NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

Data Science and Information Technologies

SPECIALIZATION

Big Data and Artificial Intelligence

MASTER'S THESIS

Automating the data acquisition of Businesses and their actions regarding environmental sustainability: The Energy Industry in Greece

Christina G. Borovilou

ATHENS

January 2023

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Επιστήμη Δεδομένων και Τεχνολογίες Πληροφορίας

ΕΙΔΙΚΕΥΣΗ

Μεγάλα Δεδομένα και Τεχνητή Νοημοσύνη

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτοματοποίηση εξαγωγής Δεδομένων του επιχειρηματικού κλάδου ως προς τις δράσεις τους για την περιβαλλοντική βιωσιμότητα: Ο τόμεας Ενέργειας στην Ελλάδα

Χριστίνα Γ. Μποροβήλου

ΑΘΗΝΑ

Ιανουάριος 2023

Master's Thesis

Automating the data acquisition of Businesses and their actions regarding environmental sustainability: The Energy Industry in Greece

Christina G. Borovilou

S.N.: DS1200008

SUPERVISORS:

Harris Papageorgiou, Research Director, Athena Research Center

EXAMINATION COMMITTEE:

Harris Papageorgiou, Research Director, Athena Research Center

Katsouros Vasilis, Research Director of ILSP, Athena Research Center

Theodore Dalamagas, Research Director, Athena Research Center

January 2023

Διπλωματική Εργασία

Αυτοματοποίηση εξαγωγής Δεδομένων του επιχειρηματικού κλάδου ως προς τις δράσεις τους για την περιβαλλοντική βιωσιμότητα: Ο τόμεας Ενέργειας στην Ελλάδα

Χριστίνα Γ. Μποροβήλου

Α.Μ.: DS1200008

ΕΠΙΒΛΕΠΟΝΤΕΣ:

Χάρης Παπαγεωργίου, Διευθυντής Ερευνών, Ερευνητικό Κέντρο Αθηνά

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Χάρης Παπαγεωργίου, Διευθυντής Ερευνών, Ερευνητικό Κέντρο Αθηνά

Κατσούρος Βασίλης, Διευθυντής Ερευνών του ΙΕΛ, Ερευνητικό Κέντρο Αθηνά

Θεόδωρος Δαλαμάγκας, Διευθυντής Ερευνών, Ερευνητικό Κέντρο Αθηνά

Ιανουάριος 2023

ABSTRACT

Climate change is one of the 17 global goals of the 2030 Agenda for Sustainable Development. Businesses in the Energy sector is a key factor responsible to help preventing climate crisis. In this work, we compile a textual collection of the Energy Industry in Greece with the aim of identifying their actions that are focused on the maintenance of a healthy and balanced ecosystem (e.g., renewable energy, reduce waste production). In addition, we also report on the automated acquisition of their published official reports regarding the Environmental, Social and Governance factors (ESG Reporting standards), if applicable.

SUBJECT AREA: Web-scraping

KEYWORDS: Web Scraping, Web Crawling, data engineering, Sustainability, Corporate Social Responsibility, CSR, ESG factors, environmental actions businesses, Greece, Energy, python

ΠΕΡΙΛΗΨΗ

Η κλιματική αλλαγή αποτελεί ένα από τους 17 στόχους σε παγκόσμιο επίπεδο της Ατζέντας για τη Βιώσιμη Ανάπτυξη 2030. Ο επιχειρηματικός κόσμος στον τομέα της ενέργειας είναι ένας παραγοντας κλειδί, υπεύθυνος να γίνει αρωγός στη προσπάθεια να εμποδίσουμε την κλιματική κρίση. Σε αυτή την εργασία συγκεντρώνουμε κείμενα από τη βιομηχανία της Ενέργειας στην Ελλάδα με σκοπό να εντοπίσουμε τις δράσεις τους που εστιάζουν στη διατήρηση ενός υγιούς και ισορροπημένου οικοσυστήματος (π.χ. Ανανεώσιμες πηγές ενέργειας, μείωση παραγωγής αποβλήτων). Επιπλέον, θα περιγράψουμε την αυτόματη εξαγωγή των επίσημων αναφορών των επιχειρήσεων οι οποίες αφορούν Περιβαλλοντικούς, Κοινωνικούς και Κυβερνητικούς παράγοντες (εκθέσεις ESG) όπου αυτό εφαρμόζεται από τις επιχειρήσεις.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Απόσπαση δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Web Scraping, Web Crawling, μηχανική δεδομένων, Βιωσιμότητα, εταιρική κοινωνική ευθύνη, δείκτες ESG, περιβαλλοντικές δράσεις επιχειρήσεις, Ελλάδα, Ενέργεια, python

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Papageorgiou Haris as well as Dr. Pappas Dimitris for guiding me and for sharing all their knowledge during the time we have been working together. I would also like to express my gratitude to my family and friends for the affection and encouragement they have given me during this dissertation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

1. INTRODUCTION

“Information is the oil of the 21st century, and analytics is the combustion engine.”
Peter Sondergaard

Nowadays, we are bombed with data derived from multiple areas of everyday life. This new resource gives us the opportunity to gain access and also handle information with respect to our purposes. In many cases information is already collected, structured and ready for use. But also, there are many other cases where interesting information can be found unstructured and scattered inside articles, audio files, websites e.t.c. In this work we will try to identify environmental actions of Greek businesses in the Energy Industry exploiting the public information they share at their websites. The method that is used to extract any kind of information in the sector of Information Technology is called Web scraping.

