UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Treball de Fi de Grau

Grau en Enginyeria Informàtica: Menció en Tecnologies de la Informació

Identificació de pàgines web mitjançant captures passives a partir d'enllaços externs a un entorn xifrat

Nicolás Montesino Córdoba
nicolas.montesino@estudiantat.upc.edu

Director: Pere Barlet Ros (pbarlet@ac.upc.edu)
Codirector: Ismael Castell Uroz (ismael.castell@upc.edu)

Curs acadèmic 2021-22 Q2

Resum

Aquest projecte planteja la creació d'eines per a poder identificar pàgines web a partir de tràfic contingut a una captura en format *pcap*. Per poder-ho fer, s'ha seguit la hipòtesi que el conjunt d'adreces IP de les quals es reben o s'envien paquets identifica una pàgina web en concret. Així doncs, s'han desenvolupat les tres eines necessàries per a l'obtenció de traces, creació del model d'aprenentatge automàtic i cerca de la pàgina web.

Resumen

Este proyecto plantea la creación de herramientas para poder identificar páginas web a partir de tráfico contenido en una captura en formato *pcap*. Para poderlo hacer, se ha seguido la hipótesis de que el conjunto de direcciones IP de las que se reciben o se envían paquetes identifica una página web en concreto. Así entonces, se han desarrollado las tres herramientas necesarias para la obtención de trazas, creación del modelo de aprendizaje automático y búsqueda de la página web.

Abstract

This project contemplates the creation of tools with the intention of identifying webpages from traffic contained in a capture with *pcap* format. In order to do so, the hypothesis of the set of IP addresses to which packets are transmitted or received identifying a specific webpage has been followed. Therefore, the three needed tools to get the traces, create the machine learning model and search the website have been developed.

Índex

Índex de figures

Índex de taules

Capítol 1

Introducció

1.1 Context en el marc de la FIB

El Treball Final de Grau és una assignatura obligatòria del Grau en Enginyeria Informàtica a la Facultat d'Informàtica de Barcelona. Es cursa en el semestre 8è i consta de 18 crèdits ECTS (3 crèdits del curs de gestió de projectes i 15 crèdits del projecte pròpiament). Es preveu que es pugui realitzar a la FIB, en una empresa o en una altra universitat. [5]

En aquest cas, donat l'enfocament d'investigació que té el present treball, es realitza dins de la modalitat A, és a dir, dins la Facultat. S'elabora en col·laboració amb el director del TFG (el professor Pere Barlet) i del codirector (el doctorand Ismael Castell).

1.1.1 Competències tècniques a tractar

Dins del TFG, s'espera que l'estudiantat tracti una sèrie de competències tècniques associades a l'especialitat que s'estigui cursant, en aquest cas, la de Tecnologies de la Informació. Vegeu a la Taula 1.1 les competències a tractar en aquest projecte en concret.

1.2 Descripció del problema

En el món digitalitzat en el qual vivim, la necessitat d'accés a Internet que la humanitat té és molt notable. Tant és així que, donades les oportunitats que aquesta tecnologia brinda pel que fa a recursos educatius i llibertat d'expressió [6], l'Organització de les Nacions Unides reconeix l'accés a Internet com un dret humà.

Donada aquesta circumstància, els proveïdors d'Internet intenten constantment millorar els seus serveis. Entre altres mètodes, un dels més utilitzats són les CDN, les quals consisteixen a apropar físicament a l'usuari els servidors des dels quals s'ofereix un determinat servei per tal de reduir la latència i afegir

Competència involucrada	No es preveu	Poc	Bastant	En profunditat
CTI1				
CTI1.1			X	
CTI1.2	X			
CTI1.3	X			
CTI1.4	X			
CTI2				
CTI2.1	X			
CTI2.2				X
CTI2.3				X
CTI3				
CTI3.1				X
CTI3.2	X			
CTI3.3	X			
CTI3.4			X	
CTI4	X			

Taula 1.1: Competències involucrades en el projecte i el seu grau de profunditat. Elaboració pròpia a partir de [3]. Per a més informació sobre les competències, consulteu l'apèndix.

certa tolerància a errors. [7]. Per a fer això, els proveïdors de CDN fan una còpia de la pàgina web en un servidor propi, de tal manera que aquest servidor pugui estar més a prop del servidor "real" de la pàgina web.

Tot i això, amb la gran adopció del protocol HTTPS que hi ha avui en dia, la qual ha augmentat a un ritme molt gran en els darrers sis anys (vegeu Figura 1.1), es dificulta la tasca per part dels proveïdors d'Internet d'apropar el servei web físicament o d'utilitzar altres mecanismes que intentin millorar l'experiència d'usuari.

En aquest cas, si l'operador no té cap conveni amb la pàgina web, amb HTTPS és impossible poder crear una CDN per part de l'operador, ja que un dels objectius d'HTTPS és autenticar el servidor [8]. D'aquesta manera, una persona malintencionada no pot crear una pàgina que simuli ser l'original, però tampoc no es permet fer el mateix procediment amb intencions legítimes (com en aquest cas, crear una còpia de la pàgina per part de l'operador d'Internet per tal que l'usuari pugui accedir més ràpidament).

Tot i que els operadors tenen altres tècniques per a intentar millorar l'experiència d'usuari en aquest context, amb l'ús generalitzat d'HTTPS també els hi dificulta la seva implementació. Per exemple, un operador podria mirar d'alleugerar les rutes que hi van fins a certes pàgines web. Per a això, l'ISP hi hauria de, primer que tot, saber quines pàgines ha de fer que carreguin més ràpidament.

Els operadors d'Internet tenen tècniques com el monitoratge dels servidors DNS per a poder saber exactament a quina pàgina es connecten els usuaris. Donat que tots els encaminadors venen configurats per defecte per a fer les

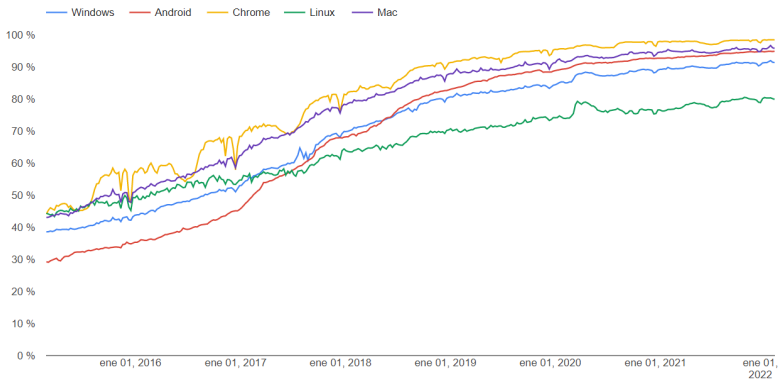


Figura 1.1: Gràfica que mostra el percentatge de pàgines carregades a través d'HTTPS a Chrome, segregat per plataforma. [2]

preguntes al servidor DNS de la companyia d'Internet, cada dia més gent està canviant els servidors DNS a uns d'altres, bé per preocupacions per la seva privadesa enfront de la seva companyia d'Internet, saltar alguns controls sobre connexions des d'un àmbit geogràfic en concret [9], o simplement per a utilitzar el protocol DNS segur. Fins i tot l'Agència Espanyola de Protecció de Dades recomana canviar-ho per aquesta darrera raó [10].

Tot i que els operadors disposen d'altres mètodes per a poder obtenir informació sobre les adreces IP a les quals es connecten els seus usuaris, aquestes eines presenten limitacions pel que fa a la quantitat de tràfic que poden processar, i per tant poden introduir desviacions en el volum real d'usuaris que utilitza un servei. Així doncs, les companyies estan interessades en altres mètodes de monitoratge per a discernir quins són els serveis web més utilitzats i optimitzar les seves xarxes i CDN de forma que millori el rendiment de la quantitat més gran d'usuaris possible.

Donada la gran quantitat de tràfic que passa per les xarxes troncals, els mètodes d'anàlisi de paquets no es poden considerar una alternativa. A més, són molt costosos tant en el sentit computacional com econòmic. S'ha de tenir en compte que fins i tot els sistemes especialitzats tenen certes limitacions a partir d'amplades de banda de 10 Gbps, fent que tinguin mal rendiment o que es perdin paquets. Si bé hi existeixen alguns mecanismes per a solucionar aquestes limitacions, són molt més costosos i complexos [11]. No obstant això, es pot evitar aquest problema treballant amb fluxos[1] en comptes d'amb paquets, ja que ocupen menys espai i per tant son més lleugers de manipular, arribant d'aquesta manera a manipular fluxos de fins a 30 Gbps amb solucions basades en software [12].

Per tot això, podem resumir que el problema de l'optimització de les xarxes

[1] Un flux consta d'una quíntupla formada per adreça IP origen i destí, port d'origen i destí i protocol.

xoca de front amb els mecanismes de seguretat que es té avui en dia. Així doncs, aquest projecte pretén saltar d'alguna manera aquests mecanismes de seguretat que ja tenim en funcionament per tal que l'operador pugui implementar solucions factibles al problema de l'alta latència de càrrega d'algunes pàgines. D'altra banda, com s'explicarà a la secció dels actors implicats és una eina que també pot fer més fàcil la seva feina a les persones malintencionades.

1.3 Estat de l'art

Els proveïdors d'Internet, en ser empreses privades, hi acostumen a donar molt poca informació pel que fa a les tecnologies que hi utilitzen per a no beneficiar a la seva competència. Així doncs, amb un repàs de la bibliografia actual resulta pràcticament impossible saber com estan resolent aquest problema. A més, alguns proveïdors disposen de departaments d'investigació propis [13], per la qual cosa no hi podem confiar en el fet que els mètodes que hi estiguin utilitzant siguin públics.

Encara que no hi tinguem cap indici que estigui sent utilitzada per cap proveïdor d'Internet, a la següent secció serà exposada una aplicació que podria ser utilitzada amb aquesta finalitat.

1.3.1 nDPI [1]

nDPI és un projecte basat en la llibreria OpenDPI [14] amb la motivació d'ampliar aquesta llibreria afegint més protocols[2] que només són suportats a la versió de pagament de la llibreria.

En el moment actual, nDPI suporta més de 246 protocols, entre ells, serveis tan populars com Google Docs, Citrix o Netflix.

En aquest sentit, després de fer una anàlisi d'aquesta eina mitjançant captures en format *pcap* de diferents pàgines, podem concloure que funciona mitjançant *Deep Packet Inspection*. Pel que fa a protocols estàndards com HTTP, és capaç d'identificar el protocol revisant la informació que conté el paquet. Per exemple, en cas d'HTTP, pot fer-ho mirant dins de la PDU de TCP, ja que s'hi trobarà un missatge HTTP. Pel que fa a HTTPS el funcionament és similar, ja que dins de la PDU de TCP es trobarà una PDU de TLS/SSL amb el valor "Application Data Protocol: http-over-tls ". Fins aquí, realment té el mateix funcionament que eines més estandarditzades com Wireshark [15]. El que té d'especial aquesta eina és que també és capaç de, en alguns casos, identificar a quin servei es realitza la connexió, encara que el tràfic estigui xifrat.

Després d'haver revisat el codi [16], podem concloure que aquesta anàlisi es realitza a partir de les adreces IP. Per exemple, en cas de tenir una connexió a una pàgina de la qual se sap el rang d'adreces IP, els paquets amb aquestes IP

[2]Hem de tenir en compte que la paraula "protocol" en aquest context no significa el que tradicionalment entenem per "protocol" a l'àmbit de les TI. En aquest sentit, es nomena "protocol" a tot el que la llibreria permet identificar, com per exemple SSH o ICMP, però també LinkedIn o Reddit.

origen o destinació les identificarà com la pàgina determinada. Això pot ser un problema en tant que les CDN existeixen, i una pàgina que estigui lliurada per una d'aquestes s'identifica com la CDN, sense proveir més informació addicional.

