ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εύρεση κρυμμένων υπηρεσιών Tor με χρήση μηχανής αναζήτησης πολυμέσων

Φιλίτσα Αντωνίου Δόβα

Επιβλέπων: **Λάζαρος Μεράκος,** Καθηγητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εύρεση κρυμμένων υπηρεσιών Tor με χρήση μηχανής αναζήτησης πολυμέσων

Φιλίτσα Α. Δόβα
Α.Μ.: M1352

ΕΠΙΒΛΕΠΩΝ: **Λάζαρος Μεράκος,** Καθηγητής

Σεπτέμβριος 2017

ΠΕΡΙΛΗΨΗ

Το Tor είναι ένα παγκόσμιο δίκτυο από εξυπηρετητές (servers) που επιτρέπει στους χρήστες του να περιηγούνται ανώνυμα στο Διαδίκτυο. Το Tor σχεδιάστηκε για να δώσει την δυνατότητα σε εκατομμύρια ανθρώπους ανά τον κόσμο να απολαύσουν την ελευθερία της πρόσβασης και έκφρασης στο Διαδίκτυο, να προστατεύσουν τον εαυτό τους και τις πηγές τους, αλλά και τις επικοινωνίες τους. Παράλληλα, παρέχει την δυνατότητα των κρυμμένων υπηρεσιών (hidden services), δηλαδή την δημοσίευση ιστοτόπων (websites) χωρίς να αποκαλύπτεται η θέση τους ή η ταυτότητά τους. Η ανωνυμία που προσφέρει το δίκτυο Tor προσελκύει πλήθος παράνομων δραστηριοτήτων, όπως είναι η μαύρη αγορά ναρκωτικών και φαρμάκων, παράνομα όπλα, στρατιωτικές υπηρεσίες και σκληρή πορνογραφία.

Στα πλαίσια των προσπαθειών καταγραφής των κακόβουλων υπηρεσιών, με σκοπό την λήψη αντιμέτρων, οι μηχανές αναζήτησης αποτελούν ένα πολύτιμο εργαλείο. Όμως, οι παραδοσιακές μηχανές αναζήτησης περιορίζονται στην ανίχνευση κειμένων και η σχετικότητα των αποτελεσμάτων εξαρτάται από τις λέξεις-κλειδιά που θα χρησιμοποιηθούν. Μιας και στο δίκτυο Tor χρησιμοποιείται ιδιαίτερη διάλεκτος ή/και διαφορετικές γλώσσες, η επιλογή των κατάλληλων λέξεων-κλειδιά μπορεί να είναι πολύ προβληματική. Σε αυτή την εργασία προτείνεται μια μηχανή αναζήτησης πολυμέσων, η οποία μπορεί να προσφέρει στην ανακάλυψη νέων κακόβουλων κρυμμένων υπηρεσιών, με βάση τα πολυμέσα που φιλοξενούνται στις ιστοσελίδες τους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανές Αναζήτησης

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: εικόνες, πολυμέσα, κατακερματισμός, ανίχνευση

ABSTRACT

Tor is a worldwide network that allows people to browse the Internet anonymously. Additionally, it provides hidden services that allow users to publish websites and services without revealing their location or identity. It was designed to give freedom of speech and a chance to avoid Internet censorship to millions of people around the world. On the other hand, Tor's anonymity attracts suspicious and illegal activities, such as black market drugs, illegal firearms and extreme pornography.

In order to clean Tor's content and get countermeasures, the discovery of illegal services is required. To enumerate all illegal services, search engines can prove to be a valuable tool. Nevertheless, traditional search engines only rely on text and their results are limited by the keyword similarity. Document search engines cannot be effective enough, due to the fact that in Tor, a special set of non-standard words is being used, in order to obscure the meanings of texts. Hence, we introduce a multimedia search engine, which aids the discovery of illegal services based on the multimedia they host.

SUBJECT AREA: Search Engines

KEYWORDS: image, multimedia, hashing, crawl, spider

Στην οικογένειά μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον καθ. κ. Λάζαρο Μεράκο και τον καθ. κ. Χρήστο Ξενάκη για την πολύτιμη συμβολή τους στην ολοκλήρωση αυτής της εργασίας, για την υπομονή τους και την πολύ καλή μας συνεργασία.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια της ολοκλήρωσης του προγράμματος μεταπτυχιακών σπουδών «Νέες Τεχνολογίες Πληροφορικής και Επικοινωνιών» του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Καποδιστριακού Πανεπιστημίου Αθηνών.

1. ΕΙΣΑΓΩΓΗ

1.1 Ανωνυμία στο Διαδίκτυο

Το Tor είναι ένα παγκόσμιο δίκτυο από εξυπηρετητές (servers) που επιτρέπει στους χρήστες του να περιηγούνται ανώνυμα στο Διαδίκτυο. Παράλληλα, είναι ένας μη-κερδοσκοπικός οργανισμός που αποτελείται από εθελοντές, με κύριο σκοπό την έρευνα και υλοποίηση εργαλείων για επίτευξη ιδιωτικότητας (privacy). Το Tor σχεδιάστηκε για να δώσει την δυνατότητα σε εκατομμύρια ανθρώπους ανά τον κόσμο να απολαύσουν την ελευθερία της πρόσβασης και έκφρασης στο Διαδίκτυο, να προστατεύσουν τον εαυτό τους και τις πηγές τους, αλλά και τις επικοινωνίες τους. Τυπικά παραδείγματα των χρηστών του Tor σύμφωνα με το Tor Project [1] είναι:

- Απλοί άνθρωποι που επιθυμούν να προστατευτούν από ασυνείδητους εμπόρους και κλέφτες ηλεκτρονικών ταυτοτήτων. Οι Πάροχοι Υπηρεσιών Διαδικτύου (Internet Service Providers – ISPs) υποστηρίζουν ότι καθιστούν ανώνυμα τα στοιχεία αποκρύπτοντας τα προσωπικά δεδομένα, αλλά έχουν κυκλοφορήσει αρκετά άρθρα που αποδεικνύουν ότι κάτι τέτοιο δεν ισχύει. Επίσης, εκτός από τους ISPs, οι ίδιες οι ιστοσελίδες (websites) και οι μηχανές αναζήτησης που επισκεπτόμαστε, εν δυνάμει φυλάσσουν αρχεία καταγραφής συμβάντων (logs) με τα ίδια δεδομένα.
- Αυτοί που επιθυμούν να προστατεύσουν τις επικοινωνίες τους από ανεύθυνες εταιρίες. Προσωπικά δεδομένα βρίσκονται σε συστήματα εταιριών που είτε μπορούν να παραβιαστούν εύκολα, είτε που δίδονται σε ερευνητές, και εν τέλει καθιστούν τις εταιρίες μη-αξιόπιστες για την φύλαξή τους.
- Ιδιώτες ή δημοσιογράφοι που επιθυμούν να ερευνήσουν ευαίσθητα για το κράτος ζητήματα, μπορούν να αποφύγουν τον στιγματισμό και τις κυρώσεις [11].
- Οποιοσδήποτε θέλει να αποφύγει την διαδικτυακή λογοκρισία. Σε αρκετά μέρη του κόσμου βασικές λειτουργικότητες του Διαδικτύου, όπως είναι το Facebook και το Youtube, είναι απαγορευμένες.

Το Tor δίκτυο είναι εξίσου χρήσιμο εργαλείο και στα χέρια των οργανισμών επιβολής του νόμου:

- Αποκρύπτουν σε διαχειριστές αμφιβόλων ιστοσελίδων ότι μια κρατική IP τους επισκέπτεται, και έτσι δεν διακόπτεται η έρευνά τους.
- Επιτρέπει την μεταφορά της συγκεκαλυμμένης δράσης αστυνομικών (undercover operations) στον διαδικτυακό κόσμο.
- Επιτρέπει την συλλογή ανώνυμων καταγγελιών. Οι άνθρωποι ενθαρρύνονται να κάνουν καταγγελίες στις Αρχές, όταν ξέρουν ότι θα διατηρηθεί η ανωνυμία τους.

Μια άλλη λειτουργικότητα που προσφέρεται είναι οι κρυμμένες υπηρεσίες (hidden services), δηλαδή η δημοσίευση ιστοτόπων (websites) χωρίς να αποκαλύπτεται η θέση τους ή η ταυτότητά τους. Αυτή η λειτουργικότητα είναι ένας από τους πρωταρχικούς λόγους που δημιουργήθηκε το δίκτυο Tor, δηλαδή η διευκόλυνση εγκατάστασης υπηρεσιών επικοινωνίας για στρατιωτικούς σκοπούς, χωρίς την δυνατότητα ανακάλυψης των κεντρικών εξυπηρετητών και την αχρήστευσή τους.

1.2 Προβλήματα

“Η εξάπλωση του Διαδικτύου έχει οδηγήσει σε μια έκρηξη στην «αγορά» της παιδικής πορνογραφίας”, είναι η φράση – κλειδί της ετήσιας αναφοράς του 2016 του FBI προς το Κογκρέσο [16].

Παρ' όλο που το Tor σχεδιάστηκε για να δώσει την δυνατότητα σε εκατομμύρια ανθρώπους ανά τον κόσμο να απολαύσουν την ελευθερία της πρόσβασης και έκφρασης στο Διαδίκτυο, η έλλειψη ελέγχου και επιβολής του νόμου αποτέλεσε πρόσφορο έδαφος για την εγκληματικότητα, αποτελώντας πόλο έλξης παραβατικών συμπεριφορών, ενώ ταυτόχρονα έδωσε παγκόσμιο πεδίο δράσης στην ψηφιακή παρανομία. Η μαύρη αγορά φαρμάκων και ναρκωτικών είναι το πιο χαρακτηριστικό παράδειγμα παράνομων υπηρεσιών που προσφέρονται μέσα στο δίκτυο Tor, ενώ η λίστα συμπληρώνεται με εμπόριο όπλων, πληροφοριών, κακόβουλου λογισμικού (malware), εκτελεστών και η διακίνηση σκληρού πορνογραφικού υλικού. Πολλοί χρήστες του Tor αντιδρούν στην έξαρση των εγκληματικών στοιχείων, και κυρίως της υποστήριξης του δικτύου παραγωγής και διακίνησης παιδικής πορνογραφίας.

Η ανακάλυψη και η δημιουργία ενός ευρετηρίου με ιστότοπους που περιέχουν υλικό κακοποίησης παιδιών θα είχε πολλαπλά οφέλη τόσο στην προσπάθεια επιβολής του νόμου, όσο και στα ίδια τα θύματα.

Αναγνώριση θυμάτων: Στην ετήσια αναφορά του 2016 του FBI προς το Κογκρέσο, αναφέρεται πως η ανάλυση μιας και μόνο συγκεκριμένης ιστοσελίδας στο Tor, η οποία περιλάμβανε επίμαχο περιεχόμενο, απέφερε περίπου 1.3 εκατομμύρια εικόνες κακοποίησης παιδιών και αναγνώρισε 73 θύματα που δεν τους ήταν ήδη γνωστά. Η τροφοδοσία της υπηρεσίας με περισσότερους ιστοτόπους θα βοηθούσε ακόμη περισσότερο στην αναγνώριση θυμάτων.

Μείωση εξάπλωσης υλικού: Όπως αναφέρεται στην κεντρική ομιλία του Crimes Against Children Conference, τα θύματα φυσικής και σεξουαλικής βίας, πολύ περισσότερο από την κακοποίηση, φοβούνται ότι οι εικόνες τους θα συνεχίσουν να διαμοιράζονται στο Διαδίκτυο. Η ανακάλυψη του υλικού αυτού μπορεί να βοηθήσει στον περιορισμό της εξάπλωσης του.