Web scraping is a technique where a application is set to extract data from public websites according to per case specified criteria. It simulates the behavior of a human user browsing a website by sending a request to its server (by its URL) and receives as response an HTML file. Having its data, its functionalities and all the architecture of the website, scrapper will be able to process the data, filter them and even use them to redirect to other websites as a human being (but way faster), a method also called *Web-crawling*.

1.1 What is the problem to examine?

Over the last 10 years, there is a clear imperative to redefine our activity related to environment to all sectors of society, starting from individuals, expanding to communities and society as a whole. United Nations Member States in 2015 adopted the 2030 Agenda for Sustainable Development that includes finding ways tackling climate change and preserve our oceans and forests. In this direction, planning teams have tried to deconstruct the problem by identifying its parameters and they have created many projects for this purpose.

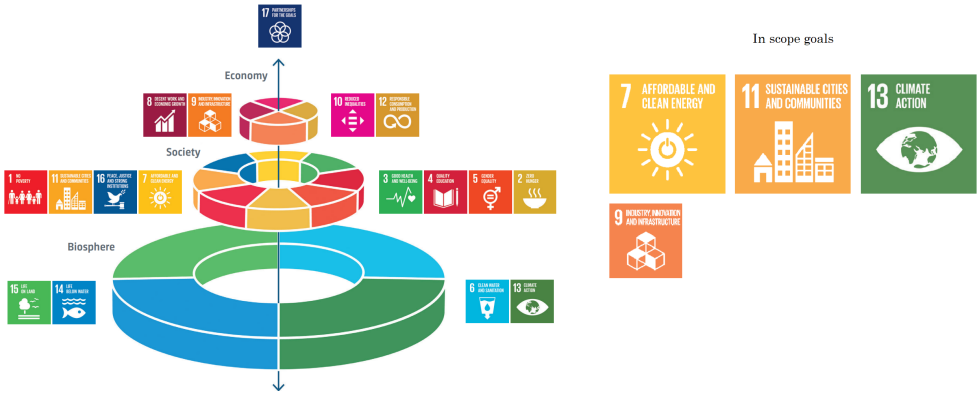

Figure 1: `Wedding cake model for the sustainable development goals`

Intelcomp is a project part of the EOSC initiative, involving multi-disciplinary teams to co-develop innovative analytics services, Natural Language Processing pipelines and Artificial Intelligence workflows. By exploiting open data, services and computational resources from the EOSC, HPC environments and federated distributed operations at the European Union, national and regional level this project will provide tools for assisting the whole spectrum of policy [1]. Data are multilingual and heterogeneous so analytics tools should be developed to each field accordingly. In this work the field of interest that we will examine will be focused on the Greek's energy industries to collect data about their environmental actions & responsibility. Specifically we will use the websites of the energy industries in Greece to take this information (if provided) as well as their ESG annual reports. These reports are supposed to be mandatory published after 01/01/2024 as European Union political agreement on the corporate sustainability reporting directive (CSRD) [1] .

[1] consilium.europa.eu - E.U. political agreement

1.2 What is ESG principles?

ESG (environmental, social, and corporate governance) is a framework designed to be embedded into an organization's strategy that considers the needs and ways in which to generate value for all of organizational stakeholders (such as employees, customers and suppliers and financiers).

ESG corporate reporting can be used by stakeholders to assess the material sustainability-related risks and opportunities relevant to an organization. Investors may also use ESG data beyond assessing material risks to the organization in their evaluation of enterprise value, specifically by designing models based on assumptions that the identification, assessment and management of sustainability-related risks and opportunities in respect to all organizational stakeholders leads to higher long-term risk-adjusted return. Organizational stakeholders include but not limited to customers, suppliers, employees, leadership, and the environment [2]

- **Environmental**

  Environmental factors refer to an organization's environmental impact(s) and risk management practices. These include direct and indirect greenhouse gas emissions, management's stewardship over natural resources, and the firm's overall resiliency against physical climate risks (like climate change, flooding, and fires). [2]

- **Social**

  The social pillar refers to an organization's relationships with stakeholders. Examples of factors that a firm may be measured against include Human Capital Management (HCM) metrics (like fair wages and employee engagement) but also an organization's impact on the communities in which it operates.

  A hallmark of ESG is how social impact expectations have extended outside the walls of the company and to supply chain partners, particularly those in developing economies where environmental and labor standards may be less robust [2].

- **Governance**

  Corporate governance refers to how an organization is led and managed. ESG analysts will seek to understand better how leadership's incentives are aligned with stakeholder expectations, how shareholder rights are viewed and honored, and what types of internal controls exist to promote transparency and accountability on the part

of leadership [2].

1.3 Web scraping challenges

General web-scraping challenges we should bear in mind are the following [3]:

- **Authorizations**

  Many web pages have ways to identify whether there is a bot that is performing the call or a human and as a result they return ”Authentication error” as response.

  - **CAPTCHA**

    CAPTCHA will not be a limitation at this work as there is no need to enter login information to access public websites.