A més, tampoc no planteja la possibilitat que hi hagi *third-parties* a les pàgines web. Així doncs, en cas d'accedir a una pàgina que tingui un bàner de Facebook, nDPI intentarà identificar el flux de la pàgina web, però també identificarà el flux de Facebook encara que no sigui realment a la pàgina a la qual ha accedit l'usuari.

En resum, nDPI resulta d'una aplicació realment útil per a classificar tràfic, però donada la seva funcionalitat i la seva programació basada en el principi del fet que una adreça IP identifica una pàgina, no té l'objectiu de poder identificar la pàgina web concreta que s'ha lliurat.

1.4 Treball previ

Aquest projecte parteix del Treball Final de Grau d'en Carlos Jiménez Bailén [17], en el qual es plantejava la creació d'una eina que identifiques els serveis web als quals es connecta un usuari a partir de les *third-parties* a les quals es connecta aquest web.

La solució proposada en aquest projecte va ser la de poder identificar, a partir de tràfic xifrat, a quina pàgina web s'estava realitzant la connexió. Es va fer a partir dels SNI recopilats als certificats HTTPS, de tal manera que el primer certificat rebut es considera el de la pàgina web a la qual accedeix l'usuari i els altres es consideren els de les *third-parties*. Així doncs, si el SNI del primer certificat rebut és identificatiu (és a dir, que en tot el *dataset* només tenim una pàgina amb aquest SNI), ja es pot saber de quina pàgina web es tracta. Altrament, es considera que el conjunt de *third-parties* identifica a la pàgina web i es fa una cerca en el *dataset* mitjançant *Doc2Vec*. Amb aquesta eina es va aconseguir un 88% de precisió. Tot i això, són notables tres inconvenients sobre aquest projecte:

1. No es va desenvolupar una eina real. Tota la investigació va partir d'una llista de les *third-parties* que contenia cada web i dels SNI d'aquestes pàgines web i de les *third-parties*. Així doncs, si bé es va analitzar l'efectivitat amb aquestes llistes, no es va desenvolupar una eina a través de la qual en un entorn real es pogués analitzar a quina pàgina web s'estava connectant l'usuari.

2. No es va analitzar la persistència del model *Doc2Vec*. En aquest sentit, és probable que les *third-parties* de les pàgines web canviïn, i que per tant, el model baixi la seva eficàcia. No es va contemplar ni analitzar aquesta possibilitat.

3. No es va contemplar que a TLSv1.3 [18] el certificat que envia el servidor va xifrat. Així doncs, amb la implementació global de TLSv1.3, l'eina proposada és inservible en l'àmbit pràctic, ja que no es poden aconseguir els SNI ni de la pàgina web principal ni de les *third-parties*.

1.5 Objectius

Els objectius genèrics d'aquest projecte són els següents:

- Conèixer si és possible, a partir d'un tràfic xifrat, identificar el servei web al qual pertany.
- En cas que es trobi alguna manera per fer-ho, crear un prototip.
- Impulsar la investigació en el camp de la ciberseguretat.

Per a poder dur-los a terme, s'hauran d'acomplir els següents subobjectius:

- Fer recerca sobre com funcionen els xifrats empleats actualment al protocol HTTPS.
- Analitzar com podem recollir informació a través de les traces xifrades de pàgines web.
- Obtenir un recull prou extens de traces obtingudes a partir de pàgines web.
- Crear un model d'aprenentatge automàtic a partir de l'anterior recull.
- Comprovar que el model anteriorment creat funciona per identificar pàgines web amb un percentatge d'encert suficient.
- Crear la memòria del TFG explicant el projecte i el context d'aquest, amb un alt nivell de detall.

Capítol 2

Metodologia

Pel que fa a la metodologia, finalment s'ha seguit una metodologia àgil tal com es va proposar a [4] en tant que s'ha segmentat tot el projecte en petites tasques i s'han seguit mecanismes d'entrega contínua.

Aquesta ha estat la metodologia ideal, ja que tal com es pot revisar al capítol d'alternatives proposades s'han valorat diferents enfocaments per al projecte. Així doncs, en provar un possible mecanisme d'anàlisi concret i no obtenir els resultats esperats, s'ha pogut reiniciar el cicle provant un altre mecanisme però reciclant codi del cicle anterior.

En conseqüència, s'ha utilitzat un mecanisme de treball totalment adaptatiu (depenent del resultat del cicle, s'ha reorganitzat el projecte).

2.1 Planificació inicial

2.1.1 Descripció de les tasques

Tota la informació en aquest en aquesta secció s'ha extret de [4]. Les tasques del projecte estan dividides en quatre seccions: gestió del projecte, desenvolupament, tests i comprovacions i documentació i seguiment.

Gestió del projecte

- GEP1. Definició de l'abast i contextualització: definició de l'abast del projecte en el context de l'estudi.
- GEP2. Planificació temporal: planificació temporal per a l'execució total del projecte, així com la descripció de les fases del projecte i els requeriments associats a cada una.
- GEP3. Gestió econòmica i sostenibilitat: es realitzarà un pressupost per a l'execució del projecte, així com un informe de sostenibilitat.

- GEP4. Lliurable final: s'uniran les tres anteriors tasques en un sol document, tenint en compte el *feedback* del professorat.

Desenvolupament

- DEV1. Trobar diferents maneres per a aconseguir un percentatge d'encert acceptable: es realitzarà una cerca a la literatura existent, així com el desenvolupament de diferents models teòrics, per tal de poder trobar mètodes per a poder dur a terme el projecte.
- DEV2. Preparació de l'entorn de programació: es descarregaran i s'instal·laran les eines necessàries per a poder començar a programar.
- DEV3. Programar l'eina de creació del model: es realitzarà la programació de l'eina en Python.
- DEV4. Programar l'eina de cerca: es realitzarà la programació de l'eina en Python
- DEV5. Trobar possibles maneres de trobar diversos fluxos concurrents: es realitzarà una cerca a la literatura existent, així com el desenvolupament de diferents models teòrics, per tal de poder trobar mètodes per a poder dur a terme aquesta tasca.
- DEV6. Implementar la millor manera de trobar fluxos concurrents: es realitzarà la implementació de la solució en Python.
- DEV7. Paral·lelitzar l'eina de creació del model: es paral·lelitzarà l'eina de creació del model per tal que pugui ser executada amb major rapidesa.
- DEV8. Paral·lelitzar l'eina de cerca: es paral·lelitzarà l'eina de cerca per tal que pugui ser executada amb major rapidesa.

Tests i comprovacions

- TST1. Comprovar el percentatge d'encert de les eines: mitjançant comprovacions pràctiques, es comprovarà el percentatge d'encert de les eines.
- TST2. Comprovar que l'eina té un nivell d'error acceptable identificant diversos fluxos: mitjançant comprovacions pràctiques, es comprovarà que quan s'afegeixen diferents fluxos l'eina continua funcionant correctament.
- TST3. Comprovar que la paral·lelització de les eines s'ha portat a terme correctament: mitjançant comprovacions pràctiques, es comprovarà que quan s'implementa la paral·lelització l'eina continua funcionant correctament.

Codi	Estimació (en hores)	Prerequisits
GEP1	24	
GEP2	8	
GEP3	9	
GEP4	18	GEP1, GEP2, GEP3
DEV1	50	
DEV2	10	
DEV3	30	DEV1, DEV2
DEV4	30	DEV1, DEV2
DEV5	30	
DEV6	10	DEV2, DEV3, DEV4, DEV5
DEV7	1	DEV2, DEV3, DEV4
DEV8	2	DEV2, DEV3, DEV4
TST1	30	DEV3, DEV4
TST2	10	DEV6
TST3	5	DEV7, DEV8
DOC1	19	
DOC2	4	
DOC3	60	TST1, TST2, TST3
DOC4	10	DOC3

Taula 2.1: Codi de la tasca, estimació de duració en hores i prerequisits per al seu començament. Elaboració pròpia a partir de [4].

Documentació i seguiment

- DOC1. Reunió amb el codirector del projecte: reunió mitjançant Google Meet amb el codirector del projecte, l'Ismael Castell.
- DOC2. Reunió amb el director del projecte: reunió mitjançant Google Meet amb el director del projecte, en Pere Barlet.
- DOC3. Realitzar la memòria del projecte: elaboració de la memòria final del projecte.
- DOC4. Realitzar la presentació del projecte: elaboració de la presentació per a la lectura del projecte.

2.1.2 Estimació i Gantt

A la Taula 2.1 es pot una estimació de les hores que necessitarà cada tasca, així com els seus prerequisits. Pel que fa a la planificació temporal, és possible visualitzar l'esquema de Gantt a l'apèndix.

2.1.3 Desviacions

Al començament del projecte es va realitzar una planificació inicial per a organitzar el temps que es dedicaria a cada tasca, però malauradament no s'ha pogut complir donat al grau nombre d'alternatives que s'han valorat.

Concretament, tal com es va plantejar al document [4], la tasca TST1 ha fallat diverses vegades, cosa la qual ha fet tornar el projecte a DEV1. Tot i que això ha endarrerit el projecte i no s'ha pogut complir la planificació inicial, gràcies a la metodologia que s'ha utilitzat la qual permet reutilitzar codi i que s'havien sobredimensionat algunes tasques, el projecte ha pogut estar finalitzat en la data prevista.

A més, a falta de temps, necessàriament s'han hagut d'eliminar algunes tasques que no afectaven els objectius del projecte. Concretament, no s'han desenvolupat les tasques DEV5, DEV6 i DEV8. Així doncs, tampoc no s'ha executat la tasca TST2.

També cal indicar que la tasca DEV7 s'ha realitzat d'una manera que no era la plantejada inicialment. Així doncs, la part de creació del model que era factible de paral·lelitzar, es fa conjuntament amb l'obtenció de les traces, que sí que es realitza de manera paral·lela. Gràcies a aquest canvi, DEV8 no s'ha considerat una tasca necessària, ja que l'eina de cerca ja s'executa de manera suficientment ràpida.

Capítol 3

Actors implicats

Quan es desenvolupa un projecte, s'ha de fer una anàlisi de l'ambient acurada per tal de poder tenir present els possibles factors econòmics, culturals o tècnics fonamentals perquè el nostre projecte pugui tenir èxit [19].

En aquest sentit, és necessari desenvolupar una anàlisi dels actors implicats com a part de l'anàlisi de l'entorn, ja que es consideren les persones beneficiades o perjudicades del producte [20].

- Persones usuàries d'Internet: es poden veure beneficiades tant que els proveïdors d'Internet poden millorar els seus serveis, influint en l'experiència d'usuari. Tot i això, també es poden veure perjudicades, ja que la seva privacitat disminueix si l'eina és utilitzada per persones malintencionades.
- Persones malintencionades (cibercriminals): es poden veure beneficiades, ja que precisament el que fa l'eina és intentar fer *bypass* d'una protecció que s'ha vist incrementada en noves versions dels protocols TLS.
- Companyies d'Internet: es poden veure beneficiades, ja que és una eina que pot fer millorar el seu servei, per tant fent que les persones usuàries estiguin més contentes. També, amb autorització d'aquestes, poden fer servir les dades recopilades per a fins estadístics i de publicitat.
- Persones propietàries de pàgines web: es poden veure beneficiades, ja que amb col·laboració de les companyies d'Internet, podrien oferir millor servei als usuaris. Tot i això, també es poden veure perjudicades en tant que les persones malintencionades poden fer servir l'eina per a *hackejar* els seus usuaris i la seva privadesa es pot veure compromesa.
- Comunitat científica: es pot veure beneficiada, en tant que aquest projecte resulta una aportació a la comunitat i es pot fer servir per a altres projectes.
- Director i codirector del projecte: es poden veure beneficiats, en tant que es pretén crear una publicació científica a partir d'aquest projecte.

