word 'description' and 'content'. If they were found, the content was extracted. If they were not found, the top 10 words in the visible text of the main page were determined, a score was determined for the occurrence of the top 10 words in all sentences on the main page and it was calculated, finally the sentence with the highest score was subsequently selected. This information was obtained from a total of 98 % of the websites (98 % for Spain, 97 % for Ireland and 98 % for Italy).

5.1.3. CONTACT ADDRESS

The website owner's contact address was searched for in the HTML-code of the main- and contact-pages. Address information was found to be one of the most challenging variables to collect. In principle, the approach implemented is the extension of an approach originally developed to extract addresses from Dutch webpages.

First, a search was performed that focussed on detecting the occurrence of a postal code specific to the country studied; this was done using a regular expression. If a postal code was found, a subsequent search was carried out for the beginning and the end of the, usually, surrounding address part. For Spain and Italy, the search focussed on finding a city name within a maximum of 100 characters after the postal code location. This required the availability of a list of cities for each country. For Ireland, this approach was adjusted, given that the postal code was found to usually indicate the end of the address section and the city name could often be found before the postal code. Next, for each country, the beginning of the address was searched for within a maximum of 100 characters, in the text preceding the postal code. For this purpose, a list of most frequently occurring parts that are common to street names, such as 'street', 'park', 'road', 'lane', etc. was created in the language of each country. If no common street name part was found, the focus was on finding the beginning of the address section by looking for specific structural information, such as multiple spaces, soft returns, dots, etc.

When no postal code was found, the focus was on finding combinations in sections of text containing common parts of street names and city names located less than 100 characters from one another. Essentially, this was done as described above, except that the search started with finding city names. As an additional check, the text extracted should include a string of characters (such as 's/n' for Spain) that is indicative of an address.

When the two approaches described above did not provide any results, the focus was put on finding a combination of city names and the country studied, or on a city name, or the name of the country alone. When multiple results were found, the city name mentioned most was used as an indication of the potential physical location of the website owner. For Irish websites, specific attention needed to be paid to exclude addresses located in Northern Ireland as this is part of the United Kingdom.

The address search took a considerable amount of time for some webpages, in particular when multiple (potential) addresses were found. Because of this, the function was limited to the 10 most frequently occurring postal codes (if they occurred) and/or to the 5 most closely located combinations of city and street names in the text. The address information was obtained for a total of 55 % of the websites (58 % for Spain, 32 % for Ireland and 61 % for Italy).

5.1.4. REGION OF LOCATION

When an address or multiple addresses were found, the postal code or city name were used to find the corresponding region in the country studied. The region could be at Nuts level 2, level 3 or, where available and different, the county or province. This required the availability of a list of city names, their corresponding Nuts2 and Nuts3 region, county or province names, and their relation to (part of) the postal code for each country. When a postal code was included in the address obtained, that code was preferred for identifying the region. The city name extracted was used as the next alternative. If no clear address information was obtained, the HMTL-code of the main and location pages were studied for the occurrence of Nuts2, Nuts3, county or province names, in those pages' visible text. This information was obtained for a total of 46 % of the websites (57 % for Spain, 16 % for Ireland and 47 % for Italy).

5.1.5. EMAIL ADDRESS

Email addresses were searched for by using a generic regular expression in the HTML-code of the main and contact pages. First, strings corresponding to email addresses were searched for in the visible text. Strings ending in '.jpg', '.jpeg', '.png', and '.webp' were excluded. If nothing was found, the HTML-code was searched, to detect any email-addresses included in the code only. This information was obtained for a total of 60 % of the websites (64 % for Spain, 53 % for Ireland and 60 % for Italy).

5.1.6. TELEPHONE NUMBERS

Telephone numbers were searched for with a combination of country-specific regular expressions, in the HTML-code of the main and contact pages. First, the search covered the visible text, with a regular expression specific for the structure of landline numbers for the country studied. Search parameters both included and excluded the country code, for example '+34' and '0034' for Spain. Next, a regular expression specific to mobile phone numbers in the country was used, both including and excluding the country code. Subsequently, a generic regular expression, with and without country codes, was used to detect telephone numbers for other countries. The latter expression – including the country code - was especially important in determining whether a website owner did not belong to the country studied. In addition, a specific search was performed to detect numbers indicated by Whatsapp. The information on telephone numbers was obtained for a total of 72 % of the websites (74 % for Spain, 59 % for Ireland and 73 % for Italy).

5.1.7. VAT NUMBER

VAT numbers were searched for with country specific regular expressions, in the visible text of the main and contact pages. A specific regular expression was used for each country. The regular expression was sufficient to detect Spanish VAT codes. However, for Irish and Italian VAT codes an additional search was included that checked the surrounding text for the acronyms 'VAT' and 'CBL' (for Ireland) or 'IVA' and 'VAT' (for Italy). Italian codes could also be preceded by 'IT'. VAT numbers are the variable that is least prevalent on the websites of all countries studied. This information was obtained for a total of 12 % of the websites (3 % for Spain, 0.8 % for Ireland and 27 % for Italy).

5.1.8. ACTIVITIES REPORTED

An attempt was made to identify a number of activities from the websites' text. The activities sought were: manufacturing, distributing, maintaining, sales, renting, training, filming/imaging, inspecting, components produced, consultancy, entertainment/race, and design. A limited number of words in English and in the native language of the country studied were used to detect each of these activities. It was first attempted to identify the activities mentioned in the website's descriptive text (see Section 5.1.2). If any activities are mentioned in that text, they often indicate the website owner's main activity. For a sample of websites, comparing findings to those reported by experts confirmed that observation.

If no activities could be found in the descriptive text, the website's main page was searched for links to pages that might provide an overview of the company's activities. English and native-language variants of the words 'activities' and 'services' were used to identify such links. Any links found were scraped and added to the set of pages already collected, e.g. the main and contact pages. Those HTML-pages were subsequently searched for the occurrence of country-specific drone synonyms, such as 'drone', 'rpas', 'uav', 'uas' and 'unmanned', and that of acronyms of the country under study's main drone organisation, i.e. 'AESA' for Spain, 'UAAI' for Ireland and 'ENAV' for Italy ([21]). When one of the drone words was found, the surrounding text - 20 words to the left and 20 words to the right - were searched for the occurrence of one of the words specific to a particular activity. The combined set of activities is reported for all pages studied. This method usually yielded a much higher number of activities than those obtained on the basis of the descriptive text alone, and also than those reported by experts. Reducing the number of words searched to the left and to the right had the disadvantage that important activities were missed. This information was obtained for a total of 77 % of the websites (81 % for Spain, 71 % for Ireland and 74 % for Italy).

([21]) Finding these words also helps with answering the question 'Is the website really is about Drones?' For further details, see Section 5.1.12.

5.1.9. SOCIAL MEDIA PRESENCE

The information on the website owner's social media accounts was obtained from social media links found in the HTML-code of the website's main and contact pages. Specifically, links indicative of Twitter, LinkedIn, Facebook, Instagram, Pinterest, YouTube, Vimeo, and GooglePlus accounts were searched for. Each link was checked to make sure it referred to an actual account, rather than to a specific message/post or to another activity related to one of those platforms. This is important as many websites contained links to messages or to videos posted on social media. This information was obtained for a total of 64 % of the websites (64 % for Spain, 59 % for Ireland and 67 % for Italy).

5.1.10. E-COMMERCE ACTIVITY

The detection of e-commerce activities on each web site was determined by looking for the occurrence of specific words in the visible text of the main and contact pages. The words used were the English and country-specific variants of the set of words applied for the same task in the ESSnet Big Data (2017). The words were checked by Statistics Netherlands staff currently involved in these studies. In principle, the words used are all variants of 'shopping cart', 'shop' and 'customer'. For each country, the English variants and their translation into a country's native language were included. If nothing could be found, the HTML-code of the main and contact pages was additionally checked. If any words were found, e-commerce activity was identified as being 'True'. If not, it was identified as being 'False'. This information was obtained for a total of 98 % of the websites (same percentage for each of the countries studied).

5.1.11. JOB ADVERTISEMENT PRESENCE

The detection of job advertisements on each web site was determined by looking for the occurrence of specific words in the visible text of the main and contact pages. The words used were the English and country-specific variants of the words 'job', 'vacancy' and 'vacancies'. For each country, the English words and their translations into the country's native language were included. If any words were found, the presence of job vacancies was identified as being 'True'. If not, it was identified as being 'False'. This information was obtained for a total of 98 % of the websites (same percentage for each of the countries studied).

5.1.12. WEBSITE START DATE

This variable was added after all information had been collected. An additional script was therefore run for the websites identified as being drone companies in a particular country (see Section 5.1.16 below). In determining a company's start date, the start data of the website's creation were taken as a proxy. This was obtained by consulting the who.is website (https://who.is) with the website's domain name as the input. The creation or registration date was available for a large part of the websites. For Spain, in 238 of the 461 Drone websites identified (52 %) the date was before February 2022, the period when the data was collected. February 2022 was ignored because it indicated the date when information was last updated. The average date was 2014-05-13 for the Spanish websites. For Ireland, a creation date was available for 63 of the 66 Drone websites (95 %). The average date was 2014-12-23 for the Irish websites. For Italy, 325 of the 353 Drone websites identified (92 %) a creation date was available. The average date was 2013-01-11 for websites in Italy. A creation date was available in 74 % of the websites studied, in the three countries together.