- **Speed**

  Crawlers scan large amounts of non relevant information which makes tools slow down their performance

- **Data quality (& quantity)**

  It is needed to be very careful on the filters that will be applied as there is always the risk to store large amount of unnecessary data that will be hard to handle. On the other side we should be careful not to be too strict and consequently filters will exclude important finding from results. The tolerance in the filters is a process that is hard to be automated for generic purposes at the moment.

- **Security**

  There are types of files on the web such as bash scripts, .exe files and other potential malware that can harm the computer

2. RELATED WORK

2.1 Software solutions for crawling

There are open source tools as well as commercial products that can serve web scraping & web crawling purposes.

2.1.1 Open source

Crawlers have been built in plenty of languages: Python, Java, C#, JavaScript, PHP, C++, C, Ruby, Rust, R, Erlang, Perl, Go, Scala and more [4]. Among the most famous are the following:

- **Scrapy** (Python)
  Scrapy is a fast high-level web crawling and web scraping framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing. It also offers cloud environment to run the scrapers [5].

- **pyspider** (Python)
  Pyspider is a package which apart from web crawling functionality, provides a framework servicing script editor, task monitor, project manager, and result viewer [6].

- **Webmagic** (Java)
  Webmagic is a framework that can covers the whole lifecycle of crawler: downloading, url management, content extraction and persistent [7].

- **Node Crawler** (Node.js)
  Node Crawler offers automatic insertion to jQuery, Prioritization and retrial mechanism [8].

- **Beautiful Soup** (Python)
  Beautiful Soup is a library used for parsing HTML and XML documents. After creating a parse tree, extracting data from the web is much easier. This is the tool we will use in this work. [9].

2.1.2 Commercial products

There are many companies offering products for specific purposes. Some examples of companies offering web-scraping solutions as a product are:

- Zyte which can offer web data extraction services, mechanisms to avoid being blocked as well as hosting of scraping tools in the Cloud.
- DiffBot provides multiple API options for extracting web data, including data about organizations, data related to retail products and data from news content and articles. The web crawler lets you turn sites into databases of information and machine-readable data into human-readable data automatically [11].
- ScraperAPI is a web service that extracts data from websites. You can use ScraperAPI with the shell interface, like Bash and Node, using the GET request, or with programming languages, including Python, PHP, Ruby and Java. The API can gather the raw HTML data, including content in browsers, CAPTCHAs and proxies. The design of ScraperAPI has customizing features to integrate into scrapers [11].

2.2 Dictionary for the Electric Power Industry

During our work, the need for a dictionary related to renewable energy & sustainability occurred. In the process of trying to find dictionaries in energy sector collected from the linguistic community failed, it was discovered that a similar task was performed by the company `https://www.epri.com/` containing about 900 words and phrases that can be used by NLP tools to understand the language in similar documents and process them for answers. The dictionary is not shared publicly [12].

Furthermore, a well-structured work has been performed by a team at Georgia Institute of Technology. The team extracted 45,595 key phrases from the sustainability documents and augmented them by collecting 3,798,530 topics from Wikipedia under the categories related to sustainability, extracting all unigrams (one-), bigrams (two-), and trigrams (three-word combinations) from each phrase. After automatically excluding the phrases are not used by business world in CSR corpus & following manual checking of 14,037 phrases, the team ended up with 614 *environmental* phrases [10]. The dictionary in English though

is also not publicly shared. For the time being we cannot try to apply the same models in Greek reports as we miss the data (reports) that our models need to learn from.

3. PROJECT PIPELINE

As we mentioned in the chapters above, our main goal is scanning websites, to extract:

1. Information about Greek businesses environmental activity
2. ESG Reports

In order to reach this goal we need:

- the list of all the Greek businesses that belong to the energy factor with their domains
- a software application that will look over all the HTML of each and every domain in our list as well as all the URLs appearing on the web page and its subdomains & will extract their content
- an extra tool that is able to remove HTML syntax and keep only the text
- Invent the way to apply filtering according to our points of interest