- Realitzador del projecte: em puc veure beneficiat, ja que necessito fer el projecte per a obtenir el títol de Grau i, si s'acabés fent una publicació científica, em donaria projecció en aquest àmbit.

Capítol 4

Alternatives proposades

Durant la realització del projecte, s'han revisat diverses alternatives per a poder complir els objectius. En pràcticament totes s'ha seguit el mateix procediment: primer es fa una anàlisi d'un conjunt de pàgines web, es realitza un model a partir d'algunes dades i després s'intenta a partir d'aquestes dades arribar a saber de quina pàgina web es tracta.

Les diferents alternatives contemplades durant el projecte, així com els seus percentatges d'encert, són mencionades a continuació. Totes les anàlisis s'han fet a partir del mateix conjunt de pàgines (N = 10000)[1].

1. Entrenament del model *Doc2Vec* a partir dels SNI dels certificats. 3598 encerts, 1867 *zero-thirds*[2]. Percentatge de precisió 41,66% (53,15% descomptant les *zero-thirds*).

2. Entrenament de model de comparació directa a partir dels SNI dels certificats. 6067 encerts, 1867 *zero-thirds*. Percentatge de precisió 70,24% (89,62% descomptant les *zero-thirds*).

3. Entrenament d'ambdós models a partir dels SNI dels certificats, tal que s'intenta fer comparació directa i en cas que cap entrada coincideixi es passa al *Doc2Vec*. 6067 encerts, 1867 *zero-thirds*. Percentatge de precisió 70,24% (89,62% descomptant les *zero-thirds*).

4. Entrenament del model *Doc2Vec* a partir dels SNI dels certificats i de les adreces HTTP a les quals es fan peticions. 3609 encerts, 1860 *zero-thirds*. Percentatge de precisió 41,71% (53,13% descomptant les *zero-thirds*).

5. Entrenament del model de comparació directa a partir dels SNI dels certificats i de les adreces HTTP a les quals es fan peticions. 6098 encerts,

[1] Encara que el nombre de pàgines sigui aquest, s'ha de tenir en compte que el nombre de pàgines realment recopilades és menor en totes les alternatives, ja que algunes estaven caigudes en el moment de fer l'experiment.

[2] *Zero-third*: Es diu d'aquella pàgina web sobre la qual no s'ha pogut recopilar cap informació. Per exemple, en la primera alternativa, es diu d'aquella pàgina de la qual no s'ha rebut cap certificat.

1860 *zero-thirds*. Percentatge de precisió 70,47% (89,77% descomptant les *zero-thirds*).

6. Entrenament d'ambdós models a partir dels SNI dels certificats i de les adreces HTTP a les quals es fan peticions, tal que s'intenta fer comparació directa i en cas que cap entrada coincideixi es passa al *Doc2Vec*. 6098 encerts, 1860 *zero-thirds*. Percentatge de precisió 70,47% (89,77% descomptant les *zero-thirds*).

7. Entrenament del model *Doc2Vec* a partir del camp *server_name* que es troba al *Client Hello* de HTTPS. 8262 encerts, 5 *zero-thirds*. Percentatge de precisió 95,94% (95,99% descomptant les *zero-thirds*).

8. Entrenament del model de comparació directa a partir del camp *server_name* que es troba al *Client Hello* de HTTPS. 8283 encerts, 5 *zero-thirds*. Percentatge de precisió 96,18% (96,24% descomptant les *zero-thirds*).

9. Entrenament d'ambdós models a partir del camp *server_name* que es troba al *Client Hello* de HTTPS, tal que s'intenta fer comparació directa i en cas que cap entrada coincideixi es passa al *Doc2Vec*. 8283 encerts, 5 *zero-thirds*. Percentatge de precisió 96,18% (96,24% descomptant les *zero-thirds*).

10. Entrenament del model *Doc2Vec* a partir dels SNI dels certificats, de les adreces HTTP a les quals es fan peticions i del camp *server_name* que es troba al *Client Hello* de HTTPS. 8294 encerts, 4 *zero-thirds*. Percentatge de precisió 95,89% (95,94% descomptant les *zero-thirds*).

11. Entrenament del model de comparació directa a partir dels SNI dels certificats, de les adreces HTTP a les quals es fan peticions i del camp *server_name* que es troba al *Client Hello* de HTTPS. 8323 encerts, 4 *zero-thirds*. Percentatge de precisió 96,23% (96,28% descomptant les *zero-thirds*).

12. Entrenament d'ambdós models a partir dels SNI dels certificats, de les adreces HTTP a les quals es fan peticions i del camp *server_name* que es troba al *Client Hello* de HTTPS, tal que s'intenta fer comparació directa i en cas que cap entrada coincideixi es passa al *Doc2Vec*. 8323 encerts, 4 *zero-thirds*. Percentatge de precisió 96,23% (96,28% descomptant les *zero-thirds*).

13. Entrenament del model *Doc2Vec* a partir de totes les adreces IP de les quals s'han rebut dades o s'han transmès. 8489 encerts, 0 *zero-thirds*. Percentatge de precisió 98,76%.[3]

[3]Donada l'esperada variabilitat de les adreces IP, s'ha calculat també el percentatge de precisió amb dues altres captures però el mateix model. El percentatge de precisió de la primera és del 79,35% i del 79,13% a la segona.

14. Entrenament del model de comparació directa a partir de totes les adreces IP de les quals s'han rebut dades o s'han transmès. 8482 encerts, 0 *zero-thirds*. Percentatge de precisió 98,67%.[4]

15. Entrenament d'ambdós models a partir de totes les adreces IP de les quals s'han rebut dades o s'han transmès, tal que s'intenta fer comparació directa i en cas que cap entrada coincideixi es passa al *Doc2Vec*. 8482 encerts, 0 *zero-thirds*. Percentatge de precisió 98,67%.[5]

16. Entrenament del model *Doc2Vec* a partir del camp *server_name* que es troba al *Client Hello* del HTTPS. Entrenament d'una base de dades SQL tal que s'indica la correspondència de *server_name* a adreça IP del servidor al qual s'envia, i cerca feta a partir de les adreces IP a les quals s'envien els *Client Hello*, però traduïdes prèviament a *server_name* mitjançant la base de dades. 4040 encerts, 0 *zero-thirds*. Percentatge de precisió 47%.[6]

17. Entrenament del model *Doc2Vec* a partir del camp *server_name* que es troba al *Client Hello* del HTTPS. Entrenament d'una base de dades SQL tal que s'indica la correspondència de *server_name* a adreça IP del servidor al qual s'envia, i cerca feta a partir de les adreces IP a les quals s'envien els *Client Hello*, però traduïdes prèviament a *server_name* mitjançant la base de dades, descartant aquelles IP que tenen correspondència amb més d'un *server_name*. 5514 encerts, 0 *zero-thirds*. Percentatge de precisió 64,15%.[7]

[4] Donada l'esperada variabilitat de les adreces IP, s'ha calculat també el percentatge de precisió amb dues altres captures però el mateix model. El percentatge de precisió de la primera és del 7,04% i del 6,94% a la segona.

[5] Donada l'esperada variabilitat de les adreces IP, s'ha calculat també el percentatge de precisió amb dues altres captures però el mateix model. El percentatge de precisió de la primera és del 79,49% i del 79,07% a la segona.

[6] Donat que s'espera que com més s'entreni la base de dades es pugui tenir un percentatge de precisió major, s'ha realitzat aquest procés quatre cops més. El percentatge de precisió al segon cop és de 52,52%, al tercer de 53,68%, al quart de 54,15% i al cinquè de 54,13%.

[7] Donat que s'espera que com més s'entreni la base de dades es pugui tenir un percentatge de precisió major, s'ha realitzat aquest procés quatre cops més. El percentatge de precisió al segon cop és de 67,97%, al tercer de 69,32%, al quart de 69,44% i al cinquè de 69,65%.

Capítol 5

Solució proposada

Finalment s'ha decidit l'opció numerada com a 13 en l'anàlisi d'alternatives.

5.1 Justificació

S'ha escollit aquesta opció perquè, conjuntament amb l'opció numerada com a 14 i l'opció numerada com a 15 són les úniques que, gràcies que treballen amb fluxos de connexions en lloc d'amb paquets, poden ser desplegades a un entorn real amb xarxes troncals i enllaços de desenes de Gbps. Totes tres alternatives tenen un percentatge d'encert similar, però l'escollida és menys costosa computacionalment.

A més, entre totes les opcions té un molt bon percentatge d'encert, i no presenta el problema de les *zero-third* que algunes altres opcions tenen.

5.2 Descripció de les eines

Per a poder desenvolupar aquest projecte, s'han creat tres eines:

- getTraces: S'encarrega de, a partir d'una llista de pàgines web a analitzar emmagatzemada a un arxiu CSV, obrir les pàgines i emmagatzemar, per a cada pàgina, les adreces IP a les quals s'ha accedit.
- Model: S'encarrega de, a partir de la llista d'adreces IP a les quals accedeix cada pàgina, construir el model *Doc2Vec* a partir del qual es realitzarà la cerca.
- Cerca: S'encarrega de, a partir de la llista d'adreces IP a les quals accedeix cada pàgina i el model *Doc2Vec* generat, endevinar a quina pàgina pertanyen unes adreces IP en concret. Donat el funcionament de tot el sistema el qual s'explicarà en el següent apartat, també és capaç d'indicar

si ha encertat o no[1], i el seu percentatge de precisió per a una execució en concret.

5.3 Funcionament

Les tres eines són codependents en el sentit en què és necessari fer servir totes per a obtenir els resultats esperats. Així doncs, es podrien classificar *getTraces* i *Model* com a les eines de creació del model (és a dir, la part que a un entorn de producció es faria de manera asíncrona), i l'eina *Cerca* com a l'eina final que ens donarà el resultat necessitat.

En aquest sentit, en tant que les eines són codependents, tenen una sèrie d'entrades i de sortides. Vegeu la Figura 5.1 per a major claredat.

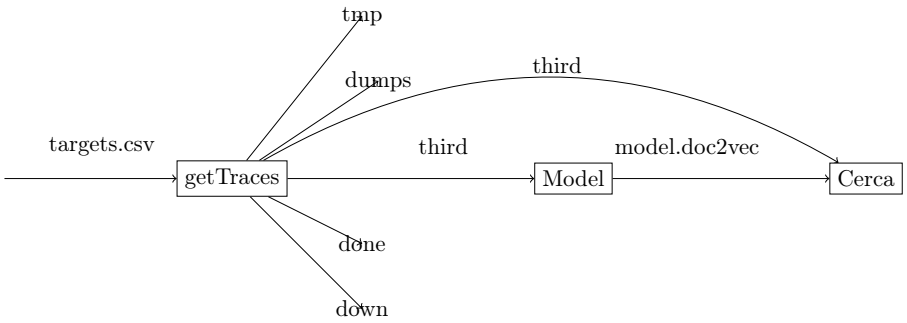


Figura 5.1: Esquema simplificat de funcionament conjunt de les tres eines, detallant les seves entrades i sortides. Elaboració pròpia.