Για να γίνει αποδοτικά η δημιουργία ενός τέτοιου ευρετηρίου, προτείνουμε την χρήση τεχνολογιών για αυτοματοποίηση της ανακάλυψης αλλά και της καταχώρησης παράνομου υλικού. Σκοπός αυτής της εργασίας είναι η μελέτη και η κατασκευή μιας μηχανής αναζήτησης, η οποία θα αναζητά σελίδες με βάση τα πολυμεσικά αρχεία που φιλοξενούνται στους ιστοτόπους των κρυμμένων υπηρεσιών.

1.3 Σχετικά συστήματα

Το Καναδικό Κέντρο Προστασίας του Παιδιού (Canadian Centre for Child Protection – C3P) και η υπηρεσία του "Cybertip.ca" έχει αναπτύξει ένα εργαλείο για να καταπολεμήσει τον πολλαπλασιασμό και την διάδοση του υλικού σεξουαλικής κακοποίησης παιδιών. Αποτελείται από έναν crawler που ονομάζεται Project Arachnid, και βοηθάει στην μείωση της διαθεσιμότητας αυτού του υλικού στο διαδίκτυο, με σκοπό να «σπάσει ο κύκλος της κακοποίησης». Ο στόχος του είναι να εντοπιστούν όλοι οι πάροχοι που προσφέρουν ένα πολυμεσικό υλικό, χωρίς την συγκατάθεση του ιδιοκτήτη ή υλικό ψηφιακού εκφοβισμού (bullying).

Το **Project Arachnid** ανιχνεύει τις συγκεκριμένες εικόνες και βίντεο, τροφοδοτώντας τον crawler με τα ψηφιακά αποτυπώματα του υλικού, το οποίο γίνεται με πολύ μεγαλύτερο ρυθμό σε σύγκριση με τις χειροκίνητες μεθόδους. Όταν αναγνωριστεί η ύπαρξη κάποιου αντιγράφου, ενημερώνονται οι πάροχοι του ιστοτόπου (hosting providers) και απαιτούν την άμεση απομάκρυνση του υλικού.

Η ανίχνευση γίνεται αποκλειστικά σε ιστοσελίδες και συνδέσμους οι οποίοι είναι ήδη καταχωρημένοι ως πάροχοι παράνομου υλικού, στις λίστες του "Cybertip.ca". Η μηχανή αναζήτησης που προτείνουμε θα δράσει θετικά στην αύξηση της λίστας των

ιστοσελίδων που υπάρχει επίμαχο υλικό, η οποία είναι πολύ σημαντική για την αποτελεσματική λειτουργία του Project Arachnid και κατ’ επέκταση της επιρροής του C3P.

2. ΥΠΑΡΧΟΥΣΕΣ ΤΕΧΝΟΛΟΓΙΕΣ

2.1 Tor

Το TOR (συντομογραφία του “The Onion Router”) είναι ένα παγκόσμιο δίκτυο που χρησιμοποιείται για την προστασία της ταυτότητας του χρήστη και επιτρέπει την ανώνυμη χρήση του Διαδικτύου ή την πρόσβαση σε σελίδες Tor (κρυμμένες υπηρεσίες). Το λογισμικό που είναι εγκατεστημένο από την πλευρά του πελάτη, δρομολογεί την διαδικτυακή κίνηση μέσω ενός παγκόσμιου δικτύου διακομιστών, οι οποίοι λειτουργούν εθελοντικά, με σκοπό να αποκρύψει την πραγματική τοποθεσία του χρήστη από οποιονδήποτε παρακολουθεί ή αναλύει την διαδικτυακή κίνηση.

Ο σκοπός του είναι να προστατεύσει την ατομική ελευθερία, την ιδιωτικότητα και την δυνατότητα διεξαγωγής εμπιστευτικών εργασιών χωρίς αυτές να καταγράφονται. Η βασική αρχή λειτουργίας του είναι η στρωματοποιημένη κρυπτογράφηση. Τα αρχικά δεδομένα επανακρυπτογραφούνται πολλές φορές, μία φορά σε κάθε Tor κόμβο που επισκέπτονται. Μια πρώτη έκδοση του λογισμικού με το δίκτυο των “onion” δρομολογητών ανακοινώθηκε τον Σεπτέμβριο του 2002 από τους Roger Dingledine, Nick Mathewson και Paul Syverson [1].

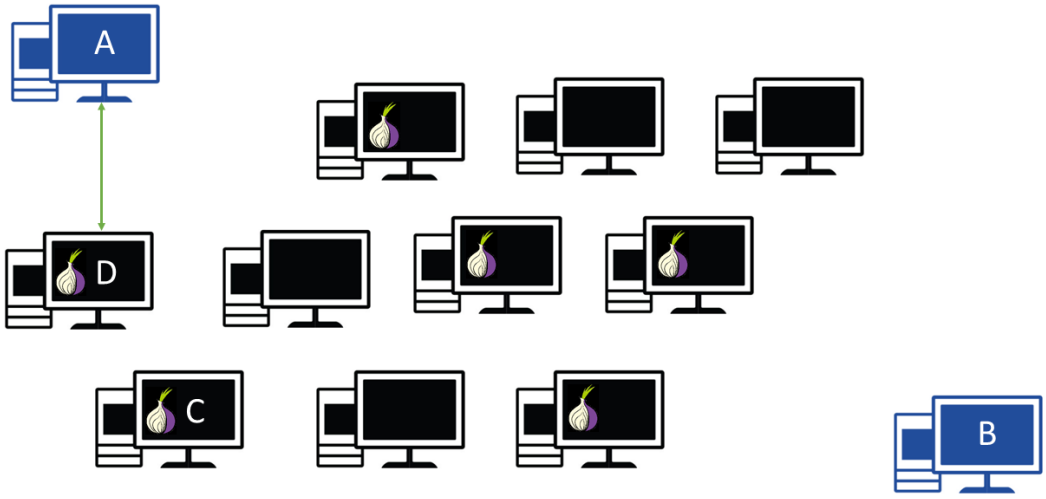
2.1.1 Λειτουργία

Η ιδέα πίσω από την λειτουργία του Tor είναι παρόμοια με την χρήση ενός διαστρεβλωμένου, δύσκολου στην παρακολούθηση μονοπατιού, για την αποφυγή κάποιου που ακολουθεί – ενώ ταυτόχρονα σβήνονται και τα ίχνη. Έτσι, αντί να υπάρχει απ’ ευθείας σύνδεση της πηγής με τον προορισμό, τα πακέτα στο Tor ακολουθούν ένα τυχαίο μονοπάτι ανάμεσα σε πολλούς εξυπηρετητές αναμετάδοσης (relay servers), με τέτοιο τρόπο ώστε κανείς παρατηρητής δεν μπορεί να καταλάβει από πού προήλθαν τα δεδομένα, ούτε για πού προορίζονται.

Για να δημιουργηθεί αυτό το ιδιωτικό μονοπάτι, το λογισμικό από την πλευρά του χρήστη, χτίζει σταδιακά ένα κύκλωμα από κρυπτογραφημένες συνδέσεις δια μέσου των εξυπηρετητών αναμετάδοσης. Το κύκλωμα αυτό επεκτείνεται κατά ένα «βήμα» κάθε φορά, και ο κάθε εξυπηρετητής αναμετάδοσης γνωρίζει μόνο τους άμεσους γείτονές του (προηγούμενο και επόμενο). Κανένας εξυπηρετητής αναμετάδοσης δεν έχει πλήρη γνώση του μονοπατιού που ακολουθούν τα δεδομένα. Για να επιτευχθεί αυτό, το λογισμικό στην πλευρά του χρήστη, διατηρεί ένα ξεχωριστό σετ κλειδιών κρυπτογράφησης για το κάθε βήμα του μονοπατιού. Έτσι, δεν μπορεί να χρησιμοποιηθεί ανάλυση κίνησης για να συνδεθεί η πηγή με τον προορισμό, ούτε από κάποιον που παρακολουθεί την κίνηση, ούτε από κάποιο κακόβουλο εξυπηρετητή αναμετάδοσης που συμμετέχει στο κύκλωμα. Μια ακόμα δικλείδα ασφαλείας που προσφέρεται είναι η ανανέωση των κυκλωμάτων κάθε 10 λεπτά. Έτσι, δεν μπορεί να υπάρξει σύνδεση μεταξύ παλιών ενεργειών με τις νέες.

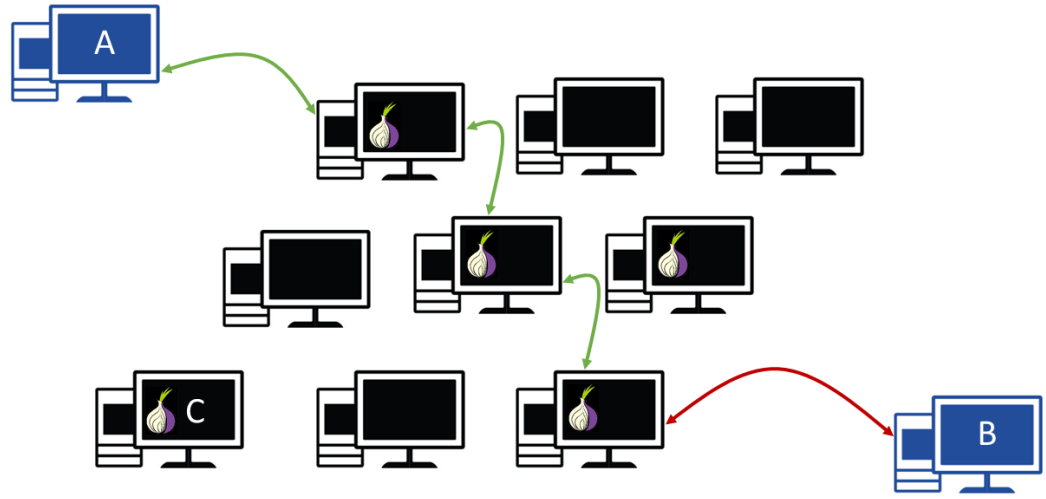
Με την εγκαθίδρυση του κυκλώματος, ξεκινάει και η επικοινωνία. Το Tor υποστηρίζει μόνο TCP ρεύματα, και μπορεί να χρησιμοποιηθεί από οποιαδήποτε εφαρμογή υποστηρίζει SOCKS (Socket Secure).

Αρχικά, ο υπολογιστής με το λογισμικό πελάτη A το οποίο επιθυμεί να συνδεθεί με το δίκτυο Tor, ζητά μια λίστα με τους κόμβους Tor από έναν εξυπηρετητή ευρετηρίου D (directory server).