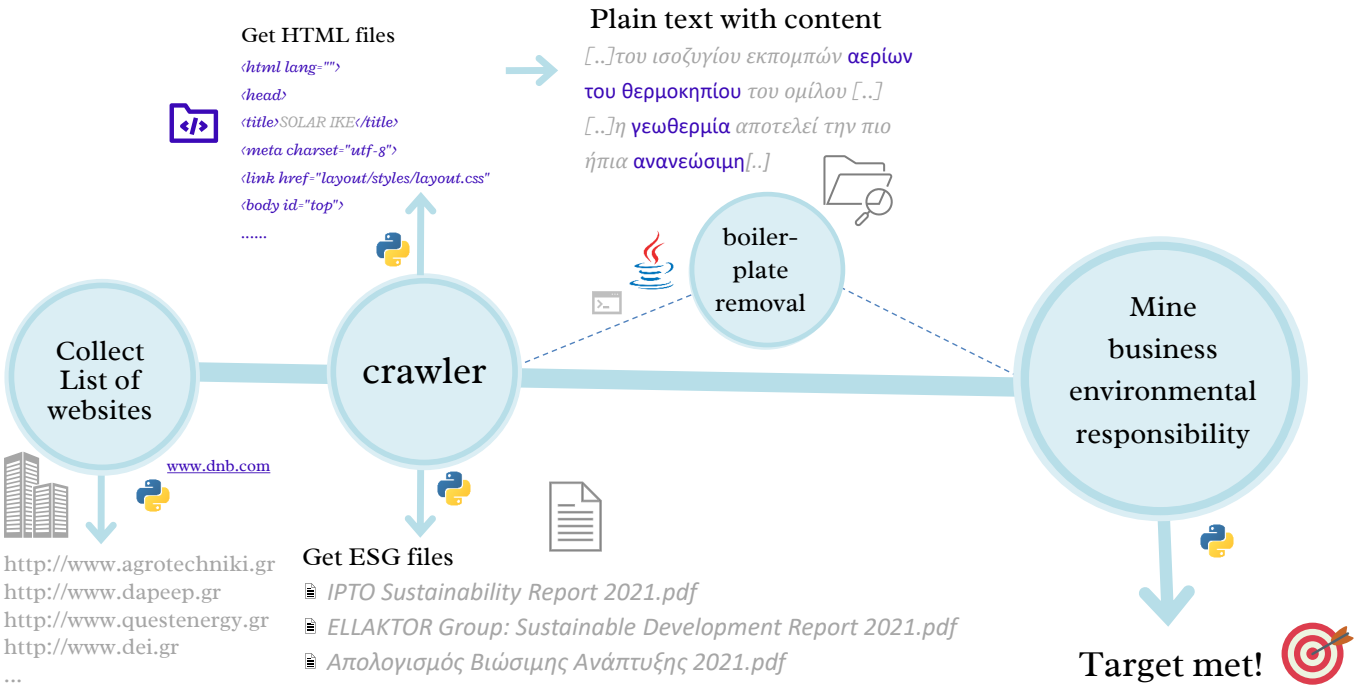

Figure 2: `Project Pipeline`

4. IMPLEMENTATION STEPS

4.1 Step 1: Find the domains of Greek Energy industry businesses

The first obstacle we need to overcome is to find a way to gather all together the names & the domains for all the Greek businesses participating in the Energy industry as there is no government's official data source which can provide this information. There is though an evident need to gather this information (even indirectly) for multiple purposes so that task can be performed in the private sector. A data company, under the name "Dun & Bradstreet's"[1] is carrying out this task & stores business data in order to consult businesses with their strategy. Data are available at company's web page and can be filtered with region & Industry, the two parameters that serve our needs[2]. The way to extract the data from the web page is using web crawling techniques and storing the key information we will need (domain, company name) in a CSV file.

Crawler

For this purpose a generic website crawler created by ATHENA R.C. is used. The functionality is simple: Given a website it collects all HTML data from the domain. The crawler operates on a Breadth-First-Search manner and stops after a specific number of crawled pages. Running the crawler we generate a list with the domains of 386 companies (list available here). The file has the following structure:

Source Page	Domain	Output folder
dnb_climate_greece	http://www.agrotechniki.gr	0
dnb_climate_greece	http://www.upsolar.com	1
dnb_climate_greece	http://www.deddie.gr	2

Table 1: Input JSON file format

Data validity: Accurate - Complete - Updated

The company itself has founded Quality Assurance (DUNSRight™) which includes over 2,000 separate automated checks — plus many manual ones — to ensure the data meet quality standards and moreover verifies data from thousands of sources on a daily basis[3].

[1] DnB company - about

[2] Electric Power Generation, Transmission And Distribution Companies In Greece

[3] Data quality of DnB

4.2 Step 2: Crawl web pages of all listed domains

4.2.1 General Idea

In this step we need to look into web pages content. In order to handle that information we need to export locally all the HTML code of each web page and some PDF files (of specific content) as well. For that purpose we need a tool able to accept as input multiple websites and export their contents in a structured way that we can handle in further steps. The tool will also be able to navigate to all desired “HTML links” (also known as “a href links”). The process of the tool that is used is described in the section below.

4.2.2 Crawler functionality

User’s Input parameters:

- The path of the input file *xyz.json*. The file has the structure of Table: 1
- N = Maximum pages to visit at each domain (optional)
- The output path where results will be exported

4.2.2.1 Export HTML mode

Output:

Results:

The program generates 3 categories of files:

- a text file which contains all the URLs that were visited
- an HTML file with the source code of the web page
- a text file returning a dictionary with the languages that were used in the web page and their frequency (in paragraphs)

What to ignore?

We need to avoid working to anything redundant in order to be more purposeful. In this task our goal is to get only web page text content; files are not in scope, neither visits to any of the standard social platforms. For this reason our algorithm will avoid:

- searching URLs having the following name extensions:
  .pdf, .3ds, .tif, .eps, .dwg, .odt, .bmp, .xml, .png, .svg, .jpg, .jpeg, .gif, .mp3, .mp4, .doc, .docx, .csv, .xls, .xlsx, .zip, .txt, .rar, .7z, .tar, .gz, .sh, .exe, .dmg, .gpg, .afdu, .psd
- common keywords redirecting websites (e.g. 'twitter', 'google', 'facebook', " ,'mailto:')

Data Storage

The results are stored in a file system; the folders are constructed using the following hierarchy:

Parent folder: Name of the *Source page* set in input file,

Middle folder: Name of the *Output folder* set in input file

Results folder: A *serial number* generated for each URL was called during the navigation of the main domain.

For example, line 1 of Table: 1 will extract the set of files on path:

`C:\Users\...\dnb_climate_greece\0\i`

with i varies from 1 to *N*: number pages to visit inside a domain.