A continuació es procedeix a explicar les entrades i les sortides de les eines:

1. getTraces
   - Entrada
     - Llista de pàgines a analitzar en format CSV.
   - Sortida
     - Fitxers amb extensió *onit*, marcats com a *tmp* a la figura. Aquests fitxers tenen com a nom la pàgina que s'està analitzant en aquell moment, així que si l'eina ha funcionat correctament, en acabar l'execució no hi hauria d'haver cap fitxer. La seva utilitat s'explica posteriorment.
     - Fitxers amb extensió *pcap*, marcats com a *dumps* a la figura. Aquests fitxers tenen com a nom la pàgina a la qual correspon, i és la captura que s'ha elaborat en accedir a aquella pàgina.

[1] Evidentment, a un entorn real no hi seria aquesta funcionalitat. S'ha incorporat al projecte per a poder analitzar el bon funcionament de les eines.

  - Fitxers amb extensió *third*, marcats com a *third* a la figura. Aquests fitxers tenen com a nom la pàgina a la qual correspon, i conté el llistat d'adreces IP amb les quals s'ha establert connexió en accedir a la pàgina. El llistat s'obté a partir de les captures justament després d'obtenir-les.
  - Fitxers amb extensió *done*, marcats com a *done* a la figura. Aquests fitxers tenen com a nom la pàgina a la qual correspon, i la seva presència indica que ja s'hi ha accedit a la pàgina i s'ha fet el procés d'obtenció de la captura i d'obtenció d'adreces IP.
  - Fitxers amb extensió *down*, marcats com a *down* a la figura. Aquests fitxers tenen com a nom la pàgina a la qual correspon, i la seva presència indica que la pàgina estava caiguda en el moment d'intentar accedir.

2. Model
   - Entrada
     - Fitxers amb extensió *third*, marcats com a *third* a la figura i elaborats per *getTraces*.
   - Sortida
     - Fitxer *model.doc2vec*, el qual conté el model *Doc2Vec* format a partir dels noms de les adreces web i les seves corresponents adreces IP.

3. Cerca
   - Entrada
     - Fitxers amb extensió *third*, marcats com a *third* a la figura i elaborats per *getTraces*.
     - Fitxer *model.doc2vec* generat per *Model*.
   - Sortida
     - Mostra per pantalla una línia per fitxer amb extensió *third*, indicant el seu nom i si ha pogut endevinar de quina pàgina es tractava mitjançant el contingut del fitxer i el model *Doc2Vec*. Finalment, mostra el nombre d'encerts i el percentatge de precisió de tota l'execució.

Després de veure el funcionament global del projecte i d'haver detallat les entrades i sortides de les eines, és possible fer una anàlisi a granularitat eina.

5.3.1 *getTraces*

Com ja s'ha mencionat anteriorment, l'eina *getTraces* és l'encarregada d'obrir les pàgines web especificades, obtenir una captura d'aquestes i analitzar la captura retornant la llista d'adreces IP a les quals s'ha connectat l'aplicació.

Per a poder fer aquest procés més ràpid, s'ha *dockeritzat* l'eina de tal manera que es pot executar diversos contenidors al mateix temps, arribant a un paral·lelisme de manera no tradicional. Aquesta ha estat la manera escollida d'assolir-ho perquè l'aplicació que s'utilitza per a l'obtenció de captures recull tot el tràfic del sistema, de tal manera que no podríem obrir més d'una pàgina a un sol sistema alhora. Com s'explica als següents paràgrafs, s'ha arribat a un mètode de comunicació entre contenidors per a poder arribar a executar l'eina d'aquesta manera.

Així doncs, el funcionament de l'eina comença per agafar el llistat en format CSV que conté totes les pàgines a analitzar. Després d'això, s'anirà agafant seqüencialment un element de la llista i s'analitzarà si existeix al directori un fitxer amb extensió *onit*, *done* o *down* amb el nom de la pàgina. En cas que existeixi, significa que un altre contenidor està o ja ha analitzat la pàgina. En cas que no existeixi cap d'aquests fitxers, es procedirà a l'anàlisi d'aquesta.

Per a començar, es crearà un arxiu amb extensió *onit* amb el nom de la pàgina com a nom de fitxer. D'aquesta manera evitem que hi hagi dos o més contenidors analitzant la mateixa pàgina. Després, s'obrirà Tcpdump [21], eina la qual farà un bolcat de tots els paquets que hi enviem o provinguin del port 80 o 443 al fitxer amb extensió *pcap*.

Un cop fet, s'obrirà Selenium Webdriver [22] amb el connector de Google Chrome [23] per a poder obrir la pàgina web. Després d'obrir-se, esperarà deu segons perquè la pàgina es pugui carregar completament i es tancaran tant Tcpdump com Selenium.

Quan s'hi hagi obtingut la traça de la pàgina web, es crearà l'arxiu amb extensió *done* i amb el nom de la pàgina web com a nom d'arxiu. En cas que la pàgina web no s'hi hagi pogut carregar (probablement perquè estigui caiguda), es crearà un arxiu amb extensió *down* amb el nom de la web com a nom d'arxiu, s'esborraran els arxius *pcap* i *onit* i es passarà a la propera pàgina web. Altrament, es passarà la captura en format *pcap* a Pyshark [24] i s'extrauran les adreces IP de les quals s'han rebut o transferit dades. Després, es bolcaran a l'arxiu amb extensió *third* i el nom de la pàgina web com a nom de l'arxiu. Finalment, s'esborrarà el fitxer *onit*.

Un cop acabat, es passarà a la següent pàgina web en cas que no s'hi hagi acabat la llista. Podeu veure aquesta informació de manera visual a la Figura 5.2.

5.3.2 Model

L'eina de creació del model, és l'encarregada de, a partir de tots els fitxers amb extensió *third*, crear un model *Doc2Vec* que pugui ser utilitzat per l'eina de cerca.

Així doncs, per a portar aquesta acció a terme, agafa tots aquests fitxers i posa tots els noms dels fitxers a una estructura *tags* i el seu contingut a una estructura *data*. D'aquesta manera, la posició i-ena de l'estructura *tags* correspon a la posició i-ena de l'estructura *data*. A partir d'aquestes estructures

Figura 5.2: Diagrama de flux de l'eina *getTraces*. Elaboració pròpia.

elaborarà un *TaggedDocument* que sigui compatible amb les funcions de *Doc2Vec* que s'utilitzen per a crear el model.

Després, crea i emmagatzema el model a un arxiu anomenat *model.doc2vec*. Cal comentar que s'ha detectat que la funció utilitzada per a guardar el model, en cas que la mida d'aquest superi un cert llindar, crea dos fitxers més: *model.doc2vec.syn1neg.npy* i *model.doc2vec.wv.vector.npy*. Això no modifica el funcionament de l'eina en tant que tinguem els tres fitxers com a entrada de la següent eina.

Podeu veure aquesta informació de manera visual a la Figura 5.3.

5.3.3 Cerca

L'eina de cerca és l'encarregada de poder fer el que finalment es desitja amb aquest projecte: a partir d'una captura, poder saber a quina pàgina s'està

connectant l'usuari.

S'ha de tenir en compte que, tal com s'ha comentat anteriorment, actualment té com a entrada els fitxers *third* i no una captura *pcap* com s'esperaria. Aquesta decisió només respon a qüestions d'eficiència pel que fa a l'anàlisi de resultats, però reutilitzant funcions de *getTraces* (concretament, la que fa agafar les adreces IP a partir de la captura en format *pcap*) es podria canviar fàcilment i s'obtindria exactament el mateix resultat.

Així doncs, la primera entrada que necessita l'eina és el fitxer *model.doc2vec*[2], amb el qual posteriorment podrà saber a quina pàgina pertany un conjunt d'adreces IP. Després d'això, el programa recorre de manera seqüencial tots els arxius *third* dels que disposa.

Per a cada arxiu, infereix el seu contingut amb el model *Doc2Vec* i en cas que el nom de l'arxiu (traient l'extensió) sigui igual que el resultat de la inferència, es considera que s'ha encertat la pàgina. També es manté un comptador intern del nombre d'encerts. Després d'haver decidit si s'ha encertat o no, es mostra per pantalla.

Un cop no hi quedin més arxius *third*, es mostra per pantalla el nombre total d'encerts i el percentatge de precisió a l'execució.

Podeu veure aquesta informació de manera visual a la Figura 5.4.

[2]Com s'ha comentat anteriorment, segons la mida del model, la llibreria que s'utilitza pot també crear dos arxius més (*model.doc2vec.syn1neg.npy* i *model.doc2vec.wv.vector.npy*). En cas que hi hagi estat així, són necessaris tots tres arxius com a entrada de l'eina de cerca.

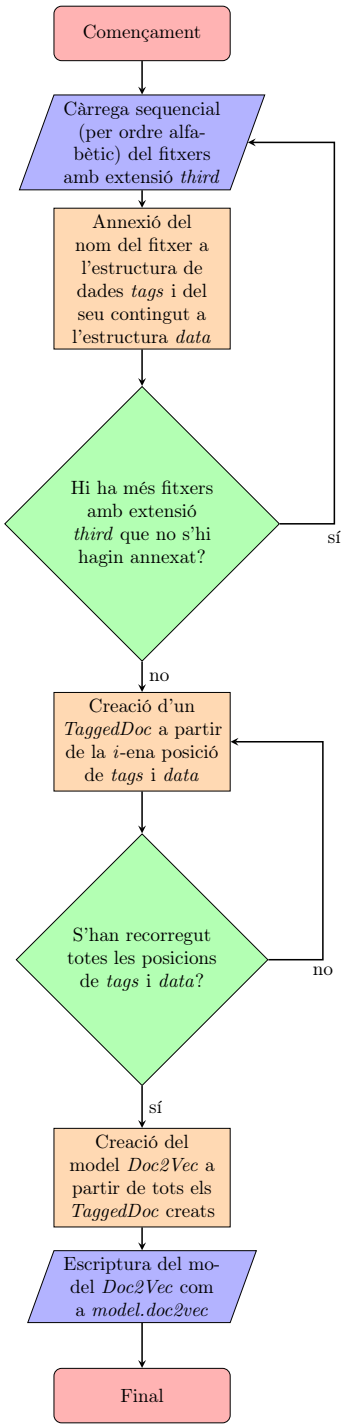


Figura 5.3: Diagrama de flux de l'eina de creació del model. Elaboració pròpia.