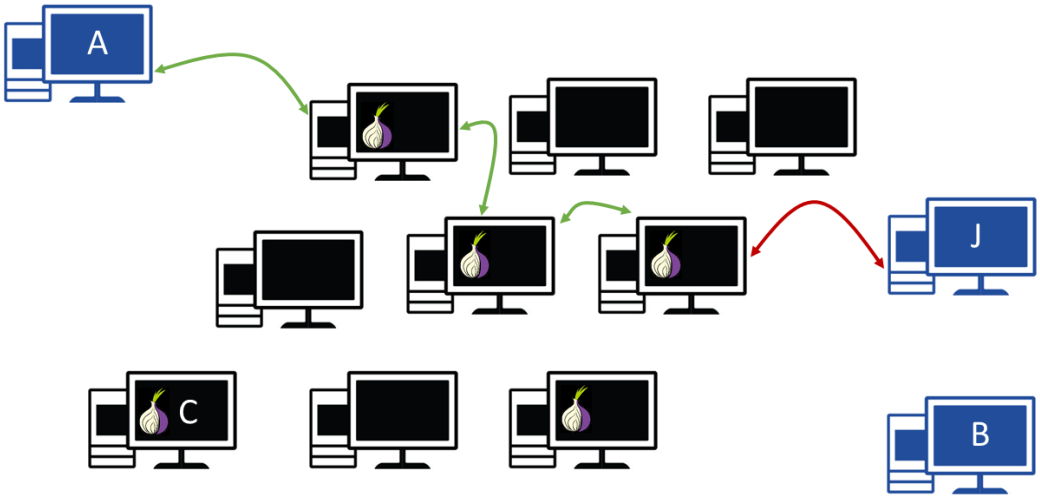
Σχήμα 1 – Αναζήτηση ευρετηρίου κόμβων Tor

Έπειτα, ο Α επιλέγει ένα τυχαίο μονοπάτι Tor κόμβων προς τον προορισμό Β. Τα πράσινα βέλη ορίζουν κρυπτογραφημένη επικοινωνία, ενώ τα κόκκινα αποτελούν μη-κρυπτογραφημένη επικοινωνία.



Σχήμα 2 – Πρόσβαση σε διαδικτυακό τόπο εκτός Tor

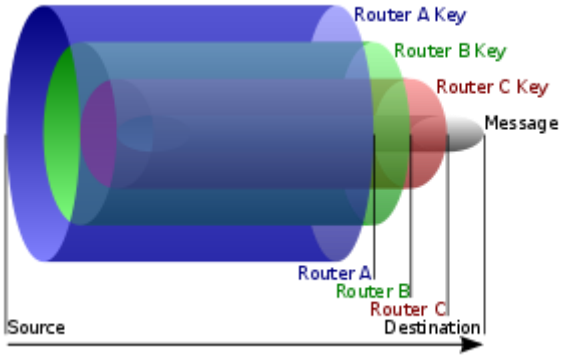
Αφού περάσει το χρονικό διάστημα των δέκα λεπτών, και εφ' όσον ο Α επιθυμεί να επισκεφθεί έναν νέο δικτυακό τόπο, το λογισμικό πελάτη του Α επιλέγει ένα δεύτερο τυχαίο μονοπάτι.



Σχήμα 3 – Πρόσβαση από διαφορετικό κύκλωμα

2.1.2 Δρομολόγηση

Η δρομολόγηση στο Tor γίνεται μέσω των εξυπηρετητών αναμετάδοσης, με μια τεχνική που ονομάζεται onion routing, η οποία ορίζει την προσθήκη ενός στρώματος κρυπτογράφησης σε κάθε βήμα της δρομολόγησης.



Εικόνα 1 – Κρυπτογράφηση πολλαπλών στρωμάτων (πηγή: Wikipedia)

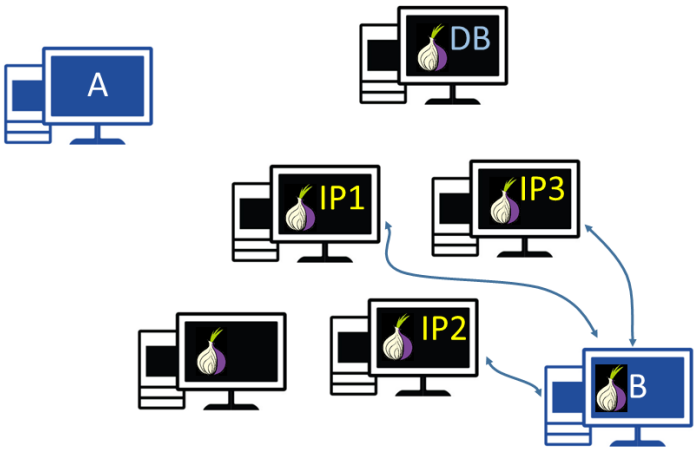
Με την τεχνική αυτή, εξασφαλίζεται τέλεια μυστικότητα προς τα εμπρός (perfect forward secrecy) μεταξύ των κόμβων, και κατ' επέκταση την ανωνυμία της πηγής και του προορισμού. Από πλευράς παραλήπτη (προορισμού), ως δημιουργός της επικοινωνίας φαίνεται ο τελευταίος κόμβος Tor (κόμβος εξόδου – exit node).

2.1.3 Κρυμμένες υπηρεσίες

Το Tor μπορεί επίσης να προσφέρει την δυνατότητα σε χρήστες να αποκρύψουν την θέση τους, ενώ προσφέρουν κάποια διαδικτυακή υπηρεσία. Οι υπηρεσίες αυτές, ονομάζονται κρυμμένες υπηρεσίες και είναι προσβάσιμες μόνο μέσω του δικτύου Tor. Αντί να αποκαλύπτουν τη διεύθυνση IP των εξυπηρετητών και άρα τη διαδικτυακή τοποθεσία τους, οι κρυμμένες υπηρεσίες είναι προσβάσιμες μέσω του Tor `.onion` ψευδό-όνομα-τομέα (pseudo top level domain – TLD). Το δίκτυο Tor αντιλαμβάνεται το

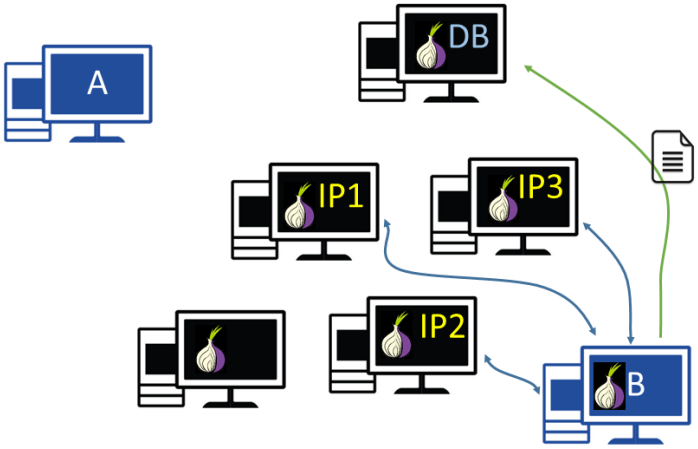
TLD και δρομολογεί δεδομένα από και προς τις κρυμμένες υπηρεσίες, ανώνυμα. Για να γίνει δυνατή η σύνδεση άλλων χρηστών του Tor στις υπηρεσίες αυτές, χρησιμοποιείται ένα άλλο είδος κόμβων που ονομάζονται “rendezvous”.

Έστω ότι ο εξυπηρετητής B θέλει να προσφέρει μια κρυμμένη υπηρεσία. Αρχικά, θα πρέπει να «διαφημίσει» την ύπαρξη μιας νέας κρυμμένης υπηρεσίας στο δίκτυο Tor, πριν να μπορέσουν οι πελάτες να συνδεθούν. Για να γίνει αυτό, ο εξυπηρετητής B επιλέγει τυχαία κάποιους εξυπηρετητές αναμετάδοσης, χτίζει κυκλώματα με αυτούς και τους ζητά να δράσουν ως σημεία εισαγωγής (introduction points – IP στο σχήμα), δίνοντας τους το δημόσιο κλειδί του.



Σχήμα 4 – Βήμα 1 Επιλογή σημείων εισόδου

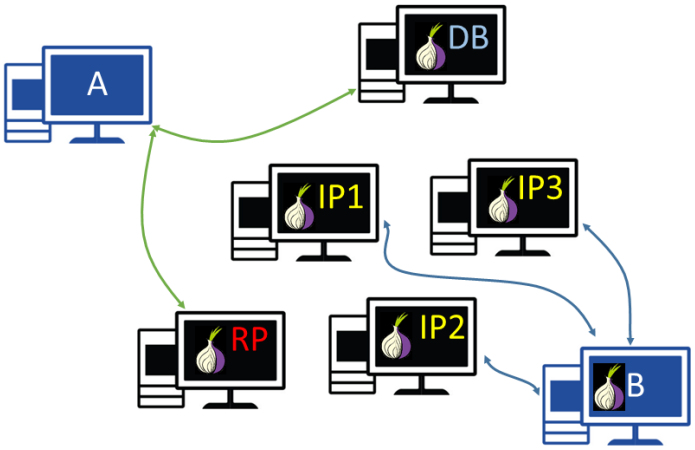
Σαν δεύτερο βήμα της διαδικασίας, ο B κατασκευάζει έναν περιγραφέα της υπηρεσίας (hidden service descriptor), ο οποίος περιέχει το δημόσιο κλειδί του, την σύνοψη των σημείων εισαγωγής και έπειτα το υπογράφει με το ιδιωτικό του κλειδί. Ο περιγραφέας αυτός προστίθεται σε έναν κατανεμημένο πίνακα κατακερματισμού. Πλέον, ο περιγραφέας είναι διαθέσιμος στους πελάτες που επιθυμούν να επισκεφθούν τον B, μέσω του 16ψήφιου χαρακτηριστικού ονόματος που έχει αποδοθεί στον B πχ `msydqstlz2kzerdg.onion`.



Σχήμα 5 – Βήμα 2 Δημοσίευση περιγραφέα

Στο τρίτο βήμα της διαδικασίας, ένας πελάτης A που θέλει να επισκεφθεί τον B, πρέπει πρώτα να μάθει την onion διεύθυνσή του. Αφού την μάθει, πιθανώς εξωσυστημικά, θα αναζητήσει με την διεύθυνση αυτή τον περιγραφέα στον πίνακα κατακερματισμού. Έτσι,

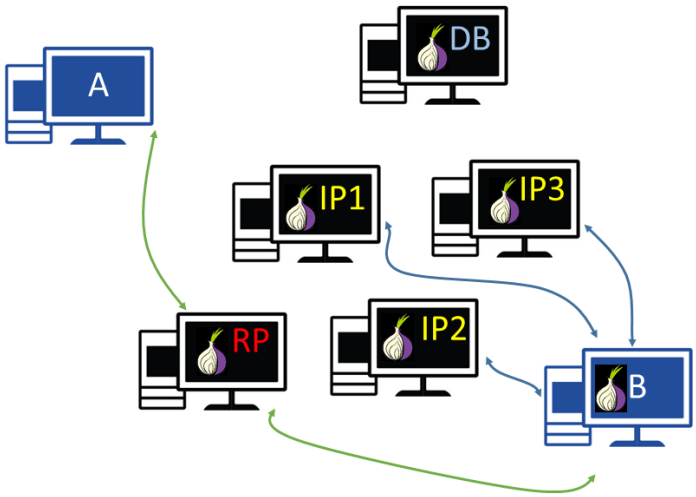
θα μάθει τα σημεία εισόδου και το δημόσιο κλειδί που θα πρέπει να χρησιμοποιήσει. Έπειτα, ο A θα εγκαθιδρύσει ένα κύκλωμα με έναν άλλο Tor κόμβο, ο οποίος θα δράσει ως “rendezvous” κόμβος.



Σχήμα 6 – Βήμα 3 Επιλογή Rendez-vous

Στο τέταρτο βήμα, ο A στέλνει ένα μήνυμα που περιέχει πληροφορία για τον κόμβο rendezvous, κρυπτογραφημένο με το δημόσιο κλειδί του B, σε ένα από τα σημεία εισόδου, ζητώντας να παραδοθεί στον B.