We should not disregard the fact that many web pages have techniques to block web crawling so it is important to keep track of successes & failures too:

Log files:

At the *Output folder* level there is the set of log files containing:

- Pages actually visited
- Pages Discarded (out of the region of interest)
- Pages returned error

4.2.2.2 Export PDF mode

Apart from plain text, we need to find and extract some official documents that some companies have published on their web site. Specifically, we need to download ESG reports when available (more information can be found in Section 1.2). Our crawler is able to perform this task applying the first filter: the file format. It can easily differentiate *Hypertext*

REFerence that is a PDF file among all other URLs. The task that we need to tackle now is to identify whether this file refers to company's activity regarding environmental sustainability.

Reading PDF files using code to ascertain their content is a time consuming process that exceeds a realistic available time margin. Companies publish many detailed financial statements that would procrastinate our process of work. So we need to find a way to filter what documents we should export in an indirect way.

How to filter ESG documents out of all PDF?

The only parameter for the filtering is text content. We will follow 2 ways to isolate PDF files according to a level of confidence that we will define here:

Therefore the idea is to build a (customizing) dictionary with keywords that will be used as event flags; when any of the keywords is found in the current HTML file, then the program will download any PDF that is found at current page (if any exists). For the purpose of our project, we gathered a set of keywords that are specific enough to ensure we are targeting the correct content. (The keywords are mentioned at Table 8 on page 39).

Filters & grades of tolerance

To make the filter less tolerate with false positive results, we demand keywords to also appear in URL itself (e.g. www.terna-energy.com/...**CSR**_Report...pdf) and store all other found in a list for further review. In this project this restriction matches our target without ignoring in scope files. That is because the keywords have great semantic distance from all other words used in the website.

The resulting name of the file will be composed using the keyword found and the domain (e.g. **CSR**_terna-energy_7.pdf). The described logic is represented in the algorithm below:

4.2.3 Crawler: The algorithm

Algorithm 1 Crawler using Breadth-first search

```
pdfModeOn ← True
HTMLModeOn ← True
for sourcepage, domain in list: do
    Create log files
    UrlsToVisit ← domain
    keywordInHtml ← False
    keywordInUrl ← False
    for Url in UrlsToVisit: do
        Store Url.response                                   ▷ get HTML
        check if keywordInHtml ← True                        ▷ condition 1
        check if keywordInUrl ← True                         ▷ condition 2
        Download PDF files where condition 1 and 2 are met   ▷ get PDF
        Store links containing PDF where condition 1 is met but not 2
        UrlsToVisit ← collect all <a> HTML tags to continue searching
        enhance list UrlsToVisit
    end for
end for
```

4.3 Step 3: Boilerplate removal

4.3.1 General Idea

In this work our focus is on the text contents of the web pages appearing in our list. Web pages (and thus their source code), include much more than plain text: advertisements, pictures, link lists, banners, navigation and more. We need to discard this noise from our HTML files for 3 main reasons:

- To enhance the performance searching less ammount of data
- To improve accuracy on results
- To make it possible to apply analysis (e.g. count of words) & process further information derived from text

The technique that is used to achieve that is called in the literature as “boilerplate removal”, “Web page segmentation” or“content extraction”. The core idea behind is that the the algorithm identifies repeated source of code and discards it.

4.3.2 Using Boilerplate removal tool

We are provided with a JAR file performing the task of boilerplate removal. The way to execute it is to run the command that takes as input the directory containing the HTML file and as output parameter the directory that will store the *clean website* in a .txt file format .

```
java -jar ilsp-boilerpipe-1.9.1-jar-with-dependencies.jar -id .\input_dir\ -od .\output_dir\
```

The obstacle here is that we need to run the tool for all the websites of our list. To do so we bring all HTML files to the same folder (per domain) (CMD command) and we use a simple python script for every website folder (created in Section 4.2) that will:

- create the corresponding directories `output_dir`
- generate corresponding commands to create a bat file with all commands

```
java -jar ilsp-boilerpipe-1.9.1-jar-with-dependencies.jar -id out\0 -od clean_websites\0
java -jar ilsp-boilerpipe-1.9.1-jar-with-dependencies.jar -id out\1 -od clean_websites\1
java -jar ilsp-boilerpipe-1.9.1-jar-with-dependencies.jar -id out\10 -od clean_websites\10
...
```

and we run all in once. The result of the process is having the plain text of all HTML files stored in structured folders per web site. The total size of all websites after boilerplate removal is 51.4 MB instead of 4.60 GB of the initial HTML files memory occupation. It would be helpful to take as example a random file for the web page `https://www.enexgroup.gr/en/pcr` and ascertain their difference in data load:
- *Initial HTML file* occupies 85 kB of memory
- *New file* after boiler-plate removal is occupying only 6 kB.