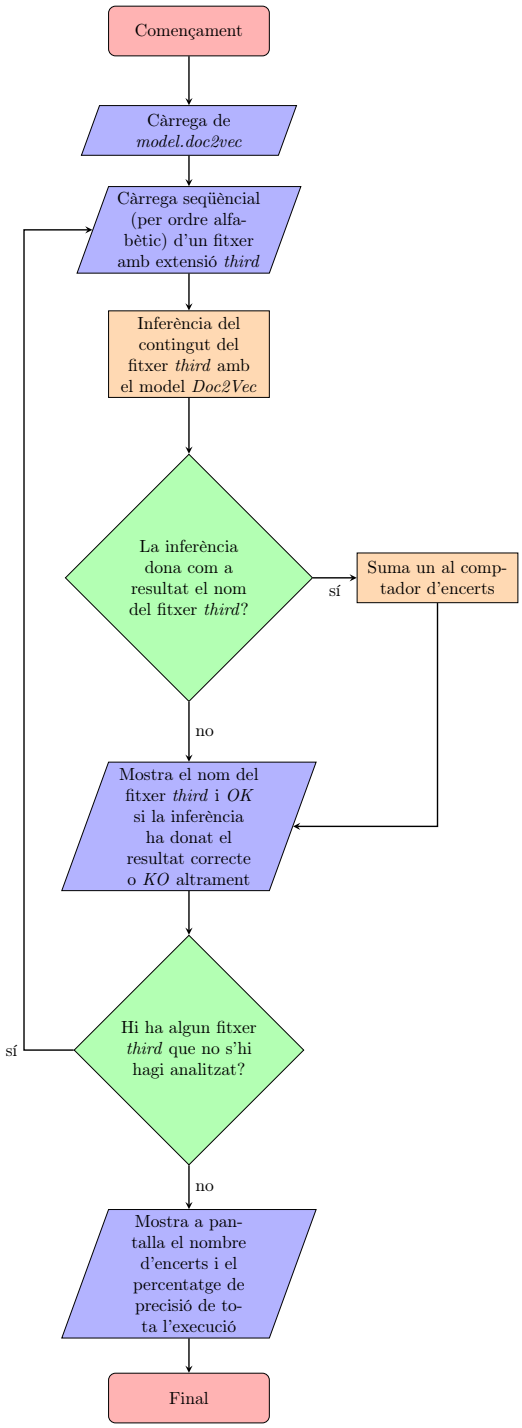


Figura 5.4: Diagrama de flux de l'eina de cerca. Elaboració pròpia.

Capítol 6

Anàlisi

6.1 Anàlisi de resultats

6.1.1 Precisió directa

Per a poder fer aquesta anàlisi, s'han utilitzat les 30000 primeres entrades de la llista Trancolist [25] amb identificador *X52GN*, generada el 5 de juny del 2022. Trancolist es tracta d'una llista utilitzada freqüentment per a investigacions relacionades amb seguretat web o Internet, i que es compon del milió de pàgines més visitades, ordenades per la seva popularitat.

En aquest sentit, s'han realitzat set execucions de l'eina *getTraces*. També s'han creat diversos models *Doc2Vec*, formats per les traces obtingudes a partir d'una o més execucions de *getTraces*. D'aquesta manera, en els models que contenen les traces de més d'una execució de *getTraces*, és possible evitar els inconvenients que el contingut de les pàgines web sigui dinàmic (i, per tant, que canviïn algunes adreces IP d'execució a execució) i podem fer una anàlisi més acurada. A més, és el que es faria a un entorn real per a adaptar-se als canvis temporals.

A continuació es comenten els *datasets* obtinguts:

1. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 6 de juny del 2022 a les 23:49 i finalitzada el dia 8 de juny del 2022 a les 11:20. De les 30000 webs que s'han analitzat, 5393 estaven caigudes.

2. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 8 de juny del 2022 a les 13:04 i finalitzada el dia 10 de juny del 2022 a les 01:16. De les 30000 webs que s'han analitzat, 5399 estaven caigudes. Hi ha 301 pàgines que sí que s'han pogut detectar, però que estaven caigudes a la primera execució.

3. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 10 de juny del 2022 a les 09:36 i finalitzada el dia 11 de juny del 2022 a les 21:04. De les 30000 webs que s'han analitzat, 5412 estaven caigudes. Hi ha 56

pàgines que sí que s'han pogut detectar, però que estaven caigudes a les execucions anteriors.

4. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 12 de juny del 2022 a les 12:45 i finalitzada el dia 13 de juny del 2022 a les 21:48. De les 30000 webs que s'han analitzat, 5398 estaven caigudes. Hi ha 29 pàgines que sí que s'han pogut detectar, però que estaven caigudes a les execucions anteriors.

5. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 14 de juny del 2022 a les 11:57 i finalitzada el dia 16 de juny del 2022 a les 03:21. De les 30000 webs que s'han analitzat, 5387 estaven caigudes. Hi ha 27 pàgines que sí que s'han pogut detectar, però que estaven caigudes a les execucions anteriors.

6. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 16 de juny del 2022 a les 08:08 i finalitzada el dia 17 de juny del 2022 a les 20:23. De les 30000 webs que s'han analitzat, 5400 estaven caigudes. Hi ha 24 pàgines que sí que s'han pogut detectar, però que estaven caigudes a les execucions anteriors.

7. *Dataset* obtingut a partir de l'execució de *getTraces* iniciada el dia 17 de juny del 2022 a les 23:26 i finalitzada el dia 19 de juny del 2022 a les 10:58. De les 30000 webs que s'han analitzat, 5422 estaven caigudes. Hi ha vuit pàgines que sí que s'han pogut detectar, però que estaven caigudes a les execucions anteriors.

El percentatge de precisió que es mencionarà a aquesta secció es tracta del nombre de pàgines que ha pogut encertar l'eina sobre el nombre de pàgines totals que hi havia a l'execució (és a dir, aplicant directament el model *Doc2Vec* sobre un dels *datasets* obtinguts). A la Taula 6.1 és possible visualitzar els percentatges obtinguts, de tal manera que les columnes indiquen els *datasets* mencionats anteriorment i les files indiquem els models generats a partir d'aquests. Així doncs, el model número tres està entrenat amb els *datasets* primer, segon i tercer. D'aquesta manera, les cel·les ombrejades signifiquen que el percentatge de precisió indicat pertany a l'entorn de proves (és a dir, que el model amb el qual s'assoleix el percentatge conté el *dataset* en qüestió) i les buides signifiquen que el percentatge indicat pertany a l'entorn de validació (és a dir, el model no ha estat entrenat amb aquell *dataset* en concret).

Així doncs, amb la Taula 6.1 podem confirmar que en afegir més *datasets* que complementen els anteriors i afegeixen adreces IP fins aquell moment desconegudes, la precisió per als *datasets* no inclosos al model augmenta. Tot i això, aquest comportament és degut que en aquest projecte s'utilitza un nombre limitat de *datasets*. A la pràctica, les adreces IP també desapareixen amb el temps degut a la naturalesa dinàmica de les pàgines web, i s'hauria de calcular la finestra de temps òptima.

Model\Dataset	1	2	3	4	5	6	7
1	98,62%	71,65%	66,42%	66,09%	66,44%	64,41%	68,66%
2	96,73%	96,94%	72,74%	73,02%	71,48%	71,26%	74,14%
3	95,16%	95,36%	96,83%	77,54%	75,49%	75,99%	77,50%
4	93,16%	93,54%	95,49%	95,59%	81,60%	82,88%	77,19%
5	93,35%	93,50%	95,06%	94,98%	95,04%	84,28%	79,92%
6	93,01%	93,07%	94,78%	94,52%	94,54%	94,53%	80,26%

Taula 6.1: Taula indicant el percentatge de precisió per a cada *dataset* amb diferents models. Elaboració pròpia.

També s'ha representat la informació de la taula al Gràfic 6.1. Així doncs, cada línia representa un model diferent, i les línies de punts indiquen quin *dataset* és el darrer amb el qual s'ha entrenat el model en concret.

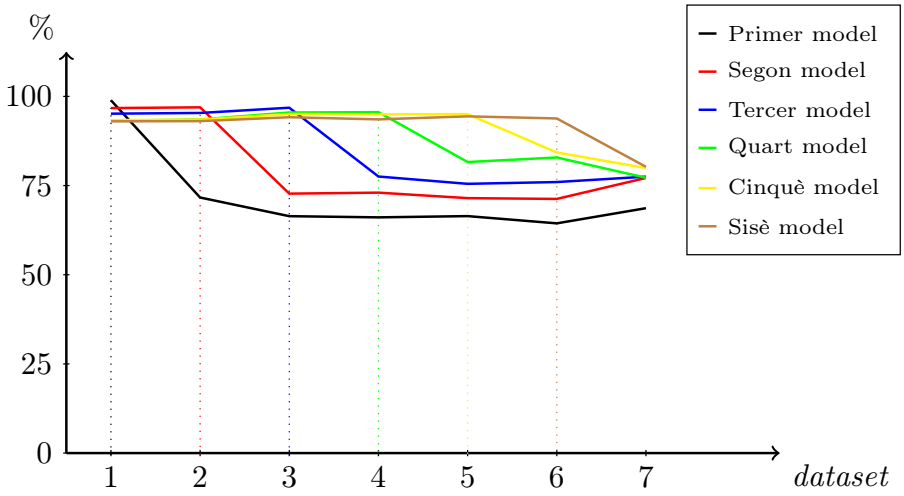


Figura 6.1: Gràfic mostrant els percentatges de precisió obtinguts amb diferents models i *datasets*. Elaboració pròpia.

6.1.2 Precisió ponderada

Com a part d'aquesta anàlisi també s'ha revisat la precisió ponderada, amb la intenció de revisar si els errors estan distribuïts de manera uniforme entre totes les pàgines, o bé si s'acumulen a les pàgines més o menys populars. En aquest sentit, si aquest percentatge fos major que la precisió directa, es pot considerar que a un entorn real podria tenir més encerts, ja que probablement els usuaris hi accedeixin més a les pàgines més populars.

Pel que fa al percentatge de precisió ponderada, es té en compte la popularitat del web en qüestió. Trancolist disposa de les pàgines ordenades per popularitat, de tal manera que la número u és la més popular. Així doncs, s'ha

calculat un factor diferent per a cada pàgina, que serà sumat al percentatge total en cas que l'eina l'hagi encertat. Per tant, el percentatge de precisió ponderada es defineix com:

$$\sum_{i \in A}(N - i + 1)/\frac{N(N+1)}{2} = \sum_{i \in A}\frac{2(N - i + 1)}{N(N+1)}$$

Sent A el conjunt de les entrades que l'eina ha pogut endevinar, N el nombre de pàgines de l'execució (és a dir, a prop de 30000) i i el número en el qual se situa a la *Trancolist*. Així doncs, $i \in A$ serà cert si la pàgina ha estat endevinada per la pàgina de cerca.

Per a poder fer l'anàlisi de precisió ponderada i saber quantes pàgines hi són presents a successives execucions i no a les anteriors (dada mostrada prèviament), s'han desenvolupat dos scripts en llenguatge de programació R amb suport de la llibreria SQLite [26].

La metodologia i els *datasets* utilitzats son els mateixos que els de la secció anterior. Es poden veure els resultats obtinguts a la Taula 6.2 de tal manera que les columnes indiquen els *datasets* mencionats anteriorment i les files indiquem els models generats a partir d'aquestos. Així doncs, el model número tres està entrenat amb els *datasets* primer, segon i tercer. D'aquesta manera, les cel·les ombrejades signifiquen que el percentatge de precisió indicat pertany a l'entorn de proves (és a dir, que el model amb el qual s'assoleix el percentatge conté el *dataset* en qüestió) i les buides signifiquen que el percentatge indicat pertany a l'entorn de validació (és a dir, el model no ha estat entrenat amb aquell *dataset* en concret).

Model\Dataset	1	2	3	4	5	6	7
1	98,64%	71,01%	65,36%	65,15%	65,07%	63,50%	67,02%
2	96,55%	96,55%	70,67%	71,59%	70,27%	69,21%	74,26%
3	94,80%	94,94%	96,35%	75,90%	74,49%	73,93%	75,88%
4	93,05%	93,00%	94,91%	94,94%	80,43%	80,96%	77,72%
5	93,05%	93,05%	94,44%	94,05%	94,86%	82,16%	79,69%
6	92,73%	92,71%	94,16%	93,56%	94,41%	93,82%	80,31%

Taula 6.2: Taula indicant el percentatge de precisió ponderada per a cada *dataset* amb diferents models. Elaboració pròpia.