Στο πέμπτο βήμα, ο B αποκρυπτογραφεί το μήνυμα του A. Ο B δημιουργεί ένα κύκλωμα με τον rendezvous κόμβο.



Σχήμα 7 – Εγκαθίδρυση Tor σύνδεσης A με B

Στο τελευταίο βήμα, ο rendezvous κόμβος ειδοποιεί τον A για την επιτυχή εγκαθίδρυση σύνδεσης με τον B.

2.2 Επίπεδα Παγκόσμιου Δικτύου

Το Παγκόσμιο Δίκτυο (World Wide Web) φιλοξενεί δισεκατομμύρια σελίδες, όμως δεν έχουν όλες το ίδιο επίπεδο πρόσβασης. Οι σελίδες χωρίζονται σε διαφορετικά επίπεδα

Δικτύου, ανάλογα με την ευκολία εύρεσης και πρόσβασης σε αυτές. Τα σημαντικότερα επίπεδα είναι:

«Παραδοσιακό» Διαδίκτυο, ή αλλιώς Clearnet [4], είναι το κομμάτι του Διαδικτύου το οποίο μπορεί και θέλουμε να καταγραφεί και να είναι διαθέσιμο προς αναζητήσεις σε ευρετήρια (indices). Σε αυτό το κομμάτι του Διαδικτύου η ανωνυμία δεν επιδιώκεται. Αντίθετα, γίνεται καταγραφή πληροφοριών για κάθε χρήστη και χρησιμοποιείται συνήθως για την εξατομίκευση και βελτιστοποίηση των παρεχόμενων υπηρεσιών. Όλες οι σελίδες που συναντώνται στο Clearnet είναι διασυνδεδεμένες μεταξύ τους, και οι μηχανές αναζήτησης χρησιμοποιούν τις διασυνδέσεις αυτές (links) για να ολοκληρώσουν το ευρετήριο (index) τους.

Deep web τυπικά ονομάζεται το κομμάτι του Διαδικτύου που δεν μπορεί να εντοπιστεί από τις μηχανές αναζήτησης, όπως είναι FTP, ηλεκτρονικές βιβλιοθήκες, βάσεις δεδομένων κλπ. Ο λόγος που το περιεχόμενο δεν βρίσκεται σε ευρετήρια είναι ότι δεν υπάρχει απ' ευθείας σύνδεσμος προς αυτό, και συνήθως έχουμε πρόσβαση κατόπιν της συμπλήρωσης κάποιας φόρμας. Θα πρέπει λοιπόν κανείς να χρησιμοποιήσει την λειτουργικότητα της αναζήτησης που προσφέρει η ίδια η αρχική σελίδα, ώστε να έχει πρόσβαση στο υλικό, όπως συμβαίνει για παράδειγμα με την βιβλιοθήκη του Πανεπιστημίου.

Dark web είναι το κομμάτι του Deep web το οποίο έχει αποκρυφθεί επί τούτου και δεν είναι προσβάσιμο με κάποιον από τους συνηθισμένους φυλλομετρητές (browsers). Το πιο γνωστό δίκτυο που ανήκει στο Dark web είναι το Tor. Οι παραδοσιακές μηχανές αναζήτησης δεν επισκέπτονται και δεν ανιχνεύουν σελίδες του.

2.3 Crawlers Διαδικτύου

Οι crawlers, που επίσης ονομάζονται και robots, bots ή spiders, είναι προγράμματα που χρησιμοποιούνται από τις μηχανές αναζήτησης για να εξερευνήσουν το Διαδίκτυο και να κατεβάσουν, με αυτοματοποιημένο τρόπο, το περιεχόμενο όλων των ιστοσελίδων που επισκέπτονται. Ταυτόχρονα, αναζητούν συνδέσμους μέσα στο περιεχόμενο, ώστε να συνεχίσουν την αναζήτησή τους σε καινούριες σελίδες. Πρόκειται δηλαδή για ένα λογισμικό που πραγματοποιεί συστηματική σάρωση στο Διαδίκτυο, και αποθηκεύει το περιεχόμενο σε μια βάση δεδομένων. Οι πιο γνωστοί crawlers ανοιχτού κώδικα είναι μεταξύ άλλων ο Apache Nutch, ο StormCrawler και το Scrapy. Η απαίτηση για δημιουργία ενός εξειδικευμένου crawler μπορεί να καλυφθεί με την χρήση κάποιου από τα πλαίσια (frameworks) ανοιχτού κώδικα για αυτό τον σκοπό.

Ο Nutch είναι πολύ επεκτάσιμος και κλιμακώσιμος crawler. Είναι βελτιστοποιημένος για να τρέχει σε κατανεμημένες εφαρμογές και σε πολλά μηχανήματα, όπως είναι τα Amazon Web Services. Έχει αρκετά πολύπλοκη αρχιτεκτονική, η οποία χρησιμοποιεί πολλές εξαρτήσεις από τεχνολογίες όπως Ant, Gora, HBase, Hadoop κλπ και είναι γραμμένο σε Java. Είναι πιο αποδοτικός σε αναζητήσεις ενός συγκεκριμένου στόχου, όπως για παράδειγμα οι ιστοσελίδες ειδήσεων, παρά στις ευρείες αναζητήσεις.

Ο StormCrawler είναι στην πράξη μια συλλογή από εργαλεία για την δημιουργία crawlers στο Apache Storm, το οποίο είναι ένα κατανεμημένο σύστημα υπολογισμού πραγματικού χρόνου.

Το Scrapy είναι ένα εύρωστο framework ανοιχτού κώδικα, που έχει σχεδιαστεί για την σάρωση δεδομένων από ποικίλες πηγές. Συγκεκριμένα, στην σάρωση ιστοσελίδων, εκμεταλλευόμενο τον σχεδιασμό του με βάση τα συμβάντα (event-based), μπορεί να εκτελεί πολλές κλήσεις (requests) ταυτόχρονα στο ίδιο νήμα (thread), μέχρις ότου κάποια δημιουργήσει κάποιο συμβάν (event), για παράδειγμα άφιξη απάντησης (HTML

response). Οι απαιτήσεις σε μνήμη είναι ανάλογες του μεγέθους των δεδομένων που διαχειρίζεται το μοναδικό του νήμα, σε αντίθεση με τις πολυνηματικές (multithread) εφαρμογές, που το κάθε νήμα προσθέτει σημαντική επιβάρυνση. Με λίγα λόγια, η σχεδίαση του Scrapy διαπρέπει όταν έχουμε να κάνουμε με απρόβλεπτες ιστοσελίδες, βάσεις δεδομένων και APIs, μιας και η καθυστέρηση τους σε απάντηση των HTTP κλήσεων δεν επηρεάζει σημαντικά την συνολική απόδοση. Μέσα στα πλαίσια του Scrapy, είναι δυνατή η ταχεία ανάπτυξη και εκτέλεση αυτόνομων crawlers.

2.4 Ανάλυση δεδομένων “Big Data”

Ο όρος “Big Data” αναφέρεται στις συλλογές δεδομένων που είναι τόσο μεγάλες και πολύπλοκες που δεν μπορούν να τις διαχειριστούν οι παραδοσιακές μηχανές επεξεργασίας δεδομένων. Οι προκλήσεις που συνοδεύουν μεγάλους όγκους δεδομένων αγγίζουν τομείς όπως η αποθήκευση, η ανάλυση, η δημιουργία ερωτημάτων και η αναζήτηση, η οπτικοποίηση, αλλά και η μεταφορά τους. Οι συλλογές αυτές έχουν μεγάλη αξία στον τομέα της αναζήτησης στο Διαδίκτυο, καθώς μπορούν να εφαρμοστούν πάνω τους τεχνικές προγνωστικής ανάλυσης (predictive analytics), ανάλυσης συμπεριφορών των ατόμων και γενικά προχωρημένες τεχνικές ανάλυσης, που επιτρέπουν την αναγνώριση μοτίβων (patterns) συσχετισμών, ώστε να εκμαιευτούν πληροφορίες.

Το Elasticsearch είναι ένα εργαλείο που αξιοποιεί την πιο εξελιγμένη βιβλιοθήκη ανάκτησης πληροφοριών για Big Data, το Apache Lucene, αποκρύπτοντας την πολυπλοκότητά της και δίνοντας στον χρήστη απλά εργαλεία αλληλεπίδρασης, μέσω ενός REST API. Κάποια από τα πλεονεκτήματα αυτού του εργαλείου είναι ότι είναι κατανεμημένο και κλιμακώσιμο, μπορεί να υποστηρίξει δομημένα ή αδόμητα δεδομένα και η δυνατότητα του να παρέχει αναζήτηση για κάθε τους πεδίο. Τα χαρακτηριστικά αυτά το κάνουν κατάλληλο για υλοποίηση μηχανών αναζήτησης, μιας και βελτιστοποιεί τους χρόνους απόκρισης τους, ενώ ταυτόχρονα μπορεί να διαχειριστεί μεγάλο όγκο δεδομένων, μέσω των μηχανισμών αναπαραγωγής (replication) και την κατανομή των δεδομένων σε πολλαπλούς κόμβους.

2.5 Υπογραφή εικόνας

Η αυξημένη διαθεσιμότητα πολυμέσων σε ψηφιακή μορφή στην εποχή της πληροφορίας, έχει οδηγήσει σε τεράστια αύξηση των εργαλείων για την διαχείριση τους. Για να διασφαλιστεί η αξιοπιστία, έχουν αναπτυχθεί τεχνικές πιστοποίησης πολυμέσων για την επαλήθευση του περιεχομένου τους και την αποτροπή της πλαστογράφησης. Εν γένει, τα προβλήματα ακεραιότητας δεδομένων αντιμετωπίζονται με συναρτήσεις κατακερματισμού, οι οποίες όμως είναι ευαίσθητες σε μικρές αλλαγές στην είσοδό τους, και επιφέρουν μεγάλες διαφορές στο αποτέλεσμα. Επομένως, οι συναρτήσεις αυτές είναι χρήσιμες όταν δεν έχει αλλοιωθεί ούτε στο ελάχιστο το αρχικό δεδομένο. Ενώ αυτή η ευαισθησία ανταποκρίνεται στην ανάγκη επαλήθευσης δεδομένων τύπου κειμένου, δεν είναι αρκετό για τον ορισμό της αυθεντικότητας στα πολυμέσα. Τα δεδομένα πολυμέσων συνήθως έχουν υποστεί μικρές αλλοιώσεις, όπως μείωση της ποιότητας, μέτρια επίπεδα φιλτραρίσματος, γεωμετρική παραμόρφωση και θόρυβο. Επομένως, η επαλήθευση ανά bit δεν είναι ικανοποιητική μέθοδος για την εξακρίβωση της ταυτότητας. Αντίθετα, ένα εργαλείο επικύρωσης με βάση το περιεχόμενο είναι πιο κατάλληλο, όπως για παράδειγμα η τεχνική αντιληπτικού κατακερματισμού (perceptual hashing) [5] που παράγει ένα είδος υπογραφής για μια πολυμεσικό δεδομένο.

Ένας απλός αλγόριθμος κατακερματισμού εικόνας με χρήση perceptual hashing, είναι αυτός που παρουσιάζεται στην δουλειά των Chi Wong, Bern και Goldberg [6]. Με βάση αυτόν τον αλγόριθμο, παρακάτω παρουσιάζεται η υλοποίησή του στην δική μας περίπτωση.