4.4 Step 4: Mine business environmental responsibility

4.4.1 General idea

We have finished the steps of preprocessing and it is the time to ”look” the data to get our answers. We have all the plain text of websites of businesses belonging to the energy sector in Greece stored in files.
The tasks are the following:

- Find all products or services invented/created by the company. Technically speaking we need to find text that includes: copyright ©, trademark: ™, Service mark ^{SM}, Registered trademark ®.
- Find actions related to environmental sustainability
- Log success score and metadata (tokens per page with & without content in scope)

Column	**example**
Folder	180
Failed entirely Boole	0
Number of Succeeded Pages	79
Succeeded pages perc	100%
Avg tokens per page	55789
Avg tokens (with content) per page	37935
Trademarks	singsong.gr © all rights reserved
Renewable energy activity	greenhouse effect

Table 2: CSV results file structure

4.4.2 Columns data explanation

The data of the columns of the exported file are described below:

1. **Folder:** correspond to a specific domain. Mapping is found in the reference file.
2. **Failure Boole:** is a flag that takes the value ’1’ if crawler was refused access entirely from the website (for the main URL and all the embedded hyperlinks as well). Otherwise takes the value ’0’.
3. **Number of Succeeded Pages:** is a counter that indicates how many hyperlinks where visited successfully without returning any error.
4. **Succeeded pages perc:** This metric returns the percentage of hyperlinks that algorithm crawled information among all hyperlinks visited.
5. **Avg tokens per page:** It is a counter for all words (tokenized) were found in every web page. This can be used as a metric on how enriched - famous is a web page.
6. **Avg tokens (with content) per page:** It is a counter for all words (tokenized) were found in every web page, ignoring stopwords.
7. **Trademarks:** Algorithm tracks sentences* containing trademarks mentioned in General idea.
8. **Renewable energy activity:** Algorithm tracks *part of the sentence* containing words that belong to Renewable energy dictionary**. (Here we set algorithm to get a region of 5 words right & 5 words left with the significant word in the center).

* Sentences are tokenized using *sent_tokenize* from *NLTK* library.

** We constructed a dictionary including all words (in Greek) related to Renewable energy. (Table 7 on page 39)

4.4.3 HTML distillation : The algorithm

Algorithm 2 Distill information of interest from HTML files

Load path with HTML files

for $domain$ in $path$: **do**

find pages returning error

Count tokens

find Sentences with trademarks

find Sentences referring to sustainability

end for

Store results in a CSV file

5. RESULTS

We have 2 branches of results of expected results:

1. Information from HTML texts about Greek businesses environmental activity
2. ESG Reports

5.1 Information obtained from HTML texts

Some statistics from the results are mentioned below (results are distilled in file) :

- **Failure Boole:**
  Almost 50% of the websites (181/386) rejected the call of the crawler (with Authentication error). Among rejected websites are included:
  `http://www.dei.gr`, `http://www.volterra.gr`.
  Considering performing a sampling check to the failed web pages, many of them are small business that do not provide adequate source of information on their web site (web pages under construction, with quite few data or even not working).

- **Number of Succeeded Pages:**
  Higher scores
  The highest score holds an outlier on the list `http://www.unigreen.gr`, that is a retail company and not an energy company (it is bad data and is rejected).
  The next most enriched in embedded URLs web pages are shown in the table below together with their scores:

Web site	Pages called
`http://www.admie.gr`	5321
`http://www.ellaktor.com`	1177
`http://www.bitros.gr`	1023
`http://www.elpedison.gr`	804

Table 3: Web pages called inside a single domain

Meanwhile the 57% of the succeeded pages (116 out of 205) called only 1 or 2

URLs. That is an indicator that many of the businesses have insufficient information digitally.

- **Succeeded pages percentage:**

  For the succeeded calls, almost all web pages (96% - 197 out of 205) visited *all* embedded hyperlink that were found (*all* meaning scored >90% in this metric).

- **Avg tokens per page:**

  Most of the web pages contain 161 tokens. There was an outlier found `http://www.heron.gr` scoring 4981 tokens per page. The outlier nevertheless is quite valid; has not eventuated from any technical discrepancy. Checking the data of the biggest files, is confirmed that they constitute plain text (and is not for example a corrupted file that misinform us with ”imaginary” words).

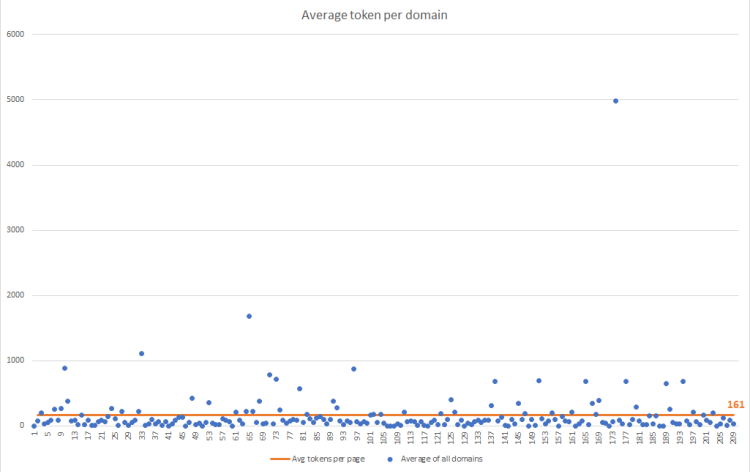

Figure 3: `Average tokens per domain`

- **Avg tokens (with content) per page:**

  This metric has of course the same shape with the distribution above. The average of all the web pages ignoring stop words is 126.

