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 is 461 billions USD, which is 2.5% of the total imports in the world trade (CIT SOURCE HERE). The report also calculates that 5% of imported good in EU are counterfeit. 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, 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%. As legitimate retailers increase their presence online, shops selling counterfeit follow suit. In 2017 the Guarda custom identified xxx worth of counterfeit goods entering the Irish border, many of those destined for the European continental market.

This tool is designed to help users to extract data related to a specific brand from online platforms such as eBay, Mercadolibre, Allegro, 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 investigations, therefore we will talk about how GDPR affects services like Linkero and what are the requirements to guarantee compliance (https://blog.scrapinghub.com/web-scraping-gdpr-compliance-guide).

1.3 Objectives

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

In relation to the first type, the goal is to build a business ready web application for brand monitoring investigations. Being *business ready* means presenting a final application that run 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 functions, 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 unlimited times completed reports from the web interface, delete old queries, and search via keyword 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

Young family committments, full time job and travel arrangements, SSL configuration, DKIM configuration, app specific issues.

1.6 Deliverables

Loren Ipsum

- 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 concerne with the art of piecing together strategic intelligence from publich sources of information. Michael Bazzel, a leading OSINT expert, defines it as:

> any intelligence produced from publicly available information that is collected, explited, 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 obteined from foreign news broad-

> casts. 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.

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 source 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 speficic 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.

These explosion of sources of information causes another problem: reliability. Unless we are sourcing information for 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 Joseph (2007), Analytic Culture in the U.S. Intelligence Community, Ed. Center for the Study of Intelligence). 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

Loren Ipsum ...

2.4 Server security standards

Loren Ipsum ...

2.5 Ergonomy of software interface

Loren Ipsum ...

2.6 Code development principles

Loren Ipsum ...

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

[Sanjeev Jaiswal(2015)] Ratan Kumar Sanjeev Jaiswal. *Learning Django Web Development*. PACKT Publishing, 2015.