Έστω ότι έχουμε την εικόνα:

Εικόνα 2 – Αρχική εικόνα

Βήμα 1: Αφαίρεση χρωματικής πληροφορίας. Από την εικόνα θα αφαιρεθούν τα χρώματα, θα γίνει δηλαδή ασπρόμαυρη.

Εικόνα 3 – Βήμα 1 της επεξεργασίας

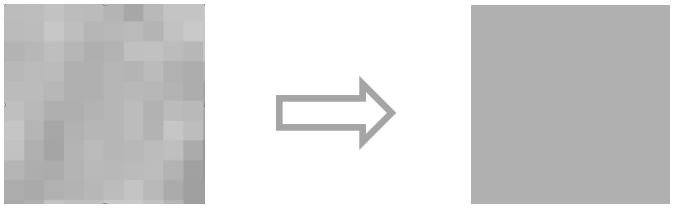
Βήμα 2: Μείωση δεδομένων, με τη χρήση μιας μάσκας σε μορφή πλέγματος. Θα παραμείνουν μόνο οι περιοχές που βρίσκονται μέσα σε κουτάκια.

Εικόνα 4 – Βήμα 2 της επεξεργασίας

Βήμα 3: Αφαιρούμε το πλέγμα και κρατάμε μόνο το δείγμα της εικόνας.

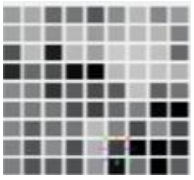
Εικόνα 5 – Βήμα 3 της επεξεργασίας

Βήμα 4: Αντικαθιστούμε κάθε κουτάκι με το μέσο χρώμα των pixels του



Εικόνα 6 – Βήμα 4 της επεξεργασίας

Βήμα 5: Δημιουργία διανυσμάτων γραμμής:



Εικόνα 7 – Βήμα 5 της επεξεργασίας

Για το κάθε pixel, ξεκινώντας από τον πάνω αριστερά γείτονα του και με φορά προς τα δεξιά, διαμορφώνουμε τα διανύσματα γραμμής, αναθέτοντας σε κάθε γείτονα μια τιμή ανάλογα με την διαφορά της απόχρωσής του, όπως ορίζεται στον Πίνακα 1.

Πίνακας 1 – Αντιστοιχίες απόχρωσης-τιμών

Απόχρωση	Τιμή
Πολύ πιο ανοιχτή	-2
Λίγο πιο ανοιχτή	-1
Παρόμοια	0
Λίγο πιο σκούρα	1
Πολύ πιο σκούρα	2

Έτσι, προκύπτουν για κάθε pixel διανύσματα της μορφής:

```
[ 1, 0, 1, 2, 1, 2, 2, 2]

[-2,-2,-1,-2, 2,-2, 2,-2]

[ 1, 0, 1, 2, 1, 2, 2, 2]
```

...

Για κάθε pixel έχουμε ένα διάνυσμα 8 διαστάσεων, αφού έχουμε μία τιμή για κάθε έναν γείτονα του. Συνδυάζοντας σε ένα διάνυσμα όλα 9x9 διανύσματα για κάθε pixel, έχουμε την υπογραφή της εικόνας, η οποία είναι τελικά ένα διάνυσμα 9x9x8, δηλαδή 648 διαστάσεων:

```
[ 1, 0, 1, 2, 1, 2, 2, 2,

-2,-2,-1,-2, 2,-2, 2,-2,

 1, 0, 1, 2, 1, 2, 2, 2,

            …

 1, 0, 1, 2, 1, 2, 2, 2,

 1, 0, 1, 2, 1, 2, 2, 2]
```

Η τεχνική του perceptual hashing είναι διαδεδομένη σε εφαρμογές του ψηφιακού κόσμου, όπως η προστασία πνευματικών δικαιωμάτων, αναζήτηση ομοιότητας ψηφιακών δεδομένων και ψηφιακή εγκληματολογία [10].

2.6 Μηχανές αναζήτησης

Οι μηχανές αναζήτησης στο Διαδίκτυο είναι οι εφαρμογές που φέρνουν σε επαφή τον χρήστη με το περιεχόμενο που αναζητά. Ενώ υπάρχουν πολλές μηχανές αναζήτησης διαθέσιμες, υπάρχουν τρία βασικά στάδια που ακολουθούν όλες, είτε πρόκειται για αναζήτηση κειμένου ή εν γένει πολυμέσων: η ανίχνευση (crawling), η τοποθέτηση σε ευρετήρια (indexing) και η ανάκτηση.

Crawling είναι η αρχική διαδικασία, κατά την οποία οι crawlers ή spiders επισκέπτονται σειριακά όλους τους συνδέσμους μιας ιστοσελίδας. Για κάθε σελίδα που επισκέπτονται, κατεβάζουν το περιεχόμενο, αν χρειάζεται το επεξεργάζονται, και το παραδίδουν στο επόμενο στάδιο, συνεχίζοντας με την ανάλυση του επόμενου συνδέσμου στην ουρά.

Indexing είναι η διαδικασία της αποθήκευσης της πληροφορίας του crawling σε βάσεις δεδομένων, και δημιουργία ευρετηρίων που επιτρέπουν την εύκολη και γρήγορη ανάκτησή της.

Ανάκτηση και ταξινόμηση αποτελεσμάτων είναι το τελευταίο στάδιο, στο οποίο υπάρχει είσοδος από τον χρήστη. Αρχικά δίνει ένα ερώτημα (query) στην μηχανή αναζήτησης, η οποία πρέπει να ανατρέξει στα ευρετήρια και να ανακτήσει σχετικό περιεχόμενο. Έπειτα, το περιεχόμενο που θα κριθεί σχετικό, πρέπει να παρουσιαστεί ταξινομημένο (από το πιο σχετικό στο λιγότερο). Ο αλγόριθμος υπολογισμού της σχετικότητας είναι επίσης πολύ σημαντικός στην αποτελεσματικότητα της μηχανής αναζήτησης.

2.6.1 Μηχανές αντίστροφης αναζήτησης εικόνων

Μια από τις πιο διαδεδομένες μηχανές αναζήτησης εικόνων είναι το Google Image Search. Η εν λόγω μηχανή δίνει την δυνατότητα στον χρήστη να δώσει μια εικόνα ως ερώτημα, και έπειτα κάνει μια εικασία σχετικά με το περιεχόμενο της εικόνας, ώστε να της αποδώσει μια περιγραφή σε μορφή κειμένου. Στη συνέχεια, χρησιμοποιεί αυτή την

περιγραφή ώστε να ψάξει παρόμοιες εικόνες, ενώ ταυτόχρονα διατηρεί τα χρωματικά κριτήρια της αρχικής εικόνας. Ο αλγόριθμος που χρησιμοποιείται ορίζει την ομοιότητα με βάση το περιεχόμενο της εικόνας και όχι την αντίληψη.

Σε αντίθεση με την Google Image Search, μια άλλη μηχανή αναζήτησης που χρησιμοποιεί αντιληπτική ομοιότητα, είναι η TinEye. Η μηχανή αυτή χρησιμοποιεί αποκλειστικά αναγνώριση εικόνας με βάση την υπογραφή της (δηλαδή το αποτέλεσμα του αλγορίθμου perceptual hashing) και όχι λέξεις-κλειδιά ή άλλα μεταδεδομένα (metadata).

2.6.2 Μηχανές αναζήτησης στο Tor

Μέσα στο δίκτυο Tor, υπάρχουν αρκετές μηχανές αναζήτησης κειμένου που προσφέρονται ως κρυμμένες υπηρεσίες ή και ως «πύλη» από το Clearnet προς το Darknet. Συνοπτικά, οι πιο σημαντικές παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 2 – Μηχανές αναζήτησης στο Tor

	Όνομα	**Περιγραφή**	**Υποστήριξη εικόνων**
1	TORCH	Μηχανή αναζήτησης κειμένου.	Όχι.
2	notEvil	Μηχανή αναζήτησης κειμένου.	Όχι.
3	DuckDuckGo	Hidden service της αντίστοιχης μηχανής αναζήτησης του Clearnet. Επιστρέφει αποτελέσματα από το Clearnet (κυρίως), αλλά και άλλα hidden services.	Ναι, με βάση το κείμενο.
4	Visitor	Μηχανή αναζήτησης κειμένου. Ευρετήριο συνδέσμων.	Όχι.
5	Candle	Μηχανή αναζήτησης κειμένου.	Όχι.
6	gram	Μηχανή αναζήτησης προϊόντων στις online αγορές του Tor.	Όχι.
7	Deep search	Εφαρμογή του Clearnet που αναζητά αποτελέσματα στο Tor.	Ναι, με βάση το κείμενο.
8	Ahmia	Εφαρμογή του Clearnet που αναζητά αποτελέσματα στο Tor. Υποστηρίζεται και το αντίστοιχο hidden service.	Όχι.

2.7 Προβλήματα

Ο σκοπός μας είναι η δημιουργία ενός εργαλείου που θα ανακαλύπτει κρυμμένες υπηρεσίες με βάση τα πολυμέσα που φιλοξενούν στις ιστοσελίδες τους. Οι υπάρχουσες λύσεις που σχετίζονται με την αναζήτηση στο Διαδίκτυο, εμπεριέχουν περιορισμούς

που δεν επιτρέπουν την χρήση τους για την κάλυψη της ανάγκης για εύρεση κρυμμένων υπηρεσιών.

Οι μηχανές αντίστροφης αναζήτησης εικόνων που είναι διαθέσιμες στο Clearnet, είναι κλειστού κώδικα. Επομένως, δεν υπάρχει τρόπος να επεκταθεί η λειτουργικότητά, πιθανώς τους με την προσθήκη κάποιας μονάδας επεξεργασίας, ώστε να υποστηρίξουν την ανακάλυψη και το indexing πολυμεσικών αρχείων. Επίσης, το πεδίο στο οποίο δραστηριοποιούνται και δημιουργούν τα ευρετήριά τους, είναι αποκλειστικά το Clearnet. Έτσι, η αναζήτηση υλικού στο Darknet μέσω αυτών των μηχανών δεν είναι δυνατή. Από την άλλη πλευρά, οι μηχανές αναζήτησης που δραστηριοποιούνται στο Darknet, δεν μπορούν να καλύψουν την ανάγκη για αναζήτηση με βάση κάποιο πολυμεσικό αρχείο, πχ εικόνα.

Επομένως, καμία από τις υπάρχουσες λύσεις δεν θα μπορούσε να χρησιμοποιηθεί ώστε να επιτευχθεί ο σκοπός μας. Έτσι, σχεδιάστηκε και κατασκευάστηκε ένα νέο εργαλείο, το οποίο συνδυάζει λειτουργικότητα μηχανής αναζήτησης με υποστήριξη εύκολης προσθήκης νέων μονάδων επεξεργασίας πολυμεσικών μέσων και πεδίο δραστηριοποίησης το Clearnet και το Darknet.

3. Η ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ RISE

Στα πλαίσια της παρούσας εργασίας, έχει κατασκευαστεί μια νέα μηχανή αντίστροφης αναζήτησης πολυμέσων, η οποία συγκεντρώνει περιεχόμενο από το Tor και διαθέτει μια διεπαφή για επικοινωνία με τον χρήστη και εμφάνιση των αποτελεσμάτων. Ονομάστηκε RISE από τα αρχικά Reverse Image Search Engine (μηχανή αντίστροφης αναζήτησης εικόνων). Με τη χρήση αυτού του εργαλείου, η ανακάλυψη επίμαχου υλικού θα γίνει πολύ πιο γρήγορη και αποδοτική, συνεισφέροντας στο έργο των αρχών.