- **Trademarks:**

  In this section our goal was to find innovative new products, patents or innovative services produced by the industrial market. The crawler did not find any trademark, copyright, service or registered mark other than the copyrights of the the company itself. A sample of the results is shown below:

Web site	Trademarks found
http://www.elperes.gr	copyright © elpe renewables 2022 - web
http://www.w-servouniou.gr	τερνα ενεργειακη © 2013 energeiaki servouniou s.a
http://www.lithosaiolos.gr	© lithos aiolos ενεργειακη α.ε.
http://www.ellinikiaioliki-sa.gr	© ελληνική αιολική ενεργειακή α.ε.
http://www.veroia1.ilg.gr	φωτοβολταικα παρκα βεροια ι αε © 2013

Table 4: Trademarks found

The total number of webpages detected with trademarks is 40.

- **Renewable energy activity:**

  Only 13 companies had references on renewable energy. Some parts of the texts of the results are shown below:

Web site	Text: region of interest
http://www.admie.gr	ηλεκτρικής ενέργειας και αναπτύσσονται στους εξής άξονες: τεχνολογίες μηδενικών ή χαμηλών εκπομπών αερίων του θερμοκηπίου για την αντιμετώπιση της κλιματικής αλλαγής παραγωγή καθαρής και φθηνής ενέργειας
http://www.ellaktor.com	το 2020 , οι εκπομπές αερίων φαινομένου του θερμοκηπίου (ghg) του ομίλου υπολογίζονται σε 97.911 tn co2 eq. , εκ των οποίων
http://www.elpedison.gr	θα βρείτε πληροφορίες που αφορούν στην παραγωγή ενέργειας από βιομάζα , γεωθερμία , υδροηλεκτρικά , αιολικά και φωτοβολταϊκά συστήματα
http://www.wattcrop.com	έργων που υποστηρίζει τη μετάβαση σε ένα κλιματικά ουδέτερο μέλλον με μηδενικές εκπομπές αερίων του θερμοκηπίου
http://www.tharros-energy.com	συνολικό ισοζύγιο διοξειδίου του άνθρακα είναι μηδενικό και η βιομάζα δε συμβάλλει στο φαινόμενο του θερμοκηπίου
http://www.soumpasis-solar.gr	οι εκπομπές διοξειδίου του άνθρακα πυροδοτούν το φαινόμενο του θερμοκηπίου και αλλάζουν το κλίμα της γης, ενώ η ατμοσφαιρική ρύπανση έχει σοβαρές επιπτώσεις

http://www.volton.gr	αφού έχει αέρια μορφή και λόγω των μειωμένων εκπομπών ρύπων επιβαρύνει λιγότερο το φαινόμενο του θερμοκηπίου
http://www.elperes.gr	ώστε να συμβάλει στην εξισορρόπηση του ισοζυγίου εκπομπών αερίων του θερμοκηπίου με μείωση του αποτυπώματος άνθρακα του ομίλου ελληνικα πετρελαια τουλάχιστον κατά 250.000 τόνους ετησίως
http://www.watt-volt.gr	στη μείωση των εκπομπών αερίων θερμοκηπίου από τις μεταφορές κατά 90
http://www.wonderplant.gr	wonderplant διακρίνεται για τον πολυδιάστατο, επιστημονικό χαρακτήρα της και περιλαμβάνει : λειτουργία υπερσύγχρονου γυάλινου θερμοκηπίου με σύστημα φυσικού δροσισμού pad & fan για απόδοση και στους θερινούς μήνες
http://www.fgrid.com	η γεωθερμία αποτελεί την πιο ήπια ανανεώσιμη πηγή ενέργειας (απε) με τον υψηλότερο συντελεστή χρήσης/λειτουργίας
http://www.beal.gr	διοξείδιο του άνθρακα (co2) , όσον αφορά στην επίδρασή του στο φαινόμενο του θερμοκηπίου
http://www.heliotop.gr	αποφεύγεται η χρήση συμβατικών ενεργειακών πόρων με αποτέλεσμα την μείωση του φαινομένου του θερμοκηπίου το βιοαέριο από την αναερόβια χώνευση δεν βελτιώνει μόνο το ενεργειακό ισοζύγιο της χώρας

Table 5: Text on web pages referring to sustainability

5.2 ESG reports

The last but mail goal is to extract ESG reports. We manage to find 47 files referring to sustainability in Greek businesses belonging to the following 9 companies:

Web sites with ESG
http://www.ellaktor.com
http://www.terna-energy.com
http://www.green.com.gr
http://www.eydap.gr
http://www.terna-energy.gr
http://www.gen-i.eu
http://www.alpiq.com
http://www.helpe.gr
http://www.admie.gr

Table 6: Companies providing ESG files

The files can be found in the folder on the link.

Note: Multiplicity of ESG files for a single company may occur because ESG reports have been published for different years.

Keeping Track of ignored URLs

We are acknowledged that the way we used on filtering PDF files found in the websites to get ESG files is not well-defined. To improve the scores on false-positive results and to re-evaluate the results, we have saved in a list all the URLs that were found in a web page with the event of an ”ESG-related” term (e.g. Sustainability) but did not satisfy the (stricter) condition to include a keyword in the URL. This file consists a very restricted range of the total information that can be exploited further. It includes many true-negative PDF, but also some false-negative cases that we did not manage to include in the automated process and is a very condensed source for the next step to improve our tool. The ignored URLs (almost 47,000 in total) can be found in this link. We named ”tolerance” in section *Export PDF mode* as a qualitative parameter to express how strict our criteria will be in order to get a file or not.

