Linkero

A Brand Monitoring Platform

This report is submitted in partial fulfilment of the requirements

for the

B.Sc. (Honours) Information Systems and Information Technology (DT249)

to the

School of Computing
College of Sciences & Health
Dublin Institute of Technology

Author: Stefano Richiardi
Student ID: D12124118
Supervisor:
Date: January 2019

Abstract

In this project I aim to build a data aggregator service for investigations and monitoring of counterfeit products online. The service will be implemented with a web interface for clients' access, an http server and specialized databases depending on data and performance requirements.

The core mission of this project is to try to simplify daily tasks of conterfeit investigations, and leverage server technologies to allow investigators to see the bigger picture painted when data from one case connect to one or multiple other cases previously unconnected.

Equally important will be the ability to deliver a product that is ready for production environment, that is: running on a virtual server accessible from the internet, protected from unauthorized intrusions and easily scalable should the user population grow significantly.

In this project I will also discuss pros and cons of data collection from online sources: from open API offered by onlines services to the community of developers, to unstructured data that can be scraped and schematized, and which legal implications need to be taken into account in light of the new European General Data Protection Regulation (aka GDPR).

Contents

Chapter 1

Introduction

1.1 Project background

The European Union Intellectual Property Office, in coolaboration with the OECD, estimated that in 2013 the value of imported counterfeit good was €338 billion, which corresponded to 2.5% of the total imports in the world trade [OECD(2016)]. The report also calculates that 5% of imported good in EU are counterfeit worth up to €85 billion. Counterfeit is an unfair commercial practice which takes advantage of the brand identity and presence on the market built by the brand owner, without incurring in the same costs of brand development (product design, quality standards and marketing). Other than a financial damage for the original brands, counterfeit products may pose safety risks for end consumers as much as for the people that work to produce them, since they evade the strict quality and safety standards set by national and international agencies around the world. From counterfeit iphone batteries that explode to counterfeit air-bags that do not trigger, they all can cause physical damage to consumers.

E-commerce has been rapidly expanding since it's early days in mid '90 when online marketplaces like eBay and Amazon were launched. Although it is hard to make a global estimate of market share for e-commerce companies, if we take the US market, the Census Bureau estimated that in 2009 online sales accounted for 4% of all retail sales, whereas in 2017 the e-commerce market share went up to 9% [of Commerce(2018)]. As legitimate retailers increase their presence online, shops selling counterfeit follow suit.

It is not surprising that many companies started to invest in fighting this trend, and that a good portion of the battleground is online. This tool is designed to help users to extract data related to a specific brands from online platforms such as eBay,

Mercadolibre, Allegro, and present it in a tabular format.

Online marketplaces offer API open to developers in order to enable automated interaction with their platforms. This tool will focus on extracting sales and business registration data. The data will be stored in a database so that users can keep historical records of all their queries. At the same time, users will be able to leverage the growing dataset to link new investigations to old cases across all platforms whose data has already been stored in the database. For instance, serching the email address of a shop that deals counterfeit Diesel jeans may return details of multiple businesses registered on eBay and Mercadolibre at different times, as well as information about the administrator of a facebook page about counterfeit Raiban glasses. Users can use this tool to estimate brand damage caused based on the volume of sales, as well as a forensics platform to facilitate identity attribution of potential counterfeiters.

1.2 Background

Identifying online counterfeit items starts with reasearches of a brand or product online presence. The aim is to identify sellers that offer a branded product sold at a price point below average retail price. An investigator would ideally prioritize sellers with higher business turnover, and ideally located in jurisdictions where legal action is a dependable and impartial option. For instance countries like Russia and China, to name few, do not always offer adequate protection for European and North American companies, making it more difficul to pursue compensation from actors located withing their borders.

There are already different online companies that offer brand monitoring services: they range from keyword web-crawlers, to consumer sentiment analytics based on social platforms, to anti-counterfeit detection. We will review those companies and their services in details in the next chapter.

GDPR is the latest Europen Union legislation about processing of personal identifiable, privacy protected data belonging to european citizens. GDPR affects the implementation of this project in two ways: first, and most obvious, because users will submit limited personal data in order to be able to access Linkero services. Something as simple as an email address required to get notification for password recovery or data collection completed confirmation, is considered a PII (personal identifiable information) and thus is protected under GDPR. But GDPR affects also the collection of public personal details available online. This is the compliance requirement that is specific to companies providing *web scraping* services. When we deal with open source intelligence and attribution, PII are the most valuable data for anti-counterfeit investi-

gations, therefore we will talk about how GDPR affects services like Linkero and what are the requirements to guarantee compliance.