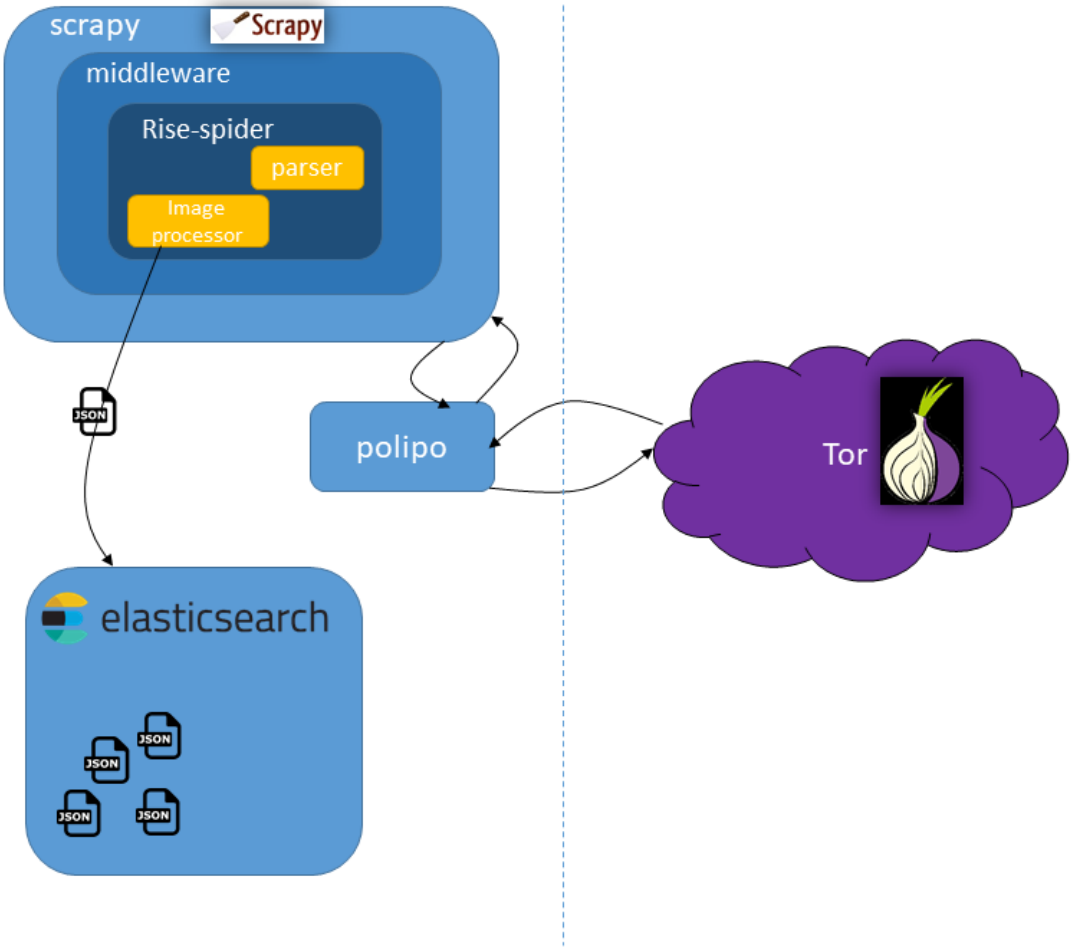
Απ' όσο γνωρίζουμε, κατά την διάρκεια συγγραφής της παρούσας εργασίας, δεν υπάρχει κάποια άλλη πλατφόρμα στην οποία να γίνεται εύκολα η εκτέλεση αλγορίθμων αναζήτησης. Το εργαλείο που παρουσιάζουμε μπορεί να ελαχιστοποιήσει τον χρόνο εκτέλεσης ενός νέου αλγόριθμου, βοηθώντας σημαντικά στην έρευνα και την ανάπτυξη νέων μεθόδων αναζήτησης πολυμέσων.

Τα ποιοτικά της χαρακτηριστικά που οδήγησαν στην επιλογή των τεχνολογιών της αρχιτεκτονικής, παρουσιάζονται στην συνέχεια:

Επεκτάσιμη για νέους τύπους πολυμέσων ή βελτίωση των υπαρχόντων αλγορίθμων. Η μηχανή αυτή έχει σχεδιαστεί χρησιμοποιώντας ανεξάρτητες μονάδες επεξεργασίας του υλικού που συγκεντρώνεται κατά τη φάση του crawling. Στην παρούσα υλοποίηση, έχει αναπτυχθεί μια μονάδα για επεξεργασία και κατακερματισμό εικόνων. Ο σχεδιασμός επιτρέπει την προσθήκη μονάδων για επεξεργασία διαφορετικών τύπων πολυμέσων (όπως βίντεο ή ήχος) ή ακόμα και την χρήση διαφορετικών αλγορίθμων κατακερματισμού για υπάρχοντες τύπους πολυμέσων. Για να επιτευχθεί αυτό, έχει αξιοποιηθεί η δυνατότητα εύκολης προσθήκης μονάδων του Scrapy Framework.

Με ελάχιστες απαιτήσεις χώρου. Εκ κατασκευής, τα αρχεία πολυμέσων είναι σημαντικά μεγαλύτερα σε μέγεθος απ' ότι τα αρχεία κειμένου. Κατά την φάση του indexing, θα ήταν χωρικά ασύμφορο να αποθηκευτεί αυτούσιο το αρχείο πολυμέσου, καθώς σύντομα θα βρισκόμασταν σε έλλειψη του «ακριβού» αποθηκευτικού χώρου. Η συμπίεση των αρχείων θα μπορούσε να λύσει προσωρινά το πρόβλημα, αλλά στην πραγματικότητα μπορούμε να αποφύγουμε εντελώς την αποθήκευση του ίδιου του αρχείου. Αντ' αυτού μπορούμε να κρατήσουμε έναν πολύ μικρότερο «περιγραφέα», ο οποίος το χαρακτηρίζει μοναδικά, με την χρήση ενός αλγορίθμου κατακερματισμού. Έτσι, ο χώρος που απαιτείται πλέον για την αποθήκευση της πληροφορίας είναι πολύ μικρότερος και ανάλογος του πλήθους των αρχείων, αφού οι αλγόριθμοι κατακερματισμού παράγουν έξοδο σταθερού μεγέθους ανεξάρτητα του μεγέθους της εισόδου τους.

Κλιμακώσιμη στην αύξηση των δεδομένων. Η μηχανή αναζήτησης αυξάνει διαρκώς τα δεδομένα που έχει επεξεργαστεί και εισαγάγει σε ευρετήρια. Επίσης, το πεδίο αναζήτησής της αυξάνεται διαρκώς, καθώς νέες δικτυακές υπηρεσίες (όπως ιστοσελίδες) αναρτώνται καθημερινά. Για το λόγο αυτό, έχει χρησιμοποιηθεί σύστημα αποθήκευσης και ανάκτησης πληροφορίας το οποίο είναι κατανεμημένο και κατάλληλο για επεξεργασία Big Data.



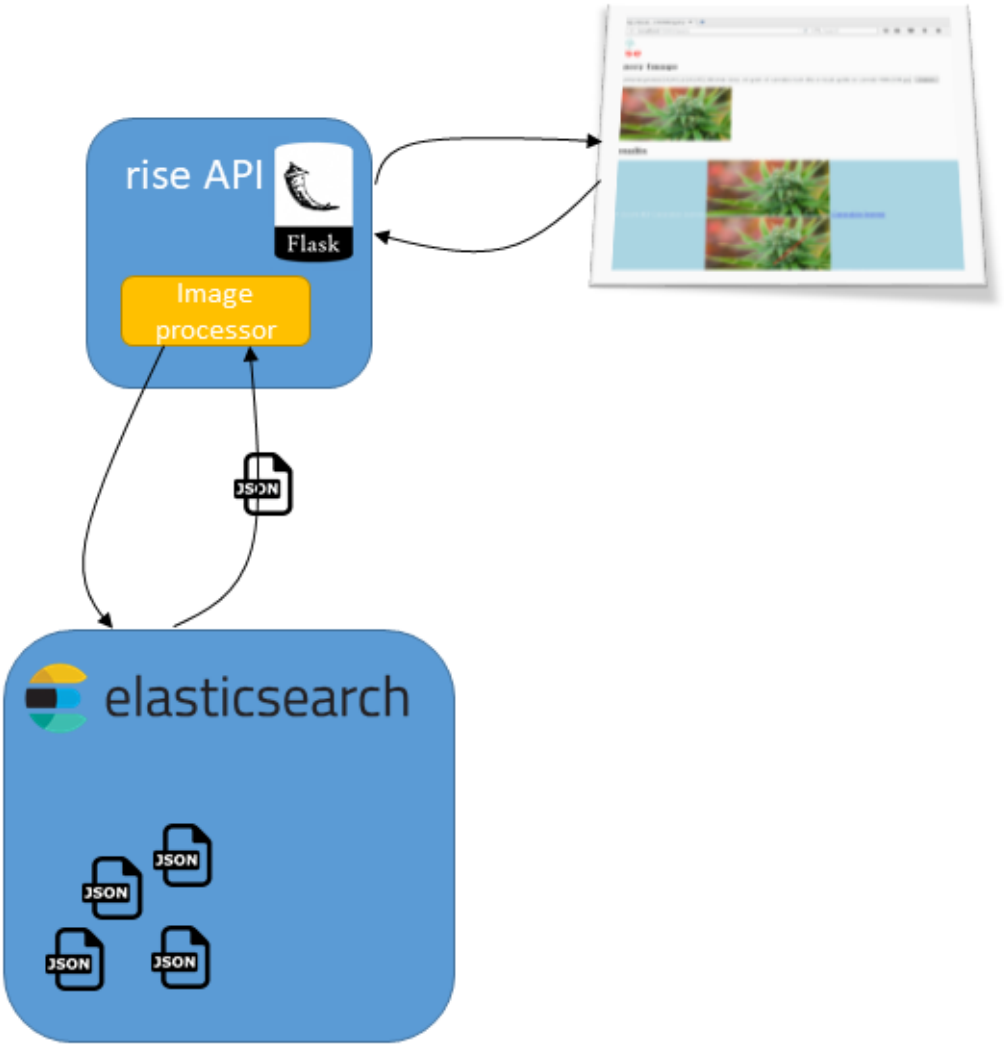
Εικόνα 8 – Αρχιτεκτονική Crawling & Indexing

3.1 Γενική εικόνα

Η μηχανή βασίζεται στο Scrapy Framework για τις ανάγκες της φάσης του crawling. Έχει υλοποιηθεί ένας spider, ο οποίος περιέχει τον επεξεργαστή εικόνων. Για την επικοινωνία με το Tor δίκτυο έχει χρησιμοποιηθεί το Tor service και ένας ενδιάμεσος εξυπηρετητής προώθησης (proxy), ο Polipo.

Ο spider, χρησιμοποιώντας ένα αρχικό σύνολο από .onion τοποθεσίες, ξεκινάει να επισκέπτεται τις κρυμμένες υπηρεσίες. Σε κάθε σελίδα, τοποθετεί στην ουρά τους νέους συνδέσμους που θα βρει, ενώ στέλνει όλες τις εικόνες στον επεξεργαστή εικόνων. Ο επεξεργαστής υπολογίζει την μοναδική υπογραφή της κάθε εικόνας και την προσθέτει στο Elasticsearch, στην φάση της δημιουργίας ευρετηρίων.