En aquest sentit, la precisió ponderada s'ha calculat amb la intenció de veure si la precisió pujava o baixava per a les pàgines més populars. Podem veure que els resultats de precisió directa i de precisió ponderada son molt similars, per la qual cosa es pot concloure que els errors estan distribuïts de manera uniforme entre totes les pàgines, i que no es concentren a una part en concret de la llista.

També s'ha representat la informació de la taula al Gràfic 6.2. Així doncs, cada línia representa un model diferent, i les línies de punts indiquen quin *dataset* és el darrer amb el qual s'ha entrenat el model en concret.

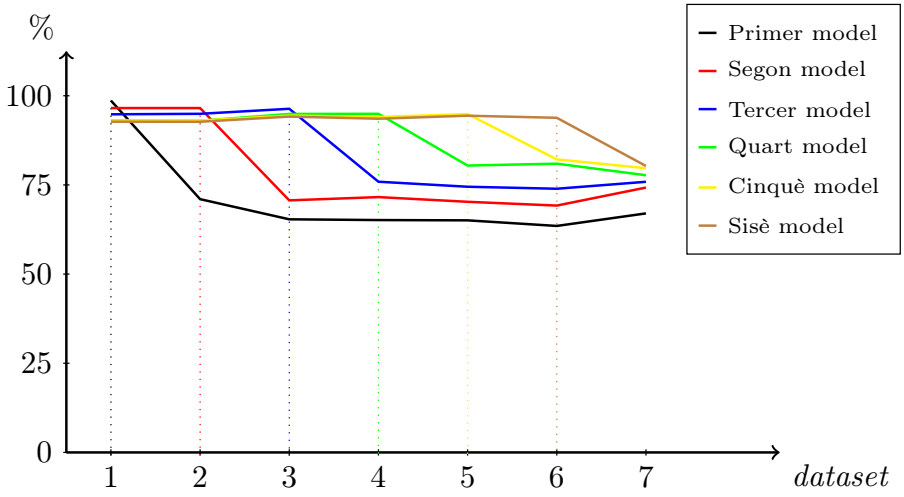


Figura 6.2: Gràfic mostrant els percentatges de precisió ponderada obtinguts amb diferents models i *datasets*. Elaboració pròpia.

6.2 Anàlisi de costos

Tot i que aquest projecte s'ha desenvolupat per l'estudiant amb suport del director i del codirector, s'ha elaborat també un pressupost en cas que aquest projecte es vulgui desplegar a una empresa. En aquest sentit, s'ha tingut la planificació inicial, ja que es considera que en aquest cas sí que seria factible executar les tasques d'identificació de diversos fluxos.

Donades les estimacions que hem fet per a les diferents tasques, podem concloure que necessitarem les següents hores de personal:

- 152 hores de *Project Manager*, corresponents a les tasques GEP1, GEP2, GEP3, GEP4, DOC1, DOC2, DOC3 i DOC4.
- 80 hores de *Knowledge Manager*, corresponents a les tasques DEV1 i DEV5.
- 10 hores de tècnic de sistemes, corresponents a la tasca DEV2.
- 133 hores de programador, corresponents a les tasques DEV3, DEV4, DEV6, DEV7, DEV8 i DOC3.
- 105 hores de *QA Manager*, corresponents a les tasques TST1, TST2, TST3 i DOC3.

Donada la planificació temporal que hem realitzat, s'haurà d'analitzar quant de temps es necessitarà pel que fa a la coordinació de l'equip (és a dir, hipotètiques tasques de reunions amb altres membres de l'equip). Per senzillesa, es pot concloure que dues hores de la jornada laboral diària de l'equip es dedicarà a

descansos i reunions i es tindrà en compte que la setmana laboral és de cinc dies.

- Pel que fa al *Project Manager*:
  - Les tasques DOC1 i DOC2 es realitzen durant tota la durada del projecte, és a dir, 21,2 setmanes. Es pot concloure que aquestes tasques hi sumaran 1,08 hores setmanals.
  - Pel que fa a DOC3, no es realitza concurrentment amb altra tasca (a part de DOC1 i DOC2), i dura 6,2 setmanes. Així doncs, durant aquest període de temps, s'haurà de sumar a l'equip de *Project Management* una càrrega de 9,68 hores setmanals.
  - Pel que fa a DOC4, no es realitza concurrentment amb altra tasca (a part de DOC1 i DOC2), i dura quatre setmanes. Així doncs, durant aquest període de temps, s'haurà de sumar a l'equip de *Project Management* una càrrega de 2,5 hores setmanals.
  - Pel que fa a GEP1, GEP2, GEP3 i GEP4, no es realitzen concurrentment amb altra tasca (a part de DOC1 i DOC2), i duren 4,2 setmanes. Així doncs, durant aquest període de temps, s'haurà de sumar a l'equip de *Project Management* una càrrega de 14,05 hores setmanals.

  Durant el temps que estiguin fent les tasques referent al bloc de GEP, es necessitaran 1,08+14,05+2·5=25,13 hores setmanals, durant 4,2 setmanes.

  Durant el temps que estiguin fent les tasques referent a DOC3, es necessitaran 1,08+9,68+2·5=20,76 hores setmanals, durant 6,2 setmanes.

  Durant el temps que estiguin fent les tasques referent a DOC4, es necessitaran 1,08+2,5+2·5=13,58 hores setmanals, durant quatre setmanes.

  La resta del temps, és a dir, 21,2-4,2-6,2-4=6,8 setmanes, es necessitaran 1,08+2·5=11,08 hores setmanals.

  Donat que en cap moment es necessiten més de 40 hores setmanals per part d'aquest equip, es conformarà per una única persona.

- Pel que fa al *Knowledge Manager*:
  - Pel que fa a DEV1, no es realitza concurrentment amb cap altra tasca i dura 1,8 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de *Knowledge Management* una càrrega de 27,78 hores setmanals.
  - Pel que fa a DEV5, no es realitza concurrentment amb cap altra tasca i dura una setmana. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de *Knowledge Management* una càrrega de 30 hores setmanals.

Durant el temps que estiguin fent les tasques referent a DEV1, es necessitaran 27,78+2·5=37,78 hores setmanals, durant 1,8 setmanes.

Durant el temps que estiguin fent les tasques referent a DEV5, es necessitaran 30+2·5=40 hores setmanals, durant una setmana.

Donat que durant la resta del temps aquest equip no té cap tasca assignada, es prescindirà d'ell. Ja que en cap moment es necessiten més de 40 hores setmanals per part d'aquest equip, es conformarà per una única persona, tot i que estaria bé ampliar-ne el nombre d'integrants a dos.

- Pel que fa al tècnic de sistemes:
  - Pel que fa a DEV2, no es realitza concurrentment amb altra tasca i dura 2,2 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de sistemes una càrrega de 4,55 hores setmanals.

  Durant el temps que estiguin fent les tasques referent a DEV2, es necessitaran 4,55+2·5=14,55 hores setmanals, durant 2,2 setmanes.

  Donat que durant la resta del temps aquest equip no té cap tasca assignada, es prescindirà d'ell. Donat que en cap moment es necessiten més de 40 hores setmanals per part d'aquest equip, es conformarà per una única persona.

- Pel que fa al programador:
  - Pel que fa a DEV3 i DEV4, es realitzen concurrentment i duren dues setmanes. Així doncs, en aquest període de temps s'haurà de sumar a l'equip de programació una càrrega de 30 hores setmanals.
  - Pel que fa a DEV6, hi ha quatre dies de concurrència amb DEV7 i DEV8 i dura 1,2 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de programació una càrrega de 8,33 hores setmanals.
  - Pel que fa a DEV7 i DEV8, es realitzen concurrentment entre elles, hi ha quatre dies de concurrència amb DEV6 i duren 1,6 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de programació una càrrega d'1,88 hores setmanals.
  - Pel que fa a DOC3, no es realitza concurrentment amb cap altra tasca i dura 6,2 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de programació una càrrega de 9,68 hores setmanals.

  Durant el temps que estiguin fent les tasques referents a DEV3 i DEV4, es necessitaran 30+2·5=40 hores setmanals, durant dues setmanes.

  Durant el temps que estiguin fent les tasques referents a DEV6, es necessitaran 8,33+2·5=18,33 hores setmanals, durant 1,2 setmanes.

Durant el temps que estiguin fent les tasques referents a DEV7 i DEV8, es necessitaran 1,88+2·5=11,88 hores setmanals, durant 1,6 setmanes.

  - Hem de tenir en compte que les tasques descrites als dos darrers paràgrafs tenen una concurrència de quatre dies, durant els quals es necessitaran 8,33+1,88+2·5=20,21 hores setmanals.

Durant el temps que estiguin fent les tasques referents a DOC3, es necessitaran 9,68+2·5=19,68 hores setmanals, durant 6,2 setmanes. Donat que durant la resta del temps aquest equip no té cap tasca assignada, es prescindirà d'ell. Ja que en cap moment es necessiten més de 40 hores setmanals per part d'aquest equip, es conformarà per una única persona, tot i que estaria bé ampliar-ne el nombre d'integrants a dos en algun punt del projecte.

- Pel que fa al *QA Manager*:

  - Pel que fa a TST1, no es realitza concurrentment amb cap altra tasca i dura 0,8 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de *QA Management* una càrrega de 30 hores setmanals.
  - Pel que fa a TST2 i TST3, no es realitzen concurrentment amb cap altra tasca i duren 1,2 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de *QA Management* una càrrega de 12,5 hores setmanals.
  - Pel que fa a DOC3, no es realitza concurrentment amb cap altra tasca i dura 6,2 setmanes. Així doncs, durant aquest període de temps s'haurà de sumar a l'equip de *QA Management* una càrrega de 9,68 hores setmanals.

  Durant el temps que estiguin fent les tasques referents a TST1, es necessitaran 30+5·2=40 hores setmanals, durant una setmana.

  Durant el temps que estiguin fent les tasques referents a TST2 i TST3, es necessitaran 12,5+5·2=22,5 hores setmanals, durant 1,2 setmanes.

  Durant el temps que estiguin fent les tasques referents a DOC3, es necessitaran 9,68+5·2=19,68 hores setmanals, durant 6,2 setmanes.

  Donat que durant la resta del temps aquest equip no té cap tasca assignada, es prescindirà d'ell. Ja que en cap moment es necessiten més de 40 hores setmanals per part d'aquest equip, es conformarà per una única persona, tot i que estaria bé ampliar-ne el nombre d'integrants a dos en algun punt del projecte.

Un cop analitzades les persones que hi treballaran al projecte, podem concloure que de material necessitarem únicament cinc ordinadors. Per tal de poder fer l'estimació dels costos, arrodonirem a l'alça els valors de les setmanes treballades i de les hores, per tal de poder tenir marge per a imprevistos.

	Preu per hora (en €)	Hores per setmana	Setmanes	Total (en €)	Total en € (amb SS)
Project Manager	21,9	26	5	2847	3416,4
		21	7	3219,3	3863,16
		14	4	1226,4	1471,68
		12	7	1839,6	2207,52
Knowledge Manager	18,07	38	2	1373,32	1647,98
		40	1	722,8	867,36
Tècnic de sistemes	15,88	15	3	714,6	857,52
Programador	13,69	40	2	1095,2	1314,24
		19	2	520,22	624,26
		12	2	328,56	394,27
		21	1	287,49	344,99
		20	7	1916,6	2299,92
QA Manager	26,29	40	1	1051,6	1261,92
		23	2	1209,34	1451,21
		20	7	3680,6	4416,72
				Total:	26439,16

Taula 6.3: Taula amb despeses de personal desglossades, classificades per lloc de treball. Elaboració pròpia.