5.1.13. IS THE WEBSITE REALLY ABOUT DRONES?

Apart from the above-mentioned variables, it was found necessary to collect information on three further variables. These are all required because one needs absolute certainty that a website: i) is about drones, ii) is active in the country studied and iii) is owned by a company. The first question is 'Is the website really about drones?'. This appears to be a trivial question to answer, given that

websites had already been classified by a drone-specific model and had been found to have a probability of 0.6 or higher. However, one must expect that some of the websites in fact correspond to so-called false-positives (see Chapter 4), i.e. non-drone websites containing words, to which high weights are assigned in the model. One can imagine that, for instance, a website about aviation scores relatively highly. Another exception would be that a non-drone website of a particular type is found that has not been included in the training set. That website could still be scoring highly, without it actually being a drone website. All in all, there is a need to make sure the websites studied are about Drones and not about other topics. For this purpose, a first check was performed of the occurrence of drone synonyms and words related to drones. This is further described in Section 5.1.8. In fact the check is performed while attempting to identify the activities the website reports on. If drone synonyms are found, the variable is assigned the value True. In all other cases, it is False. This information was obtained for all websites. Of all websites scraped, 85 % concerned drones (86 % for Spain, 87 % for Ireland and 83 % for Italy).

5.1.14. IS THE WEBSITE'S OWNER ACTIVE IN THE COUNTRY STUDIED?

The next question was to determine, whether the website is actually active in the country under study. This was necessary as, toward project beginning, it was observed that many search results may also include websites that are not located in the country studied (see Chapter 3). An attempt was made to exclude such websites, as early in the process as possible, but one may expect that a number of them remained, even after classification (see Chapter 4). Telephone numbers, the address, the HTML-codes of the main and contact-pages, and the URL were subsequently used to derive this essential information.

In order to determine the country, in which the website was active, the extracted telephone numbers found were first checked for the occurrence of the country code of the country studied, i.e. '+34' and '0034' for Spain, '+353' and '00353' for Ireland and '+39' and '0039' for Italy, respectively. If other country codes were found, which should begin with a '+' or '00', the website was clearly active in another country. If no decision could be made, the address(es) extracted were checked for the occurrence of a city's name, in the country studied, or of the name of the country studied. Is subsequently nothing was found, the visible text in the HTML-code of the main and contact pages were checked for the occurrence of a city or the name of the country. As a last resort, if still nothing could be found, the domain names' country-specific extension was looked for: '.es' for Spain, '.ie' for Ireland and '.it' for Italy. If the resulting location could be assigned to the country studied, the value of the variable became 'True'. In all other cases it was 'False'. This information was obtained for all websites. Of all the websites scraped, 53 % were located in the country studied (62 % for Spain, 35 % for Ireland and 50 % for Italy).

5.1.15. IS IT A COMPANY WEBSITE?

The next question requiring an answer was, whether the website is owned by a company. One must be aware that not all websites found are those of companies. News websites and blogs were certainly also included. This question was answered by looking at the VAT, URL, descriptive text, and the combination of address, telephone number and either the descriptive text or the main and contact pages' HTML-code. If a VAT number of the country studied was found, the website had to be of a company. If the URL or the descriptive text contained the word 'blog' was included, the website was certainly not that of a company. In all other undecided cases, the combination of finding an address and phone number, together with the occurrence of company words (either in English or in the native language, e.g. 'company' or 'business') in i) either the descriptive text or in ii) the visible text of the main or contact pages, it was considered to be indicative of a company. However, if the visible text contained many mentions to company words, yet included the word 'news' or 'blog', it was not considered to be a company website. The step was necessary to exclude a number of news and blog websites containing much information on different companies. This information was obtained for all websites. Of all websites scraped, 35 % were identified as being those of companies (37 % for Spain, 17 % for Ireland and 41 % for Italy).

5.1.16. IS IT THE WEBSITE OF A DRONE COMPANY ACTIVE IN THE COUNTRY?

The fact that identified websites are actually those of Drone companies active in the country studied, i.e. the combination of the Booleans described in Sections 5.1.12, 5.1.13 and 5.1.14, was subsequently considered. This information was obtained for all websites. Of all websites scraped, 22 % were identified as being the websites of companies that are active in the country (26 % for Spain, 10 % for Ireland and 23 % for Italy). This resulted in a total of 461 Spanish drone company websites, 66 Irish drone company websites and 353 Italian drone company websites.

5.1.17. COMPARISON OF FINDINGS

The URLs of the 461 and 353 companies found, respectively, for Spain and Italy were compared with those included in the lists of known Drone companies for Spain and Italy (see Chapter 3 and 4). This confirmed the finding of 312 new URLs for Spain and 267 new URLs for Italy. This indicates that a considerable number of new drone companies could be detected in both countries, a very successful result. The reader should nevertheless be aware that each URL needs to be manually checked, to make absolutely sure that it is indeed the website of a Drone company. The new URLs found for Spain are listed in Annex IV, those found for Italy in Annex V and those found for Ireland are included in Annex VI.

As a last checking step, samples of a 100 Spanish and 100 Italian Drone websites were drawn and findings were manually checked and compared. A total of 66 websites were studied, in the same way, for Ireland. This revealed that, of the URLs found, 85 %, 85 % and 78 % respectively were correctly assigned to drone companies active in the country, for Spain, Ireland and Italy. The accuracy of the individual variables obtained, when researching, whether the website was about drones, whether the website was that of a company and whether the website was active in the country, were considerably higher than the aggregated result, with values of 87 %, 86 % and 82 % displayed by Spain, Ireland and Italy, respectively. Results indicate that the combination of the three variables used to determine, whether a website is of a Drone company active in the country, should be improved somewhat, especially for Italy.

Table 13: Percentage of results obtained for the variables extracted for Spain, Ireland, Italy and countries combined

	Spain	Ireland	Italy	Country results combined
1. Name	98	96	98	98
2. Short description	98	97	98	98
3. Contact address	58	32	61	55
4. Region in country	57	16	47	46
5. Email address	64	53	60	60
6. Phone number	74	59	73	72
7. VAT number	3	0.8	27	12
8. Activities	81	71	74	77
9. Social media presence	64	59	67	64
10. E-commerce activity	98	98	98	98
11. Job advertisement	98	98	98	98
12. Start date of website*	52	95	92	74
13. About drones	86	87	83	85
14. Active in the country	62	35	50	53
15. Company website	37	17	41	35
16. Drone company in country	26	10	23	22

*Values only determined for the websites identified as drone companies active in the country studied

5.2. Results

The methodology developed ultimately targeted the collection of data for the assessment of the civil drone sector, by using the internet as a source of data for the automatic retrieval of information. The scripts and results obtained from extracting specific information from the classified drone websites have been described in this chapter. In particular, the results reported in this section aim to inform on the type of activity drone companies are performing in the supply chain (see Section 5.1.8) and the location of those companies.

An overview of the main results of the web search strategy and the data extracted for drone companies in Italy, Spain and Ireland on the activities reported is provided in Table 14.

Table 14: Results of the web search strategy and data extractions for Italy, Spain and Ireland

	Italy	Ireland	Spain
Total number of URLs retrieved	53 785	14 568	26 067
Websites classified as Drone companies (class. prob ≥ 0.6)	1 608	681	1 874
Drone company websites selected (of which new drone company websites found)*	353 (267)	66 (n/a)	461 (312)
Drone company websites for which 'Type of activity' could be extracted	287	56	388

**see Section 5.1.17*

Company activity could be ascertained for 388 of the 461 companies detected in Spain (84.2 %), 287 out of 353 in Italy (81.3 %) and 56 out of 66 in Ireland (84.8 %). It is important to mention that specific activities were searched for, which might limit the types of activities found. Moreover, results depend on an accuracy rate that is not 100 % (see Section 5.1.17). Besides, the results obtained refer to the set of companies identified by the methodology developed and not to the entire sector.

Analysis of the activities reported by drone companies on their websites allows for their classification according to their position in the industry value-chain. As drone sector companies may simultaneously perform different activities, analysing the co-occurrence of different activities informs on the 'service packages' drone companies offer, in the three countries under study. A high co-occurrence of certain services, such as inspection, filming and surveying, can clearly be seen in the following three tables. For example, Table 15 would indicate that 6 companies in Spain focused on distribution (i.e. distributing and sales) while also offering training.