Για τις ανάγκες της φάσης της ανάκτησης πληροφορίας, έχει υλοποιηθεί μια διαδικτυακή εφαρμογή και ένα REST API. Ο χρήστης δίνει ως ερώτημα μια εικόνα, η οποία περνάει από τον επεξεργαστή εικόνων και υπολογίζεται η υπογραφή της. Έπειτα, υπολογίζεται η απόσταση της υπογραφής από τις άλλες που βρίσκονται στο Elasticsearch, και έπειτα επιστρέφονται σε ταξινομημένη λίστα οι πιο σχετικές εικόνες που βρέθηκαν.



Εικόνα 9 – Αρχιτεκτονική εφαρμογής

Όλα τα κομμάτια της εφαρμογής, ο spider, το API και ο επεξεργαστής εικόνων, έχουν γραφτεί σε Python 2.7.9, ενώ η διεπαφή του χρήστη είναι γραμμένη σε HTML και Javascript.

3.2 RISE spider

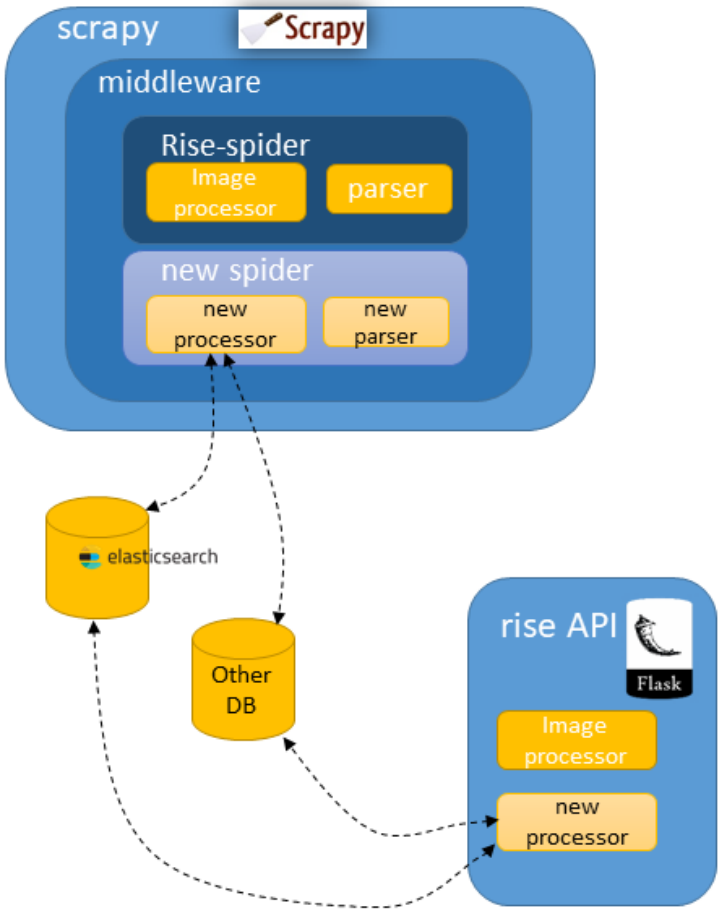
Οι spiders στο Scrapy είναι κλάσεις, που ορίζουν με ποιον τρόπο θα γίνει η ανίχνευση και η απόσπαση δομημένων δεδομένων από τις σελίδες ενός ιστοτόπου.

Ένας τυπικός κύκλος που κάνει ο spider, ακολουθεί τα βήματα:

1. Δημιουργία κλήσεων (requests) για να ανιχνευτούν τα αρχικά URLs (μέθοδος `start_requests()`) και ορισμός μιας μεθόδου (`parse()`) η οποία θα κληθεί όταν έρθει η απάντηση.
2. Στην μέθοδο parse, αναλύεται η απάντηση (η οποία είναι μια σελίδα του ιστότοπου), με χρήση των μηχανισμών που ονομάζονται Selectors και σαν έξοδος επιστρέφονται δομές με τα δεδομένα.
3. Οι δομές επιστρέφονται για να αποθηκευτούν σε κάποια βάση δεδομένων ή σε αρχείο.

Με χρήση του παραπάνω, μπορούν να γραφτούν άλλες μονάδες επεξεργασίας (modules) οι οποίες θα ανιχνεύουν και θα αναλύουν διάφορες μορφές πολυμέσων. Έπειτα, όλες αυτές οι μονάδες επεξεργασίας θα αξιοποιούν τον μηχανισμό CrawlerRunner χρονοπρογραμματισμού εκτέλεσης πολλαπλών spiders του Scrapy Framework.

Παρακάτω, παρουσιάζεται ο τρόπος για την ανάπτυξη και προσθήκη των νέων μονάδων:



Σχήμα 8 – Νέα μονάδα επεξεργασίας

Αρχικά, θα πρέπει να αναπτυχθεί μια βιβλιοθήκη η οποία:

- να υλοποιεί μια μέθοδο **Create_signature (multimedia file)**. Θα πρέπει να παίρνει ως είσοδο ένα πολυμεσικό αρχείο και να υπολογίζει την υπογραφή του μέσω μιας συνάρτησης κατακερματισμού.
- Θα ορίζει μια συνάρτηση ομοιότητας, πχ απόσταση μεταξύ των διανυσμάτων των υπογραφών δύο αρχείων.
- Θα ορίζει ένα κατώφλι ομοιότητας, πέρα από το οποίο δύο εικόνες δεν θεωρούνται όμοιες.

Έπειτα, θα πρέπει να οριστεί ο τρόπος εξαγωγής δεδομένων από την σελίδα (parser), όπως ορίζει το Scrapy Framework. Ο parser θα πρέπει να υλοποιεί τη μέθοδο:

- **Extract_source (text)**: Θα πρέπει να παίρνει ως όρισμα την HTML σελίδα και να ανακτά το URI του πολυμέσου στον εξυπηρετητή που φιλοξενείται. Επιστρέφει το URI.

Τέλος, θα πρέπει να αναπτυχθεί ένα spider που θα αξιοποιεί τις δύο προαναφερθείσες μονάδες και θα συντονίζει τις ενέργειες της ανίχνευσης, σύμφωνα με τον ακόλουθο ψευδοκώδικα:

Πίνακας 3 – Αλγόριθμος spider

```
parse(response):
  title = response.title;
  page = response.url;
  for multimedia_file in response.extract(multimedia_type)
    url = extract_source(multimedia_file);
    metadata = {url, title, page};
    signature = create_signature(multimedia_file);
    add_to_storage(storage,signature,metadata);
```

όπου η μέθοδος add_to_storage ορίζεται ως ακολούθως:

- **Add_to_storage (storage, signature, metadata)**: Θα πρέπει να προσθέτει μια καινούρια εγγραφή στην επιλεγμένη βάση δεδομένων (είτε Elasticsearch, είτε κάποια καινούρια). Παίρνει ως είσοδο την βάση δεδομένων, την υπογραφή του πολυμέσου και τα μεταδεδομένα. Αυτά ορίζονται ως ένα αντικείμενο που περιέχει τον τίτλο της σελίδας, το URL της σελίδας και το URI του πολυμέσου.

Αυτά τα βήματα αρκούν για το στάδιο της ανίχνευσης νέου τύπου πολυμέσων. Έπειτα, για την ανάκτηση πληροφορίας και την αναζήτηση μέσα στην βάση δεδομένων, πρέπει να υλοποιηθεί η μέθοδος search_in_storage, που θα χρησιμοποιηθεί από το RISE API.

- **Search_in_storage (storage, signature)**: θα πρέπει να λαμβάνει ως είσοδο την βάση δεδομένων και μία υπογραφή, και να επιστρέφει μια λίστα με τα πιο σχετικά αποτελέσματα. Η συνάρτηση ομοιότητας καθώς και το κατώφλι ομοιότητας θα έχουν οριστεί μέσα στην βιβλιοθήκη της επεξεργασίας του πολυμέσου.

Συγκεκριμένα, ο αλγόριθμος επεξεργασίας εικόνων που έχει υλοποιηθεί στον RISE spider, αρχικά διαβάζει τα αρχικά URLs από ένα αρχείο κειμένου, και ξεκινάει η ανίχνευση. Η μέθοδος ανάλυσης των σελίδων χρησιμοποιεί τον μηχανισμό των XPath επιλογέων (selectors) για να ανακτήσει όλα τα `img` HTML tags και από αυτά να εξάγει το χαρακτηριστικό (HTML attribute) `src`, το οποίο αποτελεί την «πηγή» της εικόνας. Η πηγή αυτή είναι ένας σύνδεσμος προς την εικόνα, ο οποίος δίδεται ως είσοδος στον επεξεργαστή εικόνας, όπου την κατακερματίζει. Έπειτα, δημιουργεί μια δομή (Item) η οποία περιέχει τον σύνδεσμο προς την εικόνα και προς την σελίδα που την περιέχει, τον τίτλο της σελίδας και την μοναδική υπογραφή της εικόνας. Η δομή αυτή αποθηκεύεται στο Elasticsearch, και ο αλγόριθμος καλεί αναδρομικά τον εαυτό του με τον επόμενο σύνδεσμο.

3.3 Επεξεργαστής εικόνων

Ο επεξεργαστής εικόνων περιέχει όλα τα απαραίτητα εργαλεία για την επεξεργασία των εικόνων σε όλες τις φάσεις του κύκλου ζωής τους:

Υπολογισμός υπογραφής εικόνας: Ο επεξεργαστής υλοποιεί τον αλγόριθμο που παρουσιάστηκε στην «Παράγραφος 3.5 – Υπογραφή Εικόνας».

Υπολογισμός ευρετηρίων: Η ανάγκη για ύπαρξη ευρετηρίων προκύπτει από την σχεδίαση των συστημάτων αποθήκευσης δεδομένων εν γένει, τα οποία είναι πολύ πιο γρήγορα στην ανάκτηση δεδομένων με χρήση ευρετηρίου παρά με την αναζήτηση στο σύνολο των πεδίων. Έτσι, ο επεξεργαστής προχωράει στον υπολογισμό ενός συνόλου χαρακτηριστικών αριθμών για κάθε διάνυσμα. Οι αριθμοί αυτοί αντιστοιχούν σε κάθε ευρετήριο και τα πολλαπλά ευρετήρια χρησιμοποιούνται για να κάνουν την ανάκτηση των δεδομένων πολύ πιο γρήγορη.

Αποθήκευση εικόνων: Ο επεξεργαστής δημιουργεί ένα Item για κάθε εικόνα, το οποίο περιέχει την υπογραφή της και τα αντίστοιχα ευρετήρια, σε μορφή μεταδεδομένων. Έπειτα αποθηκεύονται στο Elasticsearch σε μορφή JSON .

Υπολογισμός ομοιότητας εικόνων: Όπως έχουμε ήδη αναφέρει, η ομοιότητα δύο εικόνων χαρακτηρίζεται από την απόσταση των αντίστοιχων διανυσμάτων τους. Ένα ακόμα κομμάτι της λειτουργίας του επεξεργαστή εικόνων λοιπόν, είναι ο υπολογισμός της απόστασης αυτής.