1.3 Objectives

The objectives set for this project fall in two main categories: business and personal.

In relation to the first type, the goal is to build a business-ready web platform for brand-monitoring investigations. Being *business ready* means presenting a final application that runs on a virtual server directly connected to the public internet; that implements all industry starndard security features to guarantee that only authorized users can access its features and data; and is scalable, designed to accomodate an exponentially gowing number of users. The other aspect of the first goal refers to the ability to provide basic case management, data extraction and keyword searches capabilities tailerd for brand monitoring and anti-counterfeit investigations.

On a personal level, I aim at gaining a hands-on understanding of specific web technologies (e.g. JQuery, Django framework, NGINX, uWsgi), NoSQL databases (MongoDB, Redis), ansychronous and non-blocking programming techniques (multi-threading, AJAX, Pjthon Celery), development methodologies and testing practices.

1.4 Scope

This project will consist of a ready to deploy and use system for case management and web scraping data using eBay public API.

Regular users will be able to login, change password, set their preferred email address for general notifications and report delivery, launch queries, download data from completed reports, delete old queries, and perform keyword searches within all data collected and stored by any user in the database. Also the system will be configured to provide security protection against external unauthorized access and use of the system. Site administrators will manage the registration, password reset and deletion of users profiles.

1.5 Challenges and learning requirements

There are technical and personal challenges that have some kind of impact on the delivery of this project. To start with I am new to most of the technology used: from the basic configuration and administration of a Unix server, NoSql databases, client side web technology, SSL and DKIM protocols, etc.... This means that part of the

time dedicated to the project will be spent bringing my knowledge and familiarity to a level that suits the project's requirements. Another element of technical challenge is determined by the use of a virtual server with only 1GB of memory and 1 single CPU. These limits are mainly dictated by financial reasons: this is the cheapest available VPS, which comes at $5.00 USD a month. Considering that I needed a VPS for the entire development and testing of the project, which took almost 2 years, anything more expensive would have impancted my family finances. The other problem is that limited memory and processing power started to show their impact and the code had to optimized accordingly.

On a presonal level I will have to take into account all the extra pressure that comes from being a father in a young family and professional responsibilities of a full time job.

1.6 Deliverables

The completed project will include:

- full documentation of the system design, development and functions;
- a working implementation of the system;
- a PowerPoint presentation of the project.

Chapter 2

Literature review

2.1 Introduction

This chapter will provide an overview of different topics that informed design decisions made during the initial stages of the project. We will start looking at the discipline of Open Source Intelligence, which refers to all those protocols and techniques used by government and private agencies to piece together intelligence reports using publicly available sources, since Linkero is a tool that facilitates the structured collection and analysis of a limited portion of open source data. The legal aspect of data collection and analysis will be explored with a reasoned summary of what is GDPR and how it impacts similar online services. Three sections will be dedicated to industry level best practices in relation to the security of servers facing the public internet, the usability of software interfaces and current guidelines to build reusable, mantainable and expandable code. Finally we will review briefly what are some of the current services already offering brand monitoring tools and how they differ from one another.

2.2 Open Source Intelligence

Open Source Intelligence (more commonly referred to as OSINT) is a relatively young discipline, that is concerned with the art of piecing together strategic intelligence from public sources of information. Michael Bazzel, a leading OSINT expert, defines it as:

> any intelligence produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement. [...] For the CIA, it may mean information obtained from foreign news

> broadcasts. For an attorney, it may mean data obtained from official government documents that are available to the public. For most people, it is publicly available content obtained from the internet [Bazzell(2015)].

As Bezzell explains, OSINT is not necessarialy based on online sources, at least in its most broad definition. Journalistic style, old fashion dossiers filled with newspaper clippings are a form of OSINT. However it is fair to say that whenever OSINT is mentioned today, it will automatically produce the expectation that a large proportion of content is sourced through the internet. Another important author in this field, Stewart K. Bertram, explains:

> older OSINT research was limited by both the coverage of its information and the ability of the researcher to focus the capability on a specific subject, be it a person, location or topic. [...] What has changed this status quo is the arrival of the Internet, and particularly the explosion in the use of social media technology circa 2000. The rise of these two technologies created a multilingual, geographycally distributed, completely unregulated publishing platform to which any user could also become an author and a publisher. [...] By increasing the coverage and focus of OSINT the Internet effectively promoted OSINT from a supporting role to finally sit alongside other more clandestine and less accessible investigative capabilities [Bertrom(2015)].