Table 15: Drone company activities in the value chain and their co-occurrence in Spain

	COMPONENTS	CONSULTANCY	DESIGN	DISTRIBUTION	ENTERTAINMENT_RACE	FILMING_IMAGING	INSPECTION	MAINTENANCE	MANUFACTURING	RENTING	SALES	TRAINING
COMPONENTS	64	2	3	3	2	39	9	9	8	0	10	21
CONSULTANCY	2	9	1	2	0	3	1	3	2	0	1	2
DESIGN	3	1	16	2	0	4	4	2	4	0	1	4
DISTRIBUTION	3	2	2	20	1	5	1	1	1	0	8	6
ENTERTAINMENT_RACE	2	0	0	1	10	5	1	4	1	1	2	4
FILMING_IMAGING	39	3	4	5	5	227	28	19	17	11	12	43
INSPECTION	9	1	4	1	1	28	48	7	11	1	1	14
MAINTENANCE	9	3	2	1	4	19	7	30	7	2	9	15
MANUFACTURING	8	2	4	1	1	17	11	7	42	1	4	12
RENTING	0	0	0	0	1	11	1	2	1	16	0	3
SALES	10	1	1	8	2	12	1	9	4	0	39	19
TRAINING	21	2	4	6	4	43	14	15	12	3	19	124

Table 16: Drone company activities in the value chain and their co-occurrence in Italy

	COMPONENTS	CONSULTANCY	DESIGN	DISTRIBUTION	ENTERTAINMENT_RACE	FILMING_IMAGING	INSPECTION	MAINTENANCE	MANUFACTURING	RENTING	SALES	TRAINING
COMPONENTS	47	3	3	5	0	19	8	2	2	2	16	9
CONSULTANCY	3	23	0	2	0	8	8	2	1	2	3	10
DESIGN	3	0	17	0	1	4	5	0	3	0	2	1
DISTRIBUTION	5	2	0	19	0	7	6	0	1	1	4	1
ENTERTAINMENT_RACE	0	0	1	0	6	3	0	0	0	0	0	1
FILMING_IMAGING	19	8	4	7	3	152	48	3	20	8	20	28
INSPECTION	8	8	5	6	0	48	82	2	8	4	4	19
MAINTENANCE	2	2	0	0	0	3	2	10	1	1	3	3
MANUFACTURING	2	1	3	1	0	20	8	1	29	2	2	5
RENTING	2	2	0	1	0	8	4	1	2	16	3	4
SALES	16	3	2	4	0	20	4	3	2	3	46	10
TRAINING	9	10	1	1	1	28	19	3	5	4	10	69

Table 17: Drone company activities in the value chain and their co-occurrence in Ireland

	CLOUD	CONSULTANCY	DISTRIBUTION	FILMING_IMAGING	INSPECTION	INSURANCE	MAPPING	MONITORING	RENTING	SOFTWARE	SURVEYING	TRAINING
CLOUD	1	1	0	0	1	0	0	0	0	0	1	0
CONSULTANCY	1	4	0	1	3	0	2	1	0	0	3	1
DISTRIBUTION	0	0	4	0	0	0	0	0	0	0	0	0
FILMING_IMAGING	0	1	0	35	15	0	11	1	0	0	15	0
INSPECTION	1	3	0	15	22	0	11	2	0	1	16	0
INSURANCE	0	0	0	0	0	1	0	0	0	0	0	0
MAPPING	0	2	0	11	11	0	15	2	0	0	11	0
MONITORING	0	1	0	1	2	0	2	2	0	0	2	0
RENTING	0	0	0	0	0	0	0	0	1	0	0	0
SOFTWARE	0	0	0	0	1	0	0	0	0	1	0	0
SURVEYING	1	3	0	15	16	0	11	2	0	0	25	0
TRAINING	0	1	0	0	0	0	0	0	0	0	0	3

A company's location can be identified, based on two of the variables extracted through the method developed (i.e. contact address and region). Contact address offers information for companies' location at postal-code level, while region provides information at Nuts3 level. Codes ([22]) were created to make a visualisation of the location of companies through maps. The maps are displayed and commented in the three following illustrations.

([22]) The codes created for the maps can be accessed here: https://github.com/eurostat/wih_drones_companies/blob/master/map_ES_IT_IE.R

Figure 6: Location of drone companies in Spain

In Spain, a high number of companies are located in Madrid, followed by Barcelona. The reasons for the location of drone companies in the country could be explained by the population of the respective provinces, but also by the existence of universities that offer engineering studies. These universities might incentivise new companies based on technologies. The visual representation of the location of drone companies in Spain clearly shows that they are dispersed among all regions of the country. A considerable share are however concentrated in provinces with higher overall populations of companies, such as Madrid and Barcelona. Analysing company locations together with the activities they specialise in, as well as the sectors they provide their services to, could provide further insight into how location is selected, together with factors related to innovation.

Figure 7: Location of drone companies in Italy

In Italy, three main poles are clearly identifiable, in Rome, Milan and Turin. A broad distribution across all regions also characterises the drone sector in this country, in a way similar to Spain.

Figure 8: Location of Drone companies in Ireland

In Ireland, drone companies are predominantly located in Dublin. Although the sector displays less presence in the rest of the country, drone companies are active in each of the country's regions.

6 Generalising the methodology and tools

6. Generalising the methodology and tools

6.1. Main considerations

Wherever possible, the methodology and tools developed in the context of this project took a two-folded generalisation perspective into consideration: across countries and across economic sectors. The study's final results proved that the method developed can be generalised to an important degree.

More in detail, Scripts 1 to 5 (see Chapter 3) can be applied to other Member States by making certain minor updates (i.e. key terms in the national language used for the search queries, e.g. 'drone' in English, 'dron' in Spanish and 'Drohne' in German). The classification model (Script 6) is generalisable across countries for English-language websites and for websites that can accurately be translated into English, whenever the features used on the drone websites of the country studied are comparable to those used in Spain, Ireland and Italy. The data-extraction script (Script 7) is country-specific, given the need to adjust it to the specific features, such as VAT and address details for example, for the correct identification and extraction of information.

In terms of generalisability across countries, the Drone sector has an advantage, the likeness of terms used across countries. The domain-specific key terms used for the search queries are very much linked to standardised English terminology (e.g. RPAS, UVS, UAV).The disadvantage of using these common terms (i.e. English terminology) is that the search explores all results. The exploratory analyses performed (see Chapter 1) were essential to defining the keywords to be used for the web-search strategy, and as well for the drone company classification model and the data extraction (see Chapters 2, 4 and 5). It should further be noted that, particularly in relation to the selection of terms for the web search strategy, search engines do not perform well if too many different words are used in a query. The selection of terms was therefore based on expert exploratory analyses performed on a small sample of representative company websites, considered as being the best starting approach.

When considering generalisation across economic sectors, the country-specific scripts would be suitable for extension to other emerging economies, subject to the adaptation of key specific terminology (i.e. key terms in the national language used for the search queries, e.g. 'drone' in English, 'dron' in Spanish, 'Drohne' in German) as well as to the extraction of information on the activities of different sectors (Script 7).

From the economic sector generalisation perspective, the identification of sector-related key terms could, in part, be tackled through using Wikipedia as a source of information on the specific domain (see Section 6.2). In addition to the key word-extraction based approach, an expert check of exploratory sectoral analysis is however recommended, for an accurate selection of key terms

relevant to other sectors.

6.2. Generic keyword extraction

This section describes a generic method for the extraction of keywords from a given document about a given topic. This approach can be applied to two tasks, in the context of finding companies that are active in a specific industry (such as the drone industry):

1. to finding additional keywords for the composition search queries. This can be done by extracting keywords from a document that describes the target industry. For the drone industry we use the Wikipedia article Unmanned Aerial Vehicle: https://en.wikipedia.org/wiki/Unmanned_aerial_vehicle. Domain experts usually know, which keywords to use, but keywords may also be found automatically, that the experts did not consider. For example, when finding keywords for the topic football, an expert may think of the following keywords: 'football', 'goal', 'field', 'player', 'referee', and 'penalty'. What the expert may not realise, however, is that most of these keywords are also general purpose words. He or she may also forget other keywords that are more specific to football, such as 'offside' or 'stadium'. Keywords found automatically should not be used as a full replacement of the keywords the expert proposes, but as a potential addition.
2. to validating, whether the companies found are indeed active in the target industry. This can be done by extracting the keywords on the company's home page and by comparing them with the list of keywords mentioned in the previous point.

In the remainder of this section, we will focus on the first application.

The document from which the keywords are being extracted is called the source document. In order to determine, whether a word in the source document is specific (and hence can potentially be used as a keyword), word occurrences of a language in general are retrieved from a language corpus, a large collection of text scraped from many sources. We use the Exquisite Corpus (https://github.com/LuminosoInsight/exquisite-corpus), which itself uses 8 different domains of text, including the entire Wikipedia, for 44 languages.

6.2.1. METHOD

A few preliminaries are needed, before describing the method step by step. Let Wx to be the set of unique words contained in document x. Let n_w^x be the occurrence of each word $w \in W_x$ in document x and let

$$N_x = \sum_{w \in W_x} n_w^x$$

be the total number words in the document. The relative frequency for each word $w \in W_x$ in document *x* is calculated as

$$f_w^x = \frac{n_w^x}{N_x}$$

The method consist of the following steps:

1. All words in the source document (which we notate as *d*) are collected, but stop-words such as pronouns, prepositions, and common verbs are excluded. The remaining set of words is notated as W_d and, using the formulae above, the relative word frequency f_w^d can be calculated for each word $w \in W_d$.
2. We notate the language corpus as *q.* Let the set of words in this language corpus be W_q and the relative word frequency f_w^q for each word $w \in W_q$ be defined as above. For the method

that we describe, we only need to retrieve the relative word frequencies of the words in W_d, i.e. the words in the source document *d*.