Using clustering on the paths of the URLs we can have an insight on what to omit, what to keep and what needs more investigation.

5.3 Challenges faced

- Some URLs are constructed dynamically inside HTML. In order to get these PDF files you need to find and concatenate the origin-domain in order to find the correct path.
- There are files that are not in PDF format such as `https://www.quest.gr/flipbook/khd.html` that look like a digital brochure and cannot be crawled.
- ESG results may appear also in tables in HTML itself, and not in a separate file, for example: Deddie: deiktes_esg, Dei: deiktes_esg
  These tables are not structured strictly as tables in the HTML file and it is needed different logic to each one to identify the form of a table in order to use code and download it.
- We accept that the tool will not include other types of information that can be also found in websites (e.g. like videos).
- Code can fall into exception that was not predictable; websites structure is evolving continuously
- Authorizations errors is a nonnegotiable obstacle
- New technologies evolving the Web development branch may quickly make obsolete web scraping tools.

6. FUTURE WORK

6.1 Automate vocabulary extraction for every kind of task

Software tools are better when they self-adapt at as many tasks as possible without any human intervention. In this work, human critical thinking was a necessary part of the process that was serving only one specific field: environmental sustainability of Businesses in Greece. We created a dictionary for Renewable Energy that is:

- language-depended
- domain-depend
- unable to apply statistical analysis to get some metrics as it is generated approximately from human experience and not from mathematical analysis tools

Nowadays we need to make machines think for humans. Using NLP tools we can avoid the error margin & variety limits that human critical thinking cannot annihilate. Starting from raw text corpus we want to get text representation, following the next process:
raw text corpus → processed text → tokenized text → corpus vocabulary → text representation.
As we mention in section 2.2, building a dictionary learned from relative sources is the ideal source that will be used as filter for extractions under similar topics.

6.2 Extract tables

As it is mention in challenges section 5.3, extracting data from tables was not an easy task, especially when tables are not strictly structured. This is an important loss for a tool that focuses on quantitative data like ESG reporting. So it is a prioritized need to overcome the various layouts and find ways to store the information accordingly.

6.3 Integrate all components of pipeline in one tool

User-experience of a software tool is better when there are few things to run. Integrating the task in one tool would be much easier for the end-user that is not quote familiar with running scripts.

7. CONCLUSIONS

There is a lot of information available in the world of web about sectors of society we are part of. Building the right tools to extract the information of our interest and using statistics to understand & analyze them, give us the access to a more accurate aspect of reality. Data science can also build the right filters to focus the tool's attention in the specific task of our interest. Impediment on this task is the processing time limit if there is a big amount of data are as well as security protection operations.

In the USA around 90% of S&P 500 firms published a sustainability report in 2020 (while only 20% published in 2011) [10]. In Greece S&P 500 is not have a well defined index to use as metric as we do not know companies which are included so that we can calculate the corresponding statistic. Empirically though we ascertain that only the leading companies in the sector of energy have published official reports while the first strict deadline for publishing the reports is set on 1/1/2024 (concerning only big companies).

A. DICTIONARIES

Table 7: Renewable Energy dictionary [EL]

Words	Words
Ανεμογεννήτρ	θαλάσσι
Ανανεώσιμ	Βιομάζα
πηγές ενέργειας	Γεωθερμική
ΑΠΕ	γεωθερμία
ήπιες μορφές	βιολογικ
μορφές ενέργειας	θερμοκηπίου
νέες πηγές ενέργειας	θερμοκήπ
πράσινη ενέργεια	φωτοβολτα
ΚΑΠΕ	υδροηλεκτρικ
Εξοικονόμησης Ενέργειας	υδροηλεκτρισμός
Ηλιακή	Περιβαλλοντικών
cres.gr	βιοκαύσιμο
οικολογικ	Βιομάζα
υδροηλεκτρισμ	Ωσμωτική
παλιρροϊκ	Υδραυλική ενέργεια
Αιολική	φιλικές προς το περιβάλλον

Table 8: ESG pdf files dictionary

keywords	keywords
ESG	viosimi
ΒΙΩΣΙΜΗ ΑΝΑΠΤΥΞΗ	Environmental Social Governance
ΒΙΩΣΙΜΗ	Sustainable
ΒΙΩΣΙΜΌΤΗΤΑ	Sustainability
viosim	CSR

REFERENCES

[1] Intelcomp.eu. Intelcomp - about.

[2] corporatefinanceinstitute.com. Esg (environmental, social and governance) explained.

[3] Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 1st edition, 2015.

[4] BruceDone. Open source crawlers.

[5] Scrapy Team. Scrapy.

[6] Roy Binux. Pyspider.

[7] Yihua Huang. Webmagic.

[8] BDA-Research. Node crawler.

[9] Leonard Richardson. Beautiful soup.

[10] Sudheer Chava, Wendi Du, and Baridhi Malakar. Do managers walk the talk on environmental and social issues? *Georgia Tech Scheller College of Business Research Paper*, (3900814), 2021.

[11] indeed.com. Web scraping tools.

[12] Jeremy Renshaw Bhavin Desai Lea Boche Yashwant Jankay Carola Gregorich, Chris Wiegand. epri-journal.com.