3.4 Polipo

Το δίκτυο Tor εκ κατασκευής, χρησιμοποιεί το SOCKS για να υλοποιήσει το πρωτόκολλο επικοινωνίας του. Όμως οι spiders διαδικτύου εκτελούν κλήσεις πάνω από το HTTP πρωτόκολλο. Έτσι, χρειάζεται η ύπαρξη ενός ενδιάμεσου που θα «μεταφράζει» τις HTTP κλήσεις σε SOCKS. Για αυτό τον σκοπό έχει χρησιμοποιηθεί ο Polipo [8], ένας χαμηλών απαιτήσεων proxy προώθησης αιτήσεων, με επιπλέον δυνατότητες, όπως είναι η προσωρινή αποθήκευση (caching). Το caching είναι ιδιαίτερα χρήσιμο σε περιπτώσεις «αργών» δικτύων, όπως είναι το Tor, καθώς προσφέρει γρηγορότερη απόκριση.

3.5 Διαδικτυακή εφαρμογή

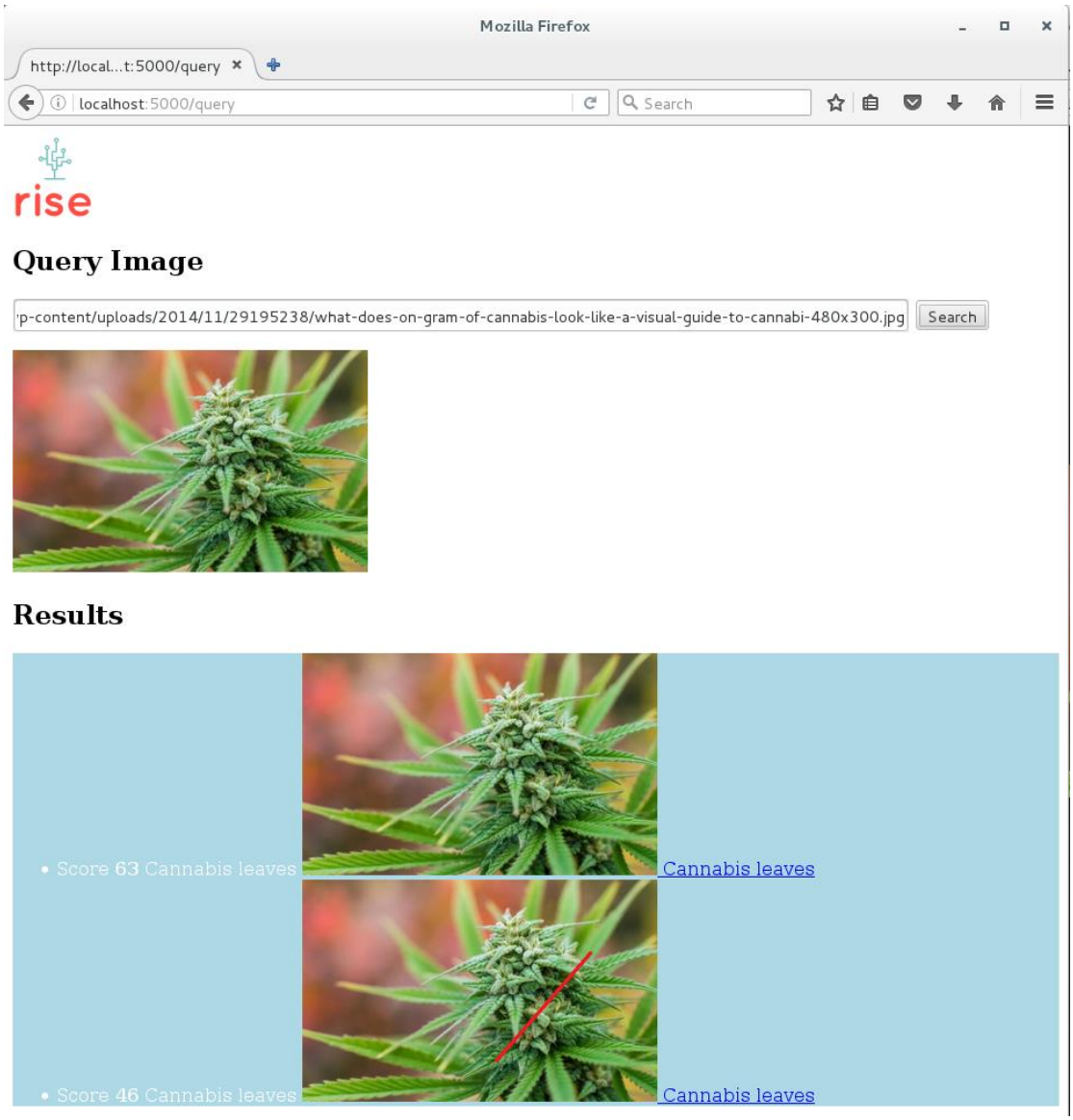
Για την χρήση όλων τον εργαλείων που προσφέρει ο βασικός μηχανισμός της ανίχνευσης, ταξινόμησης σε ευρετήρια και ανάκτησης δεδομένων, έχει υλοποιηθεί μια διαδικτυακή εφαρμογή που δρα και ως διεπαφή του χρήστη.

Η εφαρμογή αυτή αποτελείται από δύο κομμάτια: το REST API και την κονσόλα του χρήστη.

3.5.1 Διεπαφή χρήστη

Η διεπαφή του χρήστη αποτελείται από ένα SPA (single page application), μια διαδικτυακή σελίδα δηλαδή, μέσω της οποίας μπορεί να πραγματοποιεί ερωτήματα προς την μηχανή αναζήτησης και να λαμβάνει τα ταξινομημένα αποτελέσματα. Είναι γραμμένη σε JavaScript και με την χρήση της βιβλιοθήκης jQuery πραγματοποιούνται κλήσεις προς το API ώστε να καταχωρηθεί ένα ερώτημα, της μορφής URI, το οποίο θα

πρέπει να καταδεικνύει ένα αρχείο εικόνας. Για την καταχώρηση (POST HTTP request) του ερωτήματος, πρέπει να πατήσει το κουμπί “Search”.



Εικόνα 10 – Διεπαφή χρήστη

3.5.2 REST API

Το REST API είναι μια εφαρμογή από την πλευρά του εξυπηρετητή (server-side application) που δημοσιοποιεί το διαθέσιμο προς κλήση σημείο (endpoint).Το API είναι γραμμένο σε Python, με χρήση της βιβλιοθήκης Flask. Μόλις καταχωρηθεί ένα POST αίτημα από την πλευρά της σελίδας, το API χρησιμοποιεί τον επεξεργαστή εικόνων ώστε να υπολογίσει την υπογραφή της εικόνας-ερώτημα. Έπειτα, κατασκευάζονται ερωτήματα προς το Elasticsearch, το οποίο με αξιοποίηση των ευρετηρίων επιστρέφει τα αποτελέσματα.

Τα αποτελέσματα αυτά μεταποιούνται στην κατάλληλη μορφή, και επιστρέφονται προς την σελίδα, ως HTTP response.

4. ΑΞΙΟΛΟΓΗΣΗ

4.1 Λειτουργικές απαιτήσεις

Οι λειτουργικές απαιτήσεις, όπως αυτές ορίστηκαν κατά την φάση του σχεδιασμού του συστήματος, αποτελούνται από τρία βασικά χαρακτηριστικά.

Επεκτασιμότητα: για την αξιολόγηση της επεκτασιμότητας, δημιουργήθηκε μια επιπλέον μονάδα επεξεργασίας. Για λόγους απλότητας, η μονάδα αυτή αφορούσε κείμενα και όχι κάποιο άλλο τύπο πολυμέσων, αλλά αρκεί για την επιβεβαίωση της επεκτασιμότητας του συστήματος.

Ελαχιστοποίηση απαιτήσεων χώρου: Μετά από χρήση της μηχανής δίκτυο Tor, παρατηρούμε ότι για 1517 αρχεία που έχουν αποθηκευτεί, χρησιμοποιούνται 7.67MB πληροφορίας. Επομένως, χρησιμοποιούμε μόλις 5 KB πληροφορίας για κάθε εικόνα, ανεξάρτητα από το αρχικό της μέγεθος.

Κλιμάκωση: καθώς αυξάνεται το πεδίο αναζήτησης, αυξάνεται και ο όγκος της αποθηκευμένης πληροφορίας. Δημιουργώντας νέους Elasticsearch κόμβους, κρατήσαμε σταθερό τον όγκο πληροφορίας ανά κόμβο, επομένως η εφαρμογή είναι κλιμακώσιμη.

4.2 Μη λειτουργικές απαιτήσεις

Η μηχανή RISE κατασκευάστηκε με σκοπό να συνεισφέρει στην αναζήτηση ενός ερωτήματος, ανακαλύπτοντας σελίδες στις οποίες οι μηχανές αναζήτησης κειμένου δεν μπορούν να κατηγοριοποιήσουν ως σχετικές. Για την πειραματική αξιολόγηση των μη λειτουργικών απαιτήσεων της μηχανής σε συνδυαστική αναζήτηση, επιλέχθηκαν δύο χαρακτηριστικά, η πληρότητα και η αποτελεσματικότητα. Η πληρότητα ορίζεται ως το ποσοστό του συνόλου των σχετικών σελίδων που ανακτήθηκαν με βάση την εικόνα. Το ποσοστό αυτό θα πρέπει να είναι θετικό για να δηλώσει επιτυχή συνεισφορά της μηχανής που κατασκευάσαμε. Η αποτελεσματικότητα μελετάει την απόδοση του αλγορίθμου ομοιότητας μεταξύ εικόνων. Ο αλγόριθμος αυτός εξετάστηκε σε 8 παραλλαγές μιας εικόνας και την κατάταξη της ομοιότητάς τους σε σχέση με την αρχική.

4.2.1 Πληρότητα

Για τον υπολογισμό του δείκτη πληρότητας ακολουθήθηκε μια παραλλαγή της μεθοδολογίας που παρουσιάζεται στην δουλειά των Chowdhury και Soboroff, “Automatic Evaluation of World Wide Web Search Services” [9], που αφορά την σύγκριση μηχανών αναζήτησης.

Η μεθοδολογία που ακολουθήθηκε ορίζει την συγκέντρωση ενός αρχικού συνόλου από σελίδες και έπειτα την πειραματική εξέταση των δύο διαφορετικών τρόπων ανάκτησης πάνω σε αυτό: με βάση το κείμενο και με βάση την εικόνα. Για την συγκέντρωση του αρχικού συνόλου, χρησιμοποιήθηκε η μηχανή αναζήτησης TinEye, η οποία απέφερε 1519 αρχεία.

Στην φάση της ανάκτησης, για την αναζήτηση με βάση το κείμενο αξιοποιήθηκε η μηχανή Lucene, πάνω στην οποία είναι υλοποιημένο το Elasticsearch, και χρήση της λέξης-κλειδί “Cannabis” ως ερώτημα. Η ανάκτηση με βάση την εικόνα έγινε με χρήση της δικής μας μηχανής, RISE, και χρήση της Εικόνας 11 ως ερώτημα.

Εικόνα 11 – Εικόνα που χρησιμοποιήθηκε ως ερώτημα

Τα αποτελέσματα της αναζήτησης με κείμενο, έδειξαν ότι ανακτήθηκαν 108 αποτελέσματα, εκ των οποίων τα 95 αποτελούσαν μοναδικά domains. Η αναζήτηση με βάση την εικόνα επέστρεψε 19 αποτελέσματα, εκ των οποίων τα 6 ήταν νέα domains, δηλαδή domains που δεν είχαν ανακτηθεί με την αναζήτηση κειμένου. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.

Πίνακας 4 – Αποτελέσματα αναζητήσεων

Ερώτημα	# Αποτελέσματα	# Νέα Domains
“Cannabis”	108	**95**
Εικόνα