3. All words in the source document are stemmed. In text mining, stemming is reducing each word to its stem, base or root-form. For instance, the words 'use', 'used', and 'using' have the same stem, namely 'use', and are therefore reduced to this stem word.
4. The words W_d are grouped per stem word. Let S_d be the set of stem words retrieved from document *d*, where $R_d(s) \subseteq W_d$ is the set of words that are reduced to stem word $s \in S_d$.
5. For each stem word $s \in S_d$, its relative frequency is calculated as the sum of relative frequencies of its original 'unstemmed' words in source document *d*:
6. $f_s^d = \sum_{w \in R_d(s)} f_w^d$
7. For each stem word $s \in S_d$, its relative frequency in the language corpus *q* is calculated in the same way:

$$f_s^q = \sum_{w \in R_d(s)} f_w^q$$

Note that we only use the relative frequencies of the unstemmed words that are found in source document *d*, and not all possible unstemmed words that can be found in the language corpus. So in the example of the stem word 'use', suppose only the words 'use', 'used', and 'using' occur in document *d*. Then only the relative frequencies of those words are counted and not those of other words that have 'use' as stem word, such as 'users'.

8. Finally, per stem word $s \in S_d$, the ratio between the relative frequency in the source document and the relative frequency in the language corpus is calculated:

$$r_s = \frac{f_s^d}{f_s^q}$$

For details on the implementation of the method, please see Annex VII.

6.2.2. RESULTS

The method was applied to Wikipedia articles describing three emerging economic sectors: drones (UAV), the circular economy, and renewable energy.

Drone (Unmanned Aerial Vehicle)

The results from the Wikipedia article about drones (Unmanned Aerial Vehicle) https://en.wikipedia.org/wiki/Unmanned_aerial_vehicle are shown in Table 14.

The top-ranked word ‘uncrewed’ is rarely used (Microsoft Word tried to auto-correct it to ‘unscrewed’), but very specific to drones.

As could be expected, the topics themselves (‘drone’ and ‘uav’) are included in the list. Interestingly, the more technical term ‘uav’ occurs more frequently in the source document and less in the language corpus than ‘drone’ does.

The list of keywords may also include the acronyms of companies or institutes. In this example, DJI is a Chinese drone manufacturer and FAA (Federal Aviation Administration) is a US-based agency.

Table 18: Output table for the Wikipedia page ‘Unmanned Aerial Vehicle’. Only the top 20 rows are shown.

stem	word(s)	count	freqSrcDoc	freqExCorp	freqRatio
uncrew	['uncrewed']	8	0.002537266	2.2e-08	115 856.9
uav	['uav', 'uavs']	127	0.040279099	1.667e-06	24 162.63
dji	['dji']	7	0.002220108	2.69e-07	8 253.19
unman	['unmanned']	17	0.00539169	1.86e-06	2 898.76
payload	['payload', 'payloads']	11	0.003488741	2.187e-06	1 595.22
drone	['drone', 'drones']	53	0.016809388	1.155e-05	1 455.36
faa	['faa']	7	0.002220108	3.24e-06	685.22
aerial	['aerial']	15	0.004757374	7.76e-06	613.06
autonom	['autonomous', 'autonomously']	11	0.003488741	8.577e-06	406.76
endur	['endurance']	7	0.002220108	6.03e-06	368.18
autonomi	['autonomy']	9	0.002854424	8.51e-06	335.42
balloon	['balloon', 'balloons']	7	0.002220108	1.176e-05	188.78
sensor	['sensor', 'sensors']	9	0.002854424	1.591e-05	179.41
aircraft	['aircraft']	38	0.012052014	7.24e-05	166.46
aviat	['aviation']	10	0.003171583	1.95e-05	162.65
remot	['remote', 'remotely']	15	0.004757374	3.218e-05	147.84
configur	['configuration', 'configurations', 'configurable']	6	0.00190295	1.3735e-05	138.55
pilot	['pilot', 'pilots', 'piloted']	20	0.006343165	5.493e-05	115.48
surveil	['surveillance']	7	0.002220108	2.04e-05	108.83
flight	['flight', 'flights']	23	0.00729464	0.000108	67.54

Circular economy

The results from the Wikipedia article about the circular economy https://en.wikipedia.org/wiki/Circular_economy are shown in Table 15.

Walter Stahel is an architect, who introduced the principles of the circular economy, and the Ellen MacArthur Foundation (EMF) is a popular foundation that supports circular-economy initiatives. A disadvantage of the method is that it is unable to separate the first and the last name; both 'ellen' and 'macarthur' occur in the list. It would be interesting to know, how useful these names are as keywords in finding companies that apply circular economy principles.

Table 19: Output table for the Wikipedia page 'Circular economy'. Only the top 20 rows are shown.

stem	word(s)	count	freqSrcDoc	freqExCorp	freqRatio
stahel	['stahel']	8	0.001360313	1.8e-08	76 422.07
circular	['circular', 'circularity']	236	0.04012923	1.0091e-05	3976.66
regen	['regenerative']	6	0.001020235	1.12e-06	910.92
reus	['reuse', 'reusing']	13	0.002210508	2.557e-06	864.49
cradl	['cradle']	15	0.002550587	3.39e-06	752.39
macarthur	['macarthur']	7	0.001190274	1.82e-06	654.0
entropi	['entropy']	7	0.001190274	1.82e-06	654.0
ce	['ce']	32	0.005441251	9.55e-06	569.76
landfil	['landfill', 'landfills']	8	0.001360313	2.447e-06	555.91
recycl	['recycling', 'recycled', 'recycle', 'recyclable']	28	0.004761095	1.1215e-05	424.53
economi	['economy', 'economies']	211	0.035878252	9.83e-05	364.99
renew	['renewable']	13	0.002210508	7.76e-06	284.86
linear	['linear']	16	0.002720626	1.17e-05	232.53
wast	['waste', 'wastes']	67	0.01139262	6.192e-05	183.99
framework	['framework', 'frameworks']	22	0.00374086	2.294e-05	163.07
sustain	['sustained', 'sustainable', 'sustainability', 'sustainably']	34	0.00578133	3.7307e-05	154.97
automot	['automotive']	6	0.001020235	7.24e-06	140.92
logist	['logistics', 'logistic']	6	0.001020235	7.54e-06	135.31
ellen	['ellen']	7	0.001190274	1.05e-05	113.36
furnitur	['furniture']	12	0.002040469	1.86e-05	109.7

Renewable energy

The results from the Wikipedia article about renewable energy https://en.wikipedia.org/wiki/Renewable_energy are shown in Table 16.

As this topic is much more technical than the circular economy, many technical terms or abbreviations appear in the list. For instance, the top-ranked keyword EGS stands for Enhanced Geothermal Systems. In finding companies active in the renewable energy sector, such words could truly serve as keywords, since those companies are most likely also technical companies.

Table 20: Output table for the Wikipedia page 'Renewable energy'. Only the top 20 rows are shown.

stem	word(s)	count	freqSrcDoc	freqExCorp	freqRatio
eg	['egs']	6	0.001006543	7.8e-08	12 970.91
geotherm	['geothermal']	32	0.005368227	1.32e-06	4 066.84
photovolta	['photovoltaic', 'photovoltaics']	26	0.004361684	1.105e-06	3 947.23
Renew	['renewable', 'renewables']	142	0.023821506	8.88e-06	2 682.6
Biofuel	['biofuels', 'biofuel']	20	0.003355142	1.251e-06	2 681.97
biomass	['biomass']	35	0.005871498	2.29e-06	2 563.97
gw	['gw']	23	0.003858413	1.58e-06	2 442.03
hydropow	['hydropower']	11	0.001845328	7.59e-07	2 431.26
biodiesel	['biodiesel']	6	0.001006543	4.68e-07	2 150.73
pv	['pv']	21	0.003522899	2.45e-06	1 437.92
ethanol	['ethanol']	23	0.003858413	3.09e-06	1 248.68
hydroelectr	['hydroelectric', 'hydroelectricity']	9	0.001509814	1.826e-06	826.84
mw	['mw']	16	0.002684113	3.31e-06	810.91
geopolit	['geopolitical', 'geopolitics']	9	0.001509814	2.047e-06	737.57
cellulos	['cellulose', 'cellulosic']	6	0.001006543	1.375e-06	732.03
solar	['solar']	110	0.01845328	3.09e-05	597.19
fossil	['fossil']	29	0.004864956	9.33e-06	521.43
har	['harnessed', 'harnessing', 'harnesses']	6	0.001006543	2.119e-06	475.01
thermal	['thermal']	24	0.00402617	1.26e-05	319.54
turbin	['turbine', 'turbines']	14	0.002348599	7.94e-06	295.79

6.3. Discussion

Keywords can be extracted from documents, using the proposed method. With standard text mining techniques, pronouns, prepositions, and common verbs are excluded, and furthermore, words that are essentially the same (that have the same stem) are grouped. Names (of companies and of persons) are however not excluded. From the work described in Chapter 3, it is clear that they can be used to find companies.

Domain expertise is required to compose a list of keywords for finding companies that are active in a specific industry. The keyword found with the proposed method can be used as a tool by domain experts, who are expected to know, which of the keywords found are useful and which are not.