This explosion of sources of information causes another problem: reliability. Unless we are sourcing information from a scientifical magazine which follows the rigorous fact checking protocol of most scientific researches, then we are facing a vast landscape of information with various degrees of veridicity: reliable fact-checked sources on one side and *fake news* at the other end of the spectrum. The work of the OSINT investigator is to move in this virtually infinite universe of news, pick only relevant information and establish how reliable they are. Again this is not new, it is called *intelligence analysis*: "[...] the application of individual and collective cognitive methods to weigh data and test hypoteses within a secret socio-cultural context" [Hayes(2007)]. Information are scored from A (reliable) down to E (unreliable), plus F (reliability unknown).

Within the domain of counterfeit investigations, OSINT can have a number of roles to play. First off and foremost, should be used to establish if the items being sold are genuine or counterfeits. Secondly, it can be used to establish the extent of profits made (and therefore loss of income for the original brand) by the merchant selling them. And finally, OSINT can provide vital help in identifying the identity of the merchant as well as that of the manifacturer.

2.3 GDPR

The General Data Protection Regulation (EU) 2016/679 (a.k.a. GDPR) is the latest legislative effort of the Europea Union, that regulates the collection and use of European citizens' data. The two main goals that guided this new legislation were already set in 2015 by the Council of the European Union, which stated:

> The twofold aim of the Regulation is to enhance data protection rights of individuals and to improve business opportunities by facilitating the free flow of personal data in the digital single market [of the Council(2015)].

Essentially the GDPR gives more control to European citizens over the use of their data, thus restricting the ability of private and public organizations to process Personally Identifiable Information (a.k.a. PII) data, and at the same time it becomes a legislative framework that standardizes personal data processing in every country member of the European Union. Let's see in more detail what this entails and what are the consequences for a product like Linkero.

First off it is important to clarify some of the terminology used in the legislation. *Personally identifiable information*, or PII, is defined in privacy laws as one single piece of data that alone can identify a single person. For instance: the first and last name of a person, an email address, a physical address, a phone number, passport serial number, credit card number, a picture portraying an individual, a combiantion of username and password. GDPR applys to the collection, processing and storage of PII. A *Data Controller* is any organization that collects and manages PII belonging to European citizens. That includes online private companies (e.g. Google, Facebook the most obvious), as well as public institutions like primary and secondary schools and hospitals to name few. Another party that is involved in the handling of private data is the *Data Processor*, which in some cases conicides with the the Data Controller itself, when data is analysed internally, or with any other organization (e.g. contractors, vendors, suppiers, etc.) that receives PII from the Data Controller and processes it in its behalf. Finally the *Data Protection Officer* is a public office that is appointed to overlook and audit every Data Controller in its area of control.

What GDPR means for European citizens is that as of May 25th 2018, when it officially came into force, it grants much more rights over their own personal data, compared to previous legislations. These rights include *data portability*, which is to say that any Data Controller has to structure the collection and storage of PII in such a way that those information can be removed, transferred and re-allocated to another Data Controller without any further modification. In other words, this is a way to guarantee interoperability between Data Controllers, preventing single Controllers to

lock-in their users. GDPR also states that EU citizens have the *right to information and transparency*, so that at any moment they can request their PII to be disclosed as they have been collected up to that point in time. Another right that generated a lot of discussion, is the *right to be forgotten*, which states that users can have their information permanently erased from a Data Controller, if they no longer require their services.

GDPR also sets more stringent constraints to the operation of organizations that fall within its scope. GDPR is self-declared *transnational in scope*, meaning that it applys to every company, anywhere in the world, as long as they hold EU citizens' PII data. Obviously, enforcement is an issue in cases such as companies that are based somewhere in the Cayman Islands. However this aspect is mainly ment at large multinational corporations, that could simply transfer their databases to another jurisdiction outside the EU in an attempt to circumvent its reach. Under the current definition of GDPR, those multinationals would still be subjected to EU law and would still be accountable as long as they have legal presence in any EU member state, regardless of where the PII data actually sits. GDPR also establishes a stronger *Data Subject consent*. This means that Data Controllers can no longer assume that their users' data can be processed and used for whatever purpose, just because PII were given to them during the registration process. Nor they can pre-tick the data consent acknoledgment box, or ask the user to tick a box only if they do not consent to the treatment of their data. On the contrary, Data Controllers have to explicitly ask and receive users' consent, and consent has to be renewed over time. Finally, GDPR requires organization to *report data breaches* within 72 hours after the breach was discovered. Data Controllers not only have to notify the supervisory authority and their users, but the notification has to carry adequate details to explain how the incident occurred, how many users are affected, and how is the organization going to respond.