Per a poder obtenir els preus per hora dels diferents llocs de treball, s'ha trobat la mitjana del salari brut anual a [27], [28], [29], [30], [31] i s'ha dividit per 1826 hores, que és la jornada màxima anual a Espanya [32]. Es pot veure l'estimació dels costos humans a la Taula 6.3

Si se suposa que l'equip està a teletreball (com veiem que és la tònica al sector TIC [33]), només s'hauria de fer la despesa dels ordinadors en cas que no hi tinguem com a recurs, i a priori no hauríem de fer cap mena de despesa general.

Pel que fa al software, es farà sevir únicament software lliure, per la qual cosa també hi haurà despesa zero en aquest sentit.

Tenint en compte un cost de 1175,02 € per cada Thinkpad T14 Gen 2 [34], es faria una despesa de 5875,10 € pel material.

Si això se suma al cost del personal, el projecte en total té una despesa de 32314,26 €. S'ha de tenir en compte que aquests equips es podrien utilitzar o revendre després. També s'ha de tenir en compte que les jornades que s'han configurat per al personal són infactibles a priori, i que el projecte està pensat per a fer dins d'una empresa en la qual aquest projecte només sigui una part del treball del personal, o bé com a *outsourcing* o *freelance*.

6.3 Anàlisi de sostenibilitat

6.3.1 Dimensió econòmica

Les despeses que aquest projecte suposa estan detallades a l'apartat de l'anàlisi dels costos. En aquest sentit, per a saber si és viable o no, s'hauria de calcular quin retorn es tindria amb l'eina.

No s'espera que el projecte tingui sobrecostos: de fet s'espera tot el contrari, que l'estimació estigui sobredimensionada. Tot i això, per a poder calcular la desviació entre el pressupost i el cost real s'hauria d'executar el projecte.

6.3.2 Dimensió ambiental

L'ordinador que s'ha escollit per desenvolupar té un consum de 65 W. En aquest cas, a partir de l'apartat d'estimació i Gantt podem assumir que, amb tot el projecte, hi tindrem 480 hores d'ordinador en ús. Així doncs, l'impacte en aquest sentit és de 31,2 kWh.

També s'ha de tenir en compte que, després de desenvolupar el projecte, es posarà en marxa en algun lloc. Es preveu que en cas de ser desenvolupat per a la seva posada en producció, faci l'anàlisi en temps real, és a dir, el servidor estarà encès tant temps com es vulgui fer l'anàlisi del tràfic. En aquest sentit, es pot considerar que el consum del servidor serà de 450 W (així que l'impacte de tenir el servidor encès durant un any serà de 3943,62 kWh per any), encara que depenent de quant tràfic hi hagi s'haurien d'incorporar servidors addicionals. Donat que només s'ha desenvolupat una prova de concepte, és complicat donar una mesura de quants servidors faran falta segons el flux de la xarxa. Probablement és possible minimitzar l'impacte només canviant les màquines proposades per unes de més eficients energèticament.

També s'ha de tenir en compte que el projecte està programat en Python, llenguatge que és molt poc eficient energèticament [35]. Es podria programar, per exemple, en el llenguatge C que és molt més eficient pel que fa a energia.

D'altra banda, donat l'enfocament del projecte de treball amb fluxos de connexió, és molt més eficient pel que fa a memòria, disc i ús de processador que l'enfocament de treballar amb paquets. [11][12].

6.3.3 Dimensió social

En aquest cas, no es destrueixen llocs de treball amb aquest projecte, ja que és un treball que presumiblement no s'està fent, i que no hi ha possibilitat de fer manualment.

Pel que fa a les persones amb necessitats especials, en interactuar a través del terminal (el nostre es tracta d'un projecte que no està destinat a un usuari final, sinó que principalment està destinat a empreses de telecomunicacions), tot dependrà de les opcions d'accessibilitat que tingui el seu ordinador configurat. Per exemple, una persona amb visibilitat reduïda ja tindrà instal·lat una eina

com Brltty [36] per a poder interactuar amb l'ordinador. L'eina estarà tan adaptada com el seu entorn de treball ho estigui.

Tot i que l'eina es pot utilitzar per al bé (i en aquest cas, en col·laboració amb els proveïdors d'internet milloraria el servei que aquests presten als seus clients), una persona malintencionada podria fer-la servir per a saber on s'està connectant un usuari, sense l'usuari saber-ho i en contra de la seva voluntat.

Es tracta d'una eina que, a més, pot portar benefici a la comunitat científica si es plantegés com a un projecte *open-source*.

6.4 Anàlisi de lleis i regulacions

A l'àmbit legal, si bé no hi ha cap regulació que afecti exactament el nostre projecte, sí que hi existeixen lleis que castiguen l'escolta d'un mitjà de comunicació sense permís explícit de les persones usuàries. En aquest cas, podrien afectar a l'empresa de telecomunicacions si les persones usuàries no donessin el consentiment necessari perquè pugui prendre aquesta informació, o bé a possibles cibercriminals que utilitzin l'eina de manera maliciosa.

En aquest cas, el text que afecta és el Capítol I *Del descubrimiento y revelación de secretos* del títol desè *Delitos contra la intimidad, el derecho a la propia imagen y la inviolabilidad del domicilio* de la Llei Orgànica 10/1995, del 23 de novembre, del Codi Penal [37].

Concretament, a l'article 197 es preveu una pena de presó d'un a quatre anys i multa de dotze a 24 mesos per a qui, per descobrir secrets o vulnerar la intimitat d'un altre sense el seu consentiment, intercepti les seves telecomunicacions o utilitzi artificis tècnics d'escolta de so, imatge, o qualsevol altre senyal de comunicació. Aquest fet es veu agreujat fins a una pena de dos a cinc anys de presó si es difonen, revelen o se cedeixen aquestes dades a tercers. A més, en cas que aquesta conducta sigui comesa per les persones encarregades o responsables dels suports informàtics o telemàtics (com podria ser el cas de les empreses de telecomunicació amb els seus usuaris o el d'una empresa amb els seus empleats i empleades), es contempla una pena de tres a cinc anys de presó. En cas que es faci amb finalitat lucrativa, s'imposen les penes en la seva meitat superior, i si a més afecta dades com poden ser la ideologia, religió, creences, salut, origen racial o vida sexual (cosa la qual és factible en cas que la persona usuària estigui accedint a alguna pàgina especialitzada en aquests àmbits), o la víctima és menor d'edat o una persona amb discapacitat que necessiti especial protecció, la pena a imposar serà de quatre a set anys de presó.

A l'article 197 quarter del mateix text també s'indica que si aquests fets són comesos per una organització criminal, s'aplicaran les penes que siguin superiors en grau.

Capítol 7

Treball futur

Donat l'estricte calendari que té aquest projecte, no s'ha pogut arribar a la implementació per tal que l'eina pugui identificar diversos fluxos. Així doncs, si s'implementés, l'aplicació seria capaç d'identificar diverses pàgines web a una mateixa captura.

També és necessària una anàlisi de resultats més profunda, especialment pel que fa al percentatge de precisió en tant que el model *Doc2Vec* es va quedant obsolet després d'una finestra de temps. En aquest sentit, no s'ha pogut fer donada la falta de recursos de computació dels quals es disposen (en la màquina en la qual s'han fet els experiments es triga de l'ordre de 36 hores a executar només l'eina de *getTraces* per a N = 30000).

Tanmateix, també resulta necessari una anàlisi més exhaustiva pel que fa a revisar amb quants *datasets* es pot obtenir el percentatge de precisió màxim.

Finalment, també podria ser útil la implementació d'un sistema que tingui emmagatzemat el model *Doc2Vec* i, en temps real, identifiqui els fluxos i a quina pàgina pertanyen. Probablement aquesta sigui l'eina que es pugui fer servir en producció a una empresa.

Capítol 8

Conclusions

En aquest projecte es proposava trobar una solució a la falta de mecanismes que tenen els proveïdors d'Internet per a identificar on es connecten els seus usuaris, cosa necessària per a poder millorar els seus serveis.

Així doncs, després d'haver introduït el problema, s'ha analitzat l'aplicació nDPI, una eina que pot servir per a aquest propòsit, tot tenint en compte les seves limitacions. També s'ha exposat un TFG anterior com a treball del qual parteix aquest projecte.

Després, s'ha exposat la metodologia que s'utilitza, la planificació inicial del projecte i les desviacions que hi ha hagut, seguida de l'anàlisi dels possibles actors implicats del projecte.

Un cop fet això, s'han exposat totes les alternatives que es van contemplar per a aquest projecte, així com la solució final que s'ha escollit, tot justificant-la. S'han descrit totes les eines desenvolupades i s'ha enfonsat en el seu funcionament.

Seguidament, s'ha elaborat una anàlisi de resultats, mitjançant el qual es pot concloure que l'eina és prou bona (arribant a un 98,66% de precisió a l'entorn de test i un 84,28% a l'entorn de validació). També s'ha detectat que els percentatges de precisió a l'entorn de test van baixant a mesura que es van afegint més *datasets* al model, però que en cap cas baixa per sota del 93,35% (amb sis *datasets*) al model, resultat el qual es pot qualificar de molt bo.

També s'evidencia que els percentatges de precisió a l'entorn de validació van pujant a mesura que es van afegint més *datasets* al model. Tot i això, com ja s'ha explicat a l'anàlisi de resultats i a l'apartat del treball futur, no es pot considerar la tònica general de l'eina, ja que els resultats ens indiquen que les pàgines tenen contingut dinàmic que va actualitzant-se a mesura que passa el temps. Així doncs, passat un temps és possible que el *dataset* més antic que s'hi hagi incorporat al model ja no sigui representatiu, i s'hagi de treure. Tanmateix, no s'ha pogut demostrar aquesta hipòtesi en aquest projecte, ja que per tal de fer-ho s'hauria d'expandir el projecte temporalment i recopilar molts *datasets* repartits en el temps.

Finalment, també s'han analitzat els costos que tindria una empresa en

desenvolupar aquest projecte de manera professional, així com la sostenibilitat en les seves tres dimensions: econòmica, ambiental i social. També s'han identificat les lleis i regulacions que podrien afectar el projecte en una hipotètica portada a producció.

Així doncs, es pot concloure que l'aproximació que el conjunt d'adreces IP de les *third-parties* (a més de la de la pàgina en qüestió) identifica una pàgina web en concret és certa, que s'ha pogut desenvolupar una prova de concepte, que s'han assolit tots els objectius del projecte de manera satisfactòria, i que s'hauria de continuar aprofundint en aquest àmbit per a poder arribar a un sistema en temps real que es pogués posar en producció.