Wikipedia is used as a source, here, because it is well-established, and articles have been written, reviewed and maintained by many dedicated people. Those articles however also have a couple of disadvantages, the main one being that a large part of Wikipedia articles is often dedicated to the topic's history, which may not be representative to the current state of the topic. Another disadvantage is that an encyclopaedic text is generally written in a different style than corporate texts are, because it intended to inform as objectively as possible, while the latter generally aims to attract potential customers or investors. Finally, a Wikipedia article is generally more theoretical, whereas company websites put more emphasis on practical applications.

Using negative keywords is often an effective strategy, in narrowing down the search results. It is recommended, when a large proportion of the search results are not relevant to the target topic, but rather to another topic that might share the same keywords. For example, a construction engineer searching for certain properties of a metal may want to exclude the search results concerning the music genre metal. A good strategy, in that case, is to add '-guitar' to the search query. To find companies that are active in a certain industry, it is worthwhile finding out, which topics undesired search results pertain to. The keyword extraction method can then be applied to those topics. Keywords found that have nothing to do with the target industry can thus be used as negative keywords.

7 Conclusions and recommendations

7. Conclusions and recommendations

The 'Web Intelligence for Drones' study has particularly targeted the development of a methodology and of the tools required to retrieve information from the www, for businesses based in EU Member States that have their main activity in the civil drones sector. The methodology and tools were developed within a perspective of generalisation, wherever possible, i.e. targeting their application to Member States not covered in this study and to other emerging economies (as described in Chapter 6).

The method and tools developed aim to identify drone companies through taking several web search approaches, without relying on known lists of drone companies. A list of drone companies in Spain was specifically built for the study, to be used as a benchmarking reference in assessing the results produced by the automatic method developed. A similar list was made available for Italy by the Drone Observatory Department at the Polytechnic University of Milan. Both lists were used exclusively to monitor the developed method's performance, at its different stages.

The main findings of the study can be resumed as follows:

The search-engine based method developed is able to find websites of drone companies active in the countries studied. Considerable numbers of new drone companies are found, when compared to the initial lists of drone companies.

This is confirmed for Spain and Italy, when comparing the URLs found with lists of known URLs of drone companies. Potentially, 312 and 267 new Drone companies were discovered for Spain and Italy, respectively. A list of known URLs was not available for Ireland, making the list of 66 URLs found a novelty. Expert checking of samples of the new URLs found revealed that, for both Spain and Ireland 85 % and for Italy 78 % of those URLs were correctly assigned to drone companies active in the country.

As reported in Chapter 3, however, the search approach developed does not detect all hitherto known drone websites. The method is nevertheless able to expand the universe of the already known drone companies. This result actually goes to confirm the sector's dynamism, one of the challenges considered at the beginning of the study. It also leads to the conclusion that commercial registers are not the best starting point for retrieving this sector's data nor, in general, that of other emerging economies.

The availability of an overview website and/or PDF-files with lists of drone companies have a positive effect on the results yielded by the search-engine based approach.

The method developed is able to find drone companies overview websites and/or PDF-files. This offers important advantages, in particular in those cases, in which there is no information about the

existence of such consolidated information. The method developed is further able to differentiate between companies and individuals if the overview files broadly list drone operators (e.g. both individuals and companies that may have a license to operate drones). The method also identifies the URLs of the companies that are identified through these overview lists (making use of the URL Search script produced in the context of the ESSnet Big Data 2017 project).

Enabling the identification of websites of drone companies active in a specific country, the classification model provided valid results, when tested on other countries.

The supervised Machine Learning model (Logistic Regression) trained on Spanish drone websites provided valid results, when applied to Irish and Italian websites (accuracy of 84 % and 85 %, respectively; sample manually checked) and setting a probability cut-off value of 0.6. These results demonstrate that the classification model developed can be applied to other countries, for which it provides findings of comparably high accuracy.

This indicates that the model could be applied to websites in other countries, if: i) websites are written in English or if a correct translation list has been created, and if ii) the features applied to a drone website for the country studied are comparable to those used for Spain, Ireland and Italy.

Important characteristics of companies can be extracted from the websites.

This is the case for a number of basic characteristics, such as the address, telephone number and email address, etc. The website's creation date was furthermore used as a proxy for the company's start date. VAT numbers were the most challenging characteristics to obtain. A number of additional characteristics such as e-commerce activity, Job vacancies, etc. can also be extracted.

Economic variables such as the number of employees and company turnover are however not readily available on company websites. That information could be extracted from online commercial registers. However, additional information may only be available for a limited number of new companies, since in online commercial registers, generally, data is only available for the main companies in a sector. This further suggests that online commercial registers are not a good starting point for retrieving data on this kind of companies.

The method developed can be generalised to an important degree.

Generalisation can be considered from two perspectives: across countries and across economic sectors (emerging economies).

Scripts 1 to 5 (see Chapter 6) can be applied to other EU countries with certain minor updates (i.e. key terms in the national language used for the search queries, e.g. 'drone' in English, 'dron' in Spanish, 'Drohne' in German). In terms of generalisability across countries, an advantage of the Drone sector is the similarity of terms used in different countries. The domain-specific key terms used for the search queries are very much linked to standardised English terminology (e.g. RPAS, UVS, UAV). The disadvantage of using of these common terms (i.e. English terminology) is that the search explores all results. From the economic sectors generalisation perspective, the identification of sector-related key terms could be partially tackled through using Wikipedia as a source of information on the domain (Chapter 6). For an accurate selection of key terms that are also relevant to other sectors, however, an expert check or an exploratory sector analysis are recommended, in addition to the key-word-extraction based approach.

The classification model (Script 6) is generalisable across countries for English-language websites and websites that can accurately be translated into English, whenever the features applied to the drone websites of the country studied are comparable to those used in Spain, Ireland and Italy.

A country-specific data-extraction script (Script 7) is prepared for the identification and extraction of information, given the need to adjust to specific national features such as VAT and address details, for example. When considering generalisation across economic sectors, the country-specific scripts would be suitable for extension to other emerging economies, subject to the adaptation of key specific terminology, as described, as well as in relation to extracting information on the activities of the respective sectors.

The method developed presents, as its main strength, the ability, through the www, to automatically identify and extract information on drone companies active in the three European countries studied (Spain, Ireland and Italy), as well as detecting a considerable number of 'not yet known' drone companies. Although, the search approach developed does not find all 'already known' drone companies, it has the advantage of providing insights into new developments in the market. Subject to the availability of lists of drone operators or drone companies (e.g. such as those published by the Spanish Aviation Safety Authority), an alternative sequence or combination of steps could be considered, to derive combined lists of drone companies, that better map the actual market in each country (i.e. full method deployment to identify new companies combined with the deployment of separate scripts to identify and extract the information for known Drone companies, e.g. Scripts 3a, 3b and 7). The aim of the current study was to develop a chain-like methodology and the relevant tools for identifying and retrieving the information from the www on businesses in the civil drones sector, that could be generalised for collecting similar information on other emerging economic trends. The chain-like methodology does not restrict the use of alternative sequences or the combination of steps.

References:

Aggarwal, C. C., (2016), Mining Text Data, In: Aggarwal, CC, editor. *Data Mining: the Textbook*, New York, Springer, pp. 429–455.

Antons, D., Grünwald, E., Cichy, P., Salge, T.O. (2020), The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3): 329–351.

Appertium (2021), A free/open-source machine translation platform, Link: http://www.apertium.org and https://wiki.apertium.org/wiki/Main_Page.

Arora, S.K., Li, Y., Youtie, J., Shapira, P. (2016), 'Using the Wayback Machine to mine website in the Social Science: A methodological resource', *Journal of the Association for Information Science and Technology*, 67(8): pp. 1904–1915.

Arora, S.K., Li, Y., Youtie, J., Shapira, P. (2020), 'Measuring dynamic capabilities in new ventures: exploring strategic change in US green goods manufacturing using website data', *The Journal of Technology Transfer*, 45: 1451–1480.

Arora, S.K., Youtie, J., Shapira, P., Gao, L., Ma, T.T. (2013), 'Entry strategies in an emerging technology: a pilot web-based study of graphene firms', *Scientometrics*, 95: 1189–1207.

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

Blaxter, L., Hughes, C., Tight, M. (2010), *How to research*, McGraw-Hill Education, New York.

Boix, R., Hervás Oliver, J.L., de Miguel Molina, B. (2015), 'Micro-geographies of creative industries clusters in Europe: from hot spots to assemblages', *Papers in Regional Science*, 94(4): pp. 753–772.

Bruni, R., Bianchi, G. (2020), 'Website categorization: a formal approach and robustness analysis in the case of e-commerce detection', Expert Systems With Applications, 142: 113001.

Daas, P.J.H., Van Der Doef, S. (2020), 'Detecting innovative companies via their website', *Statistical Journal of the IAOS*, 36(4): pp. 1239–1251.

Daas, P.J.H., van der Doef, S. (2020), 'Detecting Innovative Companies via their Website', *Statistical Journal of IAOS*, 36(4), pp. 1239–1251, doi/10.3233/SJI-200627.