With all these new rights and obligations in place, the last part of the legislation concernes to the *sanctions* for organizations that fail to comply. There are three levels of penalties that can be imposed:

1. a warning letter for the first time an organization is found in breach of GDPR, and if the breach is non-intentional;

2. a more serious breach could result in the order to commit to regular periodic data protection audits;

3. and finally, if the breach is deemed to be serious enough, GDPR the prescribes two sets of financial fines depending of which obligations were missed by the Data Controller, the harsher of which could amount to up to €20 millions or up to 4% of the annual turnover estimated from the prior financial year.

The question now is: how does GDPR affect online services that for their nature collect and process bits of PII that are publicly available? First and foremost, GDPR applys only to PII belonging to EU citizens, therefore we can say that an online service that collects PII can keep on providing its functions without any further thought as long as it stays clear of EU citizens But let's say some users will be pointing the service to PII of European users, will that be sufficient to be found in breach of GDPR? The answer to this latter question is not a clear-cut *no*, there are actually use-cases where collection can still be an option. Obviously those cases do not include when Data Subjects who expressed consent to their PII to be collected and stored by anyone online, that is not how data scraping for intelligence purpose works, especially in the context of open source investigations. Although it is too early for similar cases to have been tried, GDPR accepts that organizations may have a *legitimate interest* in scraping, storing and processing publicly available PII. In the specific context in which Linkero is designed, we are talking about an Intellectual Property owner, or an agent in their behalf, that has a vested interest in gathering information about an online shop that sells their products, under market value, or without being an officially approved reseller. Having said that, GDPR definitely limits and restricts, for better or worse, the ability to extract, store and process public PII.

To conclude this section, it is interesting to mention how ensuring adequate privacy protection is a difficult balance and has its own trade-offs. This is the case of the WHOIS protocol maintained by ICANN (i.e. Internet Corporation for Assigned Names and Numbers), where internet domain registrants' PII are kept and made public. With the introduction of GDPR, personal details of the person who registered a domain will have to be hidden. Details of the domain registrant are valuable information in order to identify the person, or the signature of the person or group behind specific internet domains. This is particularly useful when security researchers try to attribute the ownership of phshing or spamming domains to a particular actor or group. Likewise, in counterfeit investigations, being able to tell how many and which domains were registered by the same actor is particularly useful, especially if you are trying to monitor the actor's activities. Information do not need to be accurate, in fact domain registrars do not verify the real identity of the person registering a domain. There are however details that need to be valid, like contact details (e.g. email address, phone number, ...) because on them depends the correct funcioning of the domain, which is in the registrant's interest. The disappearence of those PII from WHOIS caused a great deal of frustration in the security community, and ended up with ICANN filing an official litigation against the European Union asking for some sort of exhemption in order to keep offering relevant WHOIS data.

GDPR was created and voted by EU member states in a period when information is

being considered the new oil of the global economy: access to personal data and consumer behaviour generates huge amounts of revenues for corporations like Facebook and Google. And the reason for that is simple: targeted ads campaigns are extremely effective and efficient, and marketing departments are willing to pay for that. The commercial application of petabyte of consumer data is well known. But there is another aspect to an unregulated access to data belonging to billions of people, which is *political influence*. As we have seen during the 2016 US election, bad actors managed to use personal data of millions of US citizens to both target specifically voters in swing states, and to analyze general voters' opinion in order to design very effective slogans. The same actors fabricated made up stories—literally, fake news—that would appeal to the most visceral human fears in order to influence undecided voters. That happended also during the British referendum of Brexit, and was attempted during the French presidential elections in 2017, and very likely in many other countries that did not receive as much media attention. At the same time we saw government bodies abusing their ability to tap into and collect indiscriminately data, just as Edward Snowden exposed how the US National Security Agency. GDPR may not be the best solution yet, but it is a much needed regulation at a time when personal information can have serious impact on the society.