Apèndix A

Competències tècniques de l'especialitat

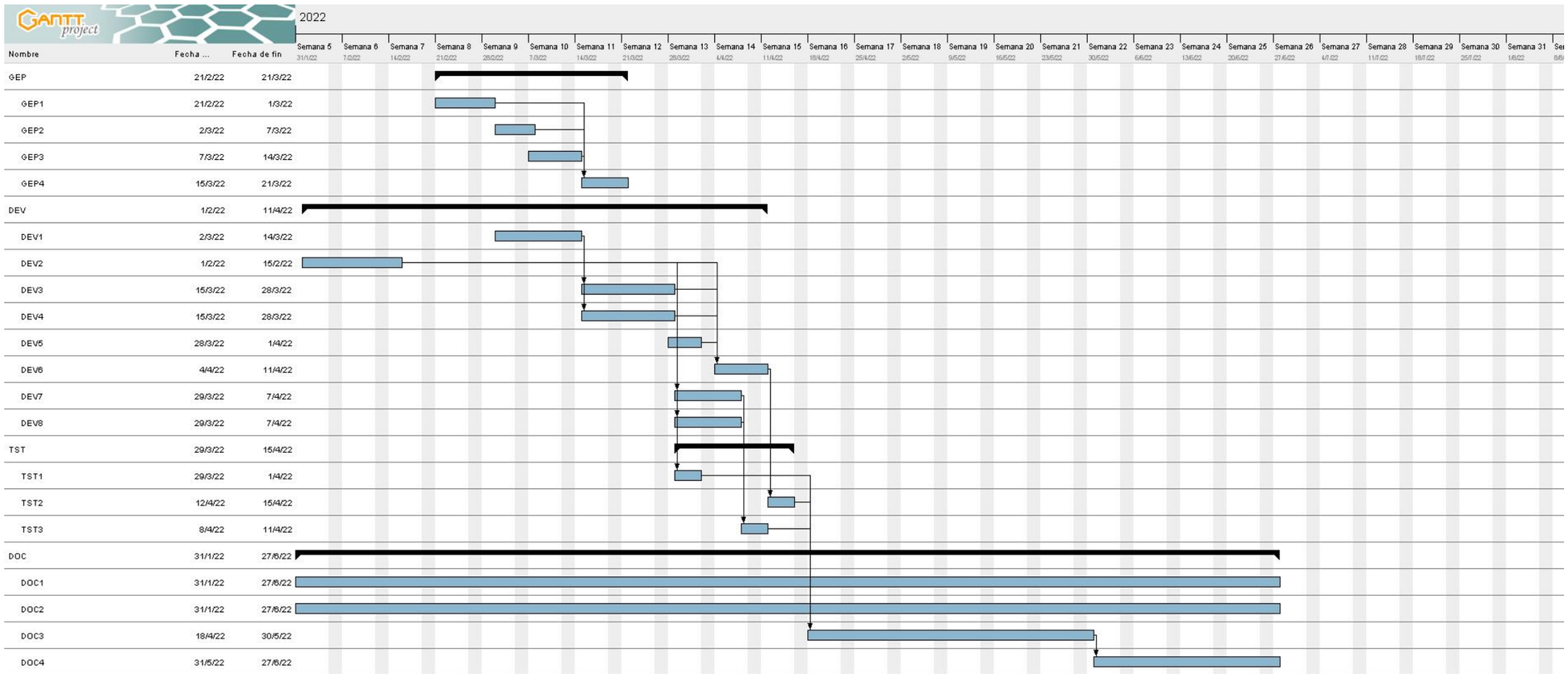
Tota la informació en aquest apèndix s'ha extret de [3].

- CTI1 Definir, planificar i gestionar la instal·lació de la infraestructura TIC de l'organització.
  - CTI1.1 Demostrar comprensió de l'entorn d'una organització i de les seves necessitats en l'àmbit de les tecnologies de la informació i les comunicacions.
  - CTI1.2 Seleccionar, dissenyar, desplegar, integrar gestionar xarxes i infraestructures de comunicacions en una organització.
  - CTI1.3 Seleccionar, desplegar, integrar i gestionar sistemes d'informació que satisfacin les necessitats de l'organització amb els criteris de cost i qualitat identificats.
  - CTI1.4 Seleccionar, dissenyar, desplegar, integrar, avaluar, construir, gestionar, explotar i mantenir les tecnologies de hardware, software i xarxes, dintre dels paràmetres de cost i qualitat adequats.
- CTI2 Garantir que els sistemes TIC d'una organització funcionen de manera adequada, són segurs i estan adequadament instal·lats, documentats, personalitzats, mantinguts, actualitzats i substituïts, i que les persones de l'organització reben un suport TIC correcte.
  - CTI2.1 Dirigir, planificar i coordinar la gestió de la infraestructura informàtica: hardware, software, xarxes i comunicacions.
  - CTI2.2 Administrar i mantenir aplicacions, sistemes informàtics i xarxes de computadors (els nivells de coneixement i de comprensió són a les competències tècniques comunes).
  - CTI2.3 Demostrar comprensió, aplicar i gestionar la garantia i la seguretat dels sistemes informàtics (CEIC6).

- CTI3 Dissenyar solucions que integrin tecnologies de hardware, software i comunicacions (i capacitat de desenvolupar solucions específiques de software de sistemes) per a sistemes distribuïts i dispositius de computació ubiqua.
  - CTI3.1 Concebre sistemes, aplicacions i serveis basats en tecnologies de xarxa, tenint en compte Internet, web, comerç electrònic, multimèdia, serveis interactius i computació ubiqua.
  - CTI3.2 Implementar i gestionar sistemes ubics (mobile computing systems).
  - CTI3.3 Dissenyar, implantar i configurar xarxes i serveis.
  - CTI3.4 Dissenyar software de comunicacions.
- CTI4 Emprar metodologies centrades en l'usuari i l'organització per al desenvolupament, l'avaluació i la gestió d'aplicacions i sistemes basats en tecnologies de la informació que assegurin l'accessibilitat, l'ergonomia i la usabilitat dels sistemes.

Apèndix B

Esquema de Gantt



Bibliografia

[1] ntop, "ndpi." [Online]. Available: https://www.ntop.org/products/deep-packet-inspection/ndpi/

[2] Google, "Cifrado https en la web," 1 2022. [Online]. Available: https://transparencyreport.google.com/https/overview

[3] FIB, "Competències tècniques de cada especialitat." [Online]. Available: https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/pla-destudis/especialitats/tecnologies-de-la-informacio#TI

[4] N. M. Córdoba, "Gestió de projectes. lliurable 4: Document final. treball de fi de grau: Identificació de pàgines web mitjançant captures passives a partir d'enllaços externs a un entorn xifrat," 3 2022.

[5] FIB, "Treball final de grau." [Online]. Available: https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/treball-de-fi-de-grau

[6] U. H. R. C. (38th sess. : 2018 : Geneva), "The promotion, protection and enjoyment of human rights on the internet : resolution / adopted by the human rights council on 5 july 2018," 7 2018. [Online]. Available: http://digitallibrary.un.org/record/1639840

[7] Cloudfare, "¿qué es una cdn?" [Online]. Available: https://www.cloudflare.com/es-es/learning/cdn/what-is-a-cdn/

[8] E. Rescorla, "Http over tls," pp. 4–5, 5 2000. [Online]. Available: doi.org/10.17487/rfc2818

[9] R. d'ADSLZone, "¿es posible saltarse el bloqueo de las webs prohibidas en españa?" 10 2021. [Online]. Available: https://www.adslzone.net/reportajes/seguridad/bloqueo-webs-espana/

[10] A. E. de Protección de Datos, "Dns privacy," pp. 9–10, 11 2019. [Online]. Available: https://www.aepd.es/sites/default/files/2019-12/nota-tecnica-privacidad-dns-en.pdf

[11] P. Emmerich, M. Pudelko, S. Gallenmuller, and G. Carle, "Flowscope: Efficient packet capture and storage in 100 gbit/s networks." IEEE, 6 2017, pp. 1–9.

[12] J. Lee, S. Lee, J. Lee, Y. Yi, and K. Park, "Flosis: A highly scalable network flow capture system for fast retrieval and storage efficiency." USENIX Association, 2015, pp. 445–457.

[13] T. S.A., "Telefónica research." [Online]. Available: https://www.telefonica.com/es/sostenibilidad-innovacion/innovacion/telefonica-research/

[14] thomasbhatia, "Opendpi," 2 2018. [Online]. Available: https://github.com/thomasbhatia/OpenDPI#readme

[15] "Wireshark." [Online]. Available: https://www.wireshark.org/

[16] ntop, "ndpi github," 6 2022. [Online]. Available: https://github.com/ntop/nDPI

[17] C. J. Bailén, "Identifiicación de servicios web a partir de tráfico cerrado," 2021. [Online]. Available: https://upcommons.upc.edu/handle/2117/344885#.YhoJPrTV0MI.mendeley

[18] E. Rescorla, "The transport layer security (tls) protocol version 1.3," 8 2018. [Online]. Available: doi.org/10.17487/rfc8446

[19] S. Burge, "An overview of the soft systems methodology," *System Thinking: Approaches and Methodologies*, pp. 1–14, 2015. [Online]. Available: https://eindhovenengine.nl/wp-content/uploads/2021/12/Soft-Systems-Methodology-source-2.pdf

[20] R. E. E. Freeman and J. McVea, "A stakeholder approach to strategic management," *SSRN Electronic Journal*, 2001. [Online]. Available: doi.org/10.2139/ssrn.263511

[21] TCPDUMP, "Home." [Online]. Available: https://www.tcpdump.org/

[22] S. F. Conservancy, "Selenium webdriver." [Online]. Available: https://www.selenium.dev/documentation/webdriver/

[23] Google, "Google chrome." [Online]. Available: https://www.google.com/intl/es_es/chrome/

[24] KiwiNewt, "pyshark github." [Online]. Available: https://github.com/KimiNewt/pyshark

[25] V. L. Pochat, T. V. Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation." Internet Society, 2019.

[26] R. developers, "Rsqlite." [Online]. Available: https://cran.r-project.org/web/packages/RSQLite/index.html

[27] Glassdoor, "¿cuánto gana un project manager?" 3 2022. [Online]. Available: https://www.glassdoor.es/Salaries/project-manager-salary-SRCH_KO0,15.htm?countryRedirect=true

[28] ——, "¿cuánto gana un knowledge manager?" 2 2022. [Online]. Available: https://www.glassdoor.es/Salaries/knowledge-manager-salary-SRCH_KO0,17.htm?countryRedirect=true

[29] ——, "¿cuánto gana un técnico de sistemas?" 3 2022. [Online]. Available: https://www.glassdoor.es/Sueldos/t%C3%A9cnico-de-sistemas-sueldo-SRCH_KO0,19.htm

[30] ——, "¿cuánto gana un programador?" 3 2022. [Online]. Available: https://www.glassdoor.es/Sueldos/programador-sueldo-SRCH_KO0,11.htm

[31] ——, "¿cuánto gana un quality assurance manager?" 3 2022. [Online]. Available: https://www.glassdoor.es/Sueldos/quality-assurance-manager-sueldo-SRCH_KO0,25.htm

[32] T. S. S. de lo Social, "Sts 6586/2007 - ecli:es:ts:2007:6586," 9 2007. [Online]. Available: https://www.poderjudicial.es/search/AN/openCDocument/e5e0cf323aea82eb34cd7e5fa7abd37c0e3d1a3bc62b49ea

[33] O. N. de Tecnología y Sociedad, "Flash datos de teletrabajo: tercer trimestre," 11 2021. [Online]. Available: https://www.ontsi.es/sites/ontsi/files/2021-12/flashdatosteletrabajotercertrimestre2021_1.pdf

[34] Lenovo, "Lenovo thinkpad t14 gen 2 (14ïntel)," 2021. [Online]. Available: https://www.lenovo.com/es/es/laptops/thinkpad/t-series/T14-G2-Intel/p/22TPT14T4N2

[35] R. Pereira, M. Couto, F. Ribeiro, R. Rua, J. Cunha, J. P. Fernandes, and J. Saraiva, "Energy efficiency across programming languages: how do energy, time, and memory relate?" ACM, 10 2017, pp. 256–267. [Online]. Available: doi.org/10.1145/3136014.3136031

[36] B. team, "Brltty."

[37] J. del Estado, "Ley orgánica 10/1995, de 23 de noviembre, del código penal." *Ley Orgánica*, vol. 10, 1995.