De Miguel Molina, B., Hervás Oliver, J.L., Boix, R., de Miguel Molina, M. (2012), 'The importance of creative industry agglomerations in explaining the wealth of European regions', *European Planning Studies*, 20(8): pp. 1263–1280.

Ellinger, A.E., Lynch, D.F., Hansen, J.D. (2003), 'Firm size, web site content, and financial performance in the transportation industry', *Industrial Marketing Management*, 32: 177–185.

ESSnet Big Data (2017), Deliverable 2.2, Methodological and IT Issues and Solutions. Work-package 2 of the ESSnet Big Data I.

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

Gentzkow, M., Kelly, B., Taddy, M. (2019), 'Text as Data', *J. Econ. Lit*. 57(3), pp. 535–574. doi:10.1257/jel.20181020.

Gök, A., Waterworth, A., Shapira, P. (2015), 'Use of web mining in studying innovation', *Scientometrics*, 102: pp. 653–671.

GOPA (2021), Deliverable D1, Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry (available at: https://ec.europa.eu/eurostat/cros/content/d1-overview-use-case-drones-and-first-draft-general-methodology_en)

GOPA (2021), Deliverable D2, Data Retrieval. Report 2 of the Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry (available at: https://ec.europa.eu/eurostat/cros/system/files/d2_dataretrieval_final.pdf)

GOPA (2022), Deliverable D3.1, Classification of Drone company websites, Report 3 of the Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry (available at: https://ec.europa.eu/eurostat/cros/content/d3dataextractionpartaclassificationmodel_en)

Grant, R.M. (2018), *Contemporary Strategy Analysis, 10th Edition*. Wiley, New Jersey.

Hillen, J. (2019), 'Web scraping for food price research', *British Food Journal*, 121(12): pp. 3350–3361.

Jo, T. (2019), 'Introduction', In: Text Mining, *Studies in Big Data*, vol. 45, Springer, Cham.

Kinne, J, Axenbeck, J. (2020), 'Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study', *Scientometrics*, 125: pp. 2011–2041.

Kotler, P. Armstrong, G. (2018), *Principios de Marketing*, 17 edición, Pearson, Madrid.

Li, Y., Arora, S., Youtie, J., Shapira, P. (2018), 'Using web mining to explore Triple Helix influences on growth in small and mid-size firms', *Technovation*, 76-77: pp. 3–14.

Libaers, D., Hicks, D. and Porter, A.L. (2010), 'A taxonomy of small firm technology commercialization', *Industrial and Corporate Change*, 25(3): pp. 371–405.

Massimino, B. (2016), 'Accessing online data: web-crawling and information-scraping techniques to automate the assembly of research data', *Journal of Business Logistics*, 37(1): pp. 34–42.

Mirtsch, M., Kinne, J., Blind, K. (2021), 'Exploring the adoption of the International Information Security Management System Standard ISO/IEC 27001: a web mining-based analysis', *IEEE Transactions on Engineering Management*, 68(1): pp. 87–100.

Patel, J.M. (2020), 'Getting structured data from the Internet: running web crawlers/scrapers on a big data production scale', *Apress*, Springer, New York.

Pedregosa et al. (2011), 'Scikit-learn: Machine Learning in Python', *JMLR 12*, pp. 2825–2830.

Porter, M.E. (2008a), *Competitive Strategy: Techniques for analyzing industries and competitors*, Free Press, New York.

Porter, M.E. (2008b), *Competitive advantage: Creating and sustaining superior performance*, Free Press, New York.

Porter, M.E. (2008c), 'The five competitive forces that shape strategy', *Harvard Business Review*, 86(1): pp. 78–93.

PUlearning (2021), Website of the pulearn python library, Link: https://pypi.org/project/pulearn/

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/

Rothaermel, F.T. (2019), *Strategic Management*, McGraw-Hill Education, New York.

Schindler, P.S. (2019), *Business research methods*, 13rd ed. McGraw-Hill Education, New York.

Shapira, P., Gök, A., Salehi, F. (2016), 'Graphene enterprise: mapping innovation and business development in a strategic emerging technology', *J Nanopart Res*, 18: 269.

Shapira, P., Gök, A., Klochikhin, E., Sensier, M. (2014), 'Probing 'green' industry enterprises in the UK: a new identification approach' *Technological Forecasting & Social Change*, 85: pp. 93–104.

Sklearn (2021), Overview of Supervised classification algorithms available in the sklearn library in Python, Link: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

vanden Broucke, S., Baesens, B. (2018a), 'Introduction', In: *Practical Web Scraping for Data Scienc*e, Apress, Berkeley, CA.

vanden Broucke, S., Baesens, B. (2018b) From Web Scraping to Web Crawling. In: *Practical Web Scraping for Data Science*. Apress, Berkeley, CA.

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

Youtie, J., Hicks, D., Shapira, P., Horsley, T. (2012), 'Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies', *Technology Analysis & Strategic Management*, 24(10): 981-995.

Annex I: Creating a list of drone company websites for Spain

In order to accurately determine, how many and which drone websites do exist in Spain, a list of drone websites was created. To construct such a list, a set of PDF documents collected by the team, provided various overviews of Spanish drone companies, and it was taken as the starting point. The documents used were:

- listado_proveedores_ats_ais_cns_certificadas.pdf
- registro_autorizaciones_operac_aereas_uas.pdf
- 20201222_Listado.Fabricantes.pdf (updated version available at: https://www.seguridadaerea.gob.es/sites/default/files/Listado_Fabricantes.pdf)
- Listado entidades reconocidas.pdf (updated version available at: https://www.seguridadaerea.gob.es/sites/default/files/Listado%20entidades%20reconocidas.pdf)
- listado_operadores.pdf
- Tabla Maestra de modelos de UAS.pdf

The relevant tables were extracted from all 6 PDF files. Only one document however contained links to the websites of drone companies; a total of 3 URLs were obtained from this document. Due of that, it was decided to extract the columns 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. The websites corresponding to those companies then needed to be found. The URLsFinder approach developed in the ESSnet Big Data (2020) was used for that task. In principle, this approach uses a search query which, in this case, includes the company's name combined with 'Espana'. The links returned indicates the potential existence of a website for that company. The payed-for Google search engine was used for this task and the first 10 links returned were collected for each search query. A very important new extension was included: in addition to the first 10 links, the searched page was also checked for the presence of a Google Knowledge Graph section. This is commonly shown in the right upper part of the webpage and it usually includes a map showing the searched-for company's location, together with further information. When this section was available, the occurence of a button with a website link was checked for. If such a button could be 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 consisting of 5 369 unique links was established. The combination of company names and links were manually checked by the team. This resulted in a final list of 1 097 unique combinations of secondary and top-domain names for drone company websites in Spain. Remarkably, this final list contained 907 of the links initially found using the extended URLsFinder approach. A total of 190 manual corrections were required; of those, 35 contained new (not yet known) domain names. That list was used to check the results provided by the different versions of Script 1 and, further, for the development of the method to detect and classify drone company websites.

Annex II: Hard- and software requirements

FOR URL RETRIEVAL

The scripts were written in Python 3.7, with the exception of the page visit browser emulation script, which was written in Bash. All scripts start 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, or the file to be processed. In addition, some scripts included optional settings. For instance Script 1 includes the option of indicating 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. Some scripts have the option to start processing at a particular row in the input file. For example, Script 7 can begin at the first record (default) or at any other line number. This is particularly useful, when processing takes considerable time, as it enables the process to be distributed over several days. The progress of Scripts 1 to 4b is stored in a log-file and it is shown on screen. The scripts were developed on computers using Ubuntu 20.04 as their operating system and they 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 Chromium driver must be installed; they are specifically needed to access the Java-based DuckDuckGo search engine and to scrape data from some social media and website pages. Xdotool also 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, in addition: pandas, googlesearch-python, timeout-decorator, nltk (make sure to subsequently import the stopwords for each language), pdfminer.six, selenium, search-engine-parser, configparser, nest-asyncio, sklearn, pickle, langdetect, and collections.

The scripts were successfully run on an Ubuntu laptop with 16 GB of memory, an Ubuntu desktop with 32 GB, an Ubuntu server with 256 GB and a raspberryP4 with 8 GB. In principle, each script has the option (in the Ini-file) to run as much of the program as possible in parallel, so as to speed things up, but in practice it became 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 the 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 high frequency, it takes the script a considerable time to complete the task. Also, a standard wait time of 20 seconds is used between queries, except for Bing (5 sec.) and for the payed Google subscription (0 sec.). This is done to prevent blocking. Scripts 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. Scripts 4a and 4b contain a browser emulation script that cannot be run in parallel; it otherwise disturbs the assignment of the browser IDs used and the data are lost. The latter, combined with the very large number of links to be processed and, in particular, the time it takes to determine whether a website really exists, are the reasons for which it takes Script 4b so long complete running. The best way to speed up Script 4b is to run it on multiple computers and to 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, spreading the queries over multiple machines, if one can ensure that those machines have non-identical IP-addresses during the period of scraping. Script 5 and 6 do have to scrape websites and they are usually finalised within a few hours. Script 7 scrapes websites but, because of the much smaller numbers of websites to visit and the fact that scraping is done in parallel, this usually takes less than 3 hours.