2.4 Server security standards

A chain is a strong as its weakest link, and a system is as secure as its most vulnerable element, as the saying goes. In today's reality, any system facing the open and public internet has to be secured against peole trying to misuse it. It is not just big banks and multinational corporations that have to put in place sci-fi grade defenses, but the same private wifi routers and connected thermostat have to be protected against bad actors that would gladly take them over and deploy them as pieces of a bigger scheme to perform some fraudulent or criminal activities. It is well known that professional computer criminals constantly scan the internet to detect new connected devices and try to add them to their personal collection of bots for different purposes. Internet security is therefore a must regardless of the complexity of the system designed. Information Secutiry literature identifies three main goals of a good system security strategy: confidentiality, integrity and availability, also referred to as the *CIA triad* [**?**]. That means that a system security strategy aims at guaranteeing that data stored in the system will be accessed only by legitimate and relevant parties, that the same data will be kept in its original state away from sources that may alter it, intercept it or erase it, and that despite of the safety measures in place, and potential attacks, the

data will still be accessible.

The industry identifies two main types of threat: internal and external. The external threat is probably the first and most popular to come to mind, often depicted like the hooded boy wearing a Guy Fawkes mask and typing away on his laptop: the hacker. Put simply, it is a person or organization that attacks a system from the outside. The other threat that does not get as much attention, but can be equally damaging if not more so, is the internal threat. The internal threat is represented by an internal user that intentionally of by mistake causes some damage to the system.

In conclusion, it is worth mentioning that there are three types of security measures that an organization can put in place to protect its servers: deterrent, preventive and reactive. Deterrent measures are meant to scare attackers off before they even consider to start their exploits. The criminal law with its system of penalties is definitely the first deterrent that comes to mind. But media articles detailing how sofisticate and complex is a detection system can offer the same purpose. Preventive measures on the other end are technologies that anticipate potential attacks and stop them before any damage occurs. Firewalls are a great example. Finally, reactive measures are the ones that are triggered after an attack occurred, and try to assess the damage and remediate where possible. Having a recent backup of all the data is a great antidote to ransomware attacks. In the next part of this section we will review in greater details some of the most common prevention technologies and solutions available.

2.4.1 Regular software updates

2.4.2 Firewalls

2.4.3 Environment isolation

2.4.4 Virtual Private Netwroks

2.4.5 SSL and TLS encryption

2.4.6 SSH

2.4.7 Users and groups permission on *need to do* basis

2.4.8 Password policy

2.4.9 Encryption of data at rest

2.4.10 Intrusion detection systems

2.4.11 Regular file, service and vulnerability audits

2.4.12 Business continuity plan

Backups, duplication...

2.4.13 Physical security

2.5 Ergonomy of software interface

Loren Ipsum ...

2.6 Code development principles

Robert Martin S.O.L.I.D. development principles.

2.7 Existing brand monitoring services

Loren Ipsum ...

Chapter 3

System design

Chapter 4

Project implementation

4.1 Security settings

4.2 Functions

4.3 Exception handling

Chapter 5

Testing and deployment

5.1 Unit testing

5.2 Integration testing

5.3 System testing

5.4 Usability testing

5.5 Deployment

Chapter 6

Evaluation

6.1 Lessons learned

6.2 Future work

6.3 Conclusions

6.4 Deployment

Bibliography

[Bazzell(2015)] Michael Bazzell. *Open Source Intelligence Techniques*. Amazon.co.uk Ltd, 2015.

[Bertrom(2015)] Stewart K. Bertrom. *The Tao of Open Source Intelligence*. IT Governance Publishing, 2015.

[Hayes(2007)] Joseph Hayes. *Analytic Culture in the U.S. Intelligence Community*. Ed. Center for the Study of Intelligence, 2007.

[OECD(2016)] EUIPO OECD. *Trade in Counterfeit and Pirated Goods: Mapping the Economic Impact*. OECD Publishing, Paris, 2016.

[of Commerce(2018)] US Department of Commerce. Quarterly retaul e-commerce sales. *US Census Bureau News*, 2018.

[of the Council(2015)] Presidency of the Council. Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *Official Journal of the European Union*, 2015.

[Pipkin(2000)] Donald Pipkin. *Information security: Protecting the global enterprise.* New York: Hewlett-Packard Company, 2000.