FOR THE CLASSIFICATION OF DRONE COMPANY WEBSITES

All scripts were written in Python 3.7.3. The following libraries were used: sklearn (version 0.21.2; Pedregosa et al., 2011), PUlearn (version 0.07), pickle (version 4.0), numpy (version 1.16.4), multiprocessing (native to Python), BeautifulSoup4 (bs4, version 4.9.1), langdetect (version 1.0.7) and NLTK (version 3.4.1). The translation program Apertium (2021) was used for the off-line translation of large volumes of texts, with the English-Spanish and English-Italian language pairs installed. All scripts were run on an Ubuntu laptop with 16 GB and 8 threads, except for the translation software which was run on an Ubuntu Workstation with 256 GB and 40 threads. The Script5_Translate_ini1.py, available in the repository, contains the code used to translate Spanish to English words. The figures included were created in Rstudio (1.25)

FOR THE DATA EXTRACTION

All scripts were written in Python 3.7.3. The following libraries were used: pandas (version 1.1.2), re (version 2.2.1), numpy (version 1.16.4), multiprocessing (native to Python), BeautifulSoup4 (bs4, version 4.9.1), langdetect (version 1.0.7) and NLTK (version 3.4.1). All scripts were run on an Ubuntu Workstation with 32 GB of memory and 8 threads. The results were visually analysed in Rstudio (version 1.25).

Annex III: Top 20 features in the Logistic Regression model

Positive features		Negative features		
Nr.	Weight	Feature	Weight	Feature
1	2.33986	audiovisual	-1.72563	companies
2	2.07578	drones	-1.44047	components
3	1.98355	engineering	-1.43618	subscribe
4	1.96928	dronF	-1.38329	Feature_language
5	1.95713	drone	-1.38040	line
6	1.69981	production	-1.23903	manufacturers
7	1.57253	per	-1.23780	national
8	1.56275	aerial	-1.22316	clothes
9	1.52524	projects	-1.22071	free
10	1.52297	environmental	-1.19948	receive
11	1.49231	dji	-1.18125	days
12	1.49225	clients	-1.17714	sep
13	1.46029	photographs	-1.15051	medical
14	1.43877	documentary	-1.10680	accessories
15	1.31980	video	-1.10312	prices
16	1.29902	fly	-1.07285	research
17	1.26139	provide	-1.06833	publish
18	1.25108	management	-1.04407	association
19	1.24898	videos	-1.04116	search
20	1.23612	com	-1.01780	sanitary

Annex IV: URLs of companies found for Spain

25drones.com
360dron.com
5gpilotosvalencia.orange.es
a3sdrones.com
academy.dronitec.es
accessdrone.es
acepdron.cat
aceroestudio.com
adhoc-digi.com
adqando.com
adronepilot.com
aereodron.es
aerialdreamz.com
aerialtronics.mobendum.com
aerialworks.es
aero-inspecciones.es
aerocamaras.es
aerofilmhd.com
aerographstudio.com
aeromedia.es
aeropic.tv
aerosportfoto.com
afjfly.com
agdronec.com
agrodato.com
agrodex.es
agrodronecr.com
ai2.upv.es
aircatglobal.com
airdronmelide.es
airelectronics.es
airevisual.es
airmagni.com
airmedia360.com
airpull.com
al-top.com
alphadrones.es
alquilerdrones.eu
alquilerdronesmadrid.es
animadrone.com
anti-drone.eu
aotecnica.com
areadron.com
arizcuren.com
artofflight.es
ascant.org
asdronica.com
asesoriadron.com
asgdrones.com
atdrones.es
aulanatura.org
avdron.es
avionesnotripulados.wordpress.com
ayresl.es
bai-sa.es
baldodrones.es
barcelonadronecenter.com
bertendsp.com
blanch-internacional.com
bluemg.eu
boscalia.org
campusrpas.com
canso.org
cartodesia.com
casa-film.com
casdron.es
centervol.es
centineldrone.com
chiclayodrones.com
compradron.com
csysc.es
ctin.es
cursosteledeteccion.com
customdrone.es
daves-films.com
davidalfaraz.com
davidsaez.net
dds-vueloexperto.com
deaparatos.com
dflyvision.com
dji-guadalajara.negocio.site
djiasturias.smartgo.es
djivalencia.smartgo.es
dondedrones.com
dron.uca.es
dronaerea.com
dronakacademy.com
dronalia.net
dronamedida.com
droncompany.com
drondrones-om-dron-drones-shop.webnode.es
drone-hunter.com
drone.episenses.com
dronealbacete.com
droneartists.eu
dronecadiz.com
droneduca.es
dronefuture.es
dronehibrido.com
dronelightshow.es
dronemadrid.com
dronepasion.com
drones.org.es
dronesafestore.com
dronesbaratosya.com
dronesbarcelona.es
dronescantabria.com
dronesceuta.es
dronescondor.es
dronesnorte.es
dronesskycam.com
droneteca.com
dronetour-oficial.webnode.es
dronetvspain.com
droneup.es
droneuropa.com
dronevision.es
droneymas.com
drongal.es
dronical.com
dronlimits.com
dronmodular.com
dronnavarra.es
dronorte.com
dronpixel.com
dronpublicidad.es
dronquixote.es
dronservice.es
dronservice.wordpress.com
drontop.com
dronvalencia.es
dronyco.negocio.site
dsbaero.com
ecapture3d.com
eco07.com
ecoespaciodrones.com
efisky.com
ekofastba.com
eldroner.com
elecnor-deimos.com
electronicarc.com
electroya.com
embention.com
emedicaldrone.es
en.ebredrone.com
en.verdedrone.com
en.xpeidrone.com
enerdrones.com
escuela-de-vuelo.com
escueladepilotosdedrones.com
escueladerpas.com
esdronia.es
euroflytec.smartgo.es
eurousc.es
exmera.com
fastfly.itg.es
filmdrone.es
fira.com
fireaviation.com
flightechspanish.weebly.com
flyanddo.com
flydronespain.es
flyingandfly.com
flyworks.es
formacion.cursodedrones.es
formadron.es
fotopro360.com
fuerteventuradrones.es
futurdrone.com
futurizable.com
fzingenieros.es
generalivalladolid.es
geodesical.com

geomati-k.com
gesingeo.es
gibraldrone.com
girodrones.es
godrone.es
gofau.es
grupobroadcast.es
grupocopisa.com
grupoforma-t.com
gruponadir.es
helifilm-foto-video-aereo-sl.negocio.site
hidrone.es
hyper-drone.org
iberdron.com
iberfdrone.es
iberodronefilms.wordpress.com
ibizabyair.com
ibizadroneworks.es
icom3d.com
ideaingenieria.es
ihobbies.es
imsdrones.com
indaldrone.es
ingecor.net
innodrone.es
internetofbusiness.com
intranet.barcelonadronecenter.com
irisdronespecialists.eu
isoairecanarias.com
itg.es
ivadrones.com
iworu.com
jessicapinto.es
joanlesan.com
jobtodron.com
josedrones.com
kreativedrone.com
kumodrones.es
librosdevuelo.com
licuas.es
livedron.com
mainaketopografia.com
mandarinak.es
martinzvara.wixsite.com
master-aviation.com
masterdronix.com
matriculasdron.es
mdrone.com
melodrone.es
militarymachine.com
milmiradas.es
miradacenital.com
miscodron.es
mldrone.es
mobus.es
moradadrones.com
moscatingenieria.com
motion-graphics.video
msolutions.es
multicoptero.com
muydrones.com
mybydrone.com
mycoordinates.org
nickstubbs.com
nordesdrons.com
objetivoaereo.com
ocdronservices.com
ofiteat.com
ojoaereo.com
omnicam4sky.tv
oneairpictures.com
opendrone.es
operadoruas.es
paintec.tech
piafmajorque.es
pildo.com
pilotadrones.es
piloto-drones.es
pirineosdrone.com
plusprointec.es
precisiondrone.es
pro-aesa-web-win.azurewebsites.net
prunetec.es
robotdronica.com
rocktoroad.com
roxudron.com
rpalabs.es
safetytude.es
salmerondrons.com
salmerondrons.smartgo.es
sdle.info
sdtstore.smartgo.es
seadrone.es
securitydron.es
segurclick.com
segurosaviacion.es
segurosdron.wordpress.com
segurosparadrones.com
servicioscondrones.com
sierradrone.es
sismodrone.smartgo.es
site4drone.com
sixarms.com
skycat.pro
skydrones.aero
skydronevision.es
skydronex.com
sparrowfour.com
spotdron.es
stgo.es
sydrone.com
technidrone.es
tecnicasdelsuelo.es
terradron.cat
thedronesland.com
tibudrones.com
tienda.avistadrone.com
tmidrones.es
todrone.com
topdron.es
topdrones.net
toylabrc.com
tueventodron.com
ualidrones.com
uastrainingcenter.com
universodrone.com
unmannedexpert.com
vdevideo.com
videoyfotocondrones.es
visiondrone.cat
visionh.es
vister.es
vuelosdrone.com
x-drone.eu
xn--airdroneespaa-tkb.com
yuneec-futurhobby.com
zenitdrones.com
zenitingenieria.com
zettadron.com

Annex V: URLs of companies found for Italy

2-way.it
3fedin.it
abdroni.com
accademiadelvolo.it
achrom.info
acl.mit.edu
advister.it
aecitalia.it
aerialmediapros.com
aeroclubancona.com
aeroclublugo.it
aeroclubroma.it
aerohabitat.org
aerokomp.com
aeronike.com
aerrobotix.com
afdrones.it
agricolturadiprecisione.info
agrosurvey.farm
airservicecenter.it
airwrx.com
algowatt.com
allservicewebagency.it
archivio.romadrone.it
assicurazione-drone.it
assistenza-droni.com
assistenza-droni.it
asterx.it
aviony.net
b2bonline.it
bearfpv.it
bigactions.com
biolineigieneambientale.it
birtimichele.com
bizmodel.it
bma-srl.it
brixiadrone.it
cabibroker.com
campiavventura.it
cardtech.davidmonetti.com
cartografiasapr.it
cec-cuneo.it
centroformazionevolo.it
comune.vimodrone.milano.it
corsodroni.com
costruzionedroni.it
creokitchens.it
d-flight.jtdrone.com
d-fligt.com
d-fly.it
dallalto.pro
data-fly.it
dauvea.it:443
defencesystem.net
degeaitalia.wixsite.com
docdrones.retedoc.net
drivedrone.it
droincompany.eu
drone-store.it
drone24hours.com
droneairview.com
dronedario.com
dronedpj.com
droneflightacademy.it
dronefly.shop
droneinitaly.it
droneitalia.net
droneitalia.online
dronelab.it
dronemaster.it
dronemotionsenterprise.com
droneprofessionalsolution.com
droneproservice.it
dronerp.com
dronersclub.com
drones-spare-parts.com
drones.horusdynamics.com
dronesafestore.com
droneserviceitaly.com
dronetopoprogram.eu
dronetopoprogram.it
dronex-roma.com
dronext.eu
dronextreme.eu
dronezero.net
dronezine.it
dronieagroecosistemi.it
dronissimo.it
dronitaly.wordpress.com
dronotica.weebly.com
dronus.com
dslrpros.com
easydroni.it
easyservicesolutions.com
eccpalestine.org
edilizianamirial.it
eipro.elettronicain.it
elifriulia.it
eliteconsulting.it
elyshop.com
enac.gov.it
enav.it:443
enontheroad.com
entoservice.it
errealcubo.com
eurousc-italia.it
extinsrl.it
eyedrone.it
f-i-d.it
fantasyland.it
flightdrone.it
fly-academy.it
flybri.it
flydron.it
flydronservice.it
flylike.it
flyvalue.eu
fotododici.com
fpvdroneracing.it
frangerini.it
free.netcurso.net
frontierprecision.com
genegis.net
genialbrand.com
geo-tech.it
geoapp-italia.com
geoapp.it
geodatalab.it
geodroneservizi.com
geometrastefanobergamini.it
geosatsrl.it
geoskylab.com
gocamera.it
gotofly.it
grafichefioroni.it
gruppomodellisticobolognese.it
hardis.it
hdaerial.com
helicenter.it
helyx.it
hetronic.com
hobbyhobby.it.clearwebstats.com
hobbymedia.it
hobbyqueenitalia.com
humanfactoritalia.com
ibtimes.com
icaroschool.com
ilmiodrone.it
imeasolar.com
imeasolar.it
infodrones.it
infomobility-italia.com
integraerospace.it
intellisystem.it
iot-italia.net
istruttoredivolo.com
ita.italdron.com
italianhotelgroup.net
italiasolutions.eu
italydrone.it
jmotion.it
jtdrone.com

keelcrab.com
keytop.it
lattoneriafrassi.com
lecasemarcieglie.com
leicesterdrones.com
levita.cloud
lifedrone.eu
lightairplanes1.com
ludicando.it
macitynet.it
mainstreamagency.it
masterdrone.it
medilifegroup.com
milanodroni.com
modelgiochi.eu
modellandia.com
modellismocrazytime.com
morriconi.com
mrkvideomaker.it
multicottero.com
my-fpv.com
mydroneacademy.net
mydronestock.com
nemeasistemi.com
nexusitalia.srl
novatest.it
nposistemi.it
nwservice.it
opendeel.com
operatori-apr.it
overit.it
p1hh.piaggioaerospace.it
pheromed.com
pipistrel-aircraft.com
pitom.eu
pmcomunicazione.com
profserv.it
projectems.it
pumasecurity.it
puntofotonline.com
quadcopternews.it
quadricottero.com
radioflyshop.com
rinaprime.rina.org
rivistageomedia.it
romadrone.it
rov-subacquei.it
sagetech.com
saidbegov.com
saltlemon.it
sapritalia.com
scaicomunicazione.com
scuoladroni.pro
seadrone.it
semprebonlux.it
servicedrone.it
services.italdron.com
serviziodroni.com
servizipa-group.it
sgd-group.com
shop.jtdrone.com
siadsrl.net
sitech-italia.com
sitodiprovadifabietto.weebly.com
skycat.pro
skycrabacademy.net
skylabstudios.it
skyters.it
softhillsrl.com
staffmillennium.it
stampaepubblicita.it
stefanotrojani.com
store.drone-zone.it
store.salentodroni.com
strumentitopografici.it
studioctl.com
studioscalisi.com
sviluppogenova.com
theaviationist.com
thedronehangarllc.com
titanpics.com
topoprogram.com
toscanadrone.com
trc-drone.com
trimble-italia.com
uas-group.com
uav.ap74.it
ugcs.com
umbriadronevision.com
unikey.it
universodroni.it
unmannedrc.com
urbandrones.com
urbevideo.com
venditadroni.it
vertworx.com
vldistribution.jimdofree.com
vrxracing.it
wearnews.it
widroneservice.com
win.new-tec.it
worldappeal.it
zanzotteraengines.com
zedprogetti.it
zeeco.com

Annex VI: URLs of companies found for Ireland

academy.safedrone.ie
aerial.ie
aerialdrone.ie
aerialeye.ie
aerialfilming.ie
aerialfilmingireland.ie
aerialmedia.ie
aerialsurvey.com
aerosky.ie
airviewmarketing.com
apexsurveys.ie
asmireland.ie
avtrain.aero
camera.ie
checkoutfast.thrivecart.com
datadrone.ie
davidclynchphotography.com
drone.ie
drone360services.com
dronebuildinginspection.ie
dronecaddy.ie
droneinsurance.ie
dronemaps24.org
droneon.ie
droneservicesireland.ie
dronesurveying.ie
dronetrainingireland.wordpress.com
dronevideohire.com
droneviews.ie
eireialcreations.ie
engineerswithdrones.ie
flymedia.ie
geoaerospace.com
geoinspect.eu
gravityconstruction.ie
idrone.ie
inspiregroup.ie
iogeomatics.ie
jpdmedia.ie
kestreldrone.ie
khdroneservices.ie
landandaerialsurveys.ie
landandmineralsconsulting.com
livedrone.ie
mcdonaldsurveys.ie
mywebsitephoto.com
output42.com
precisionaerialservices.ie
radiocontrolledshop.ie
rentauav.com
skyfab.ie
skypix.ie
standardfire.ie
superfly.ie
thebasementmedia.com
thedroneguys.ie
thefreedictionary.com
thonon-modelisme.com
topdrone.ie
uavservices.ie
ugcs.com
unchartedhorizons.ie
uspacefinland.com
protectstat.com
videobase.ie
vmg.ie
website.informer.com

Annex VII: Generic keyword extraction: implementing the method

The method is implemented in the Python function extract_keyword_profile. The implementation is built on the well-established Python libraries **bs4** (Beautiful Soup) for extracting text from a webpage, **nltk** for text mining techniques, and **wordfreq** for word frequencies from the Exquisite Corpus.

This function has the following input parameters:

- **url** the URL of the source document
- **min_count** the minimum count that is required for stemmed words to be contained in the output list (this is useful when excluding names). The default is set to 5.
- **filename** the filename of the output file
- **language** the language of the source document.

The output is a table containing the following columns:

- **stem** the stemmed words
- **words** the original words that share the same stem
- **count** the occurrences of the original words in the source document
- **freqSrcDoc** the relative frequencies of those words in the source document
- **freqExCorp** the relative frequencies of those words in the Exquisite Corpus
- **freqRatio** the ratio between those two frequencies, which is used as an indicator of how much a word is specific to the source document.

For details, please see the Python code documentation.

GETTING IN TOUCH WITH THE EU

In person
All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email
Europe Direct is a service that answers your questions about the European Union. You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications at: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents
For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU
The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

KS-TC-22-004-EN-N

Web intelligence for measuring emerging economic trends: the drone industry

The 'Web Intelligence for Drones' project targeted the development of a method to retrieve information from the web on businesses based in EU countries that have their main activity in the sector of civil unmanned aerial systems (UAS) also known as 'drones'.

The study presents a novel methodology and the tools developed to identify drone companies through the web and to extract company-relevant information from their websites. The method was developed with a perspective of generalisation in mind (across countries and across economic sectors) wherever possible. It has already been fully applied to three European countries (Spain, Italy and Ireland), and the results are presented in this report.

For more information

https://ec.europa.eu/eurostat/

Publications Office of the European Union

ISBN 978-92-76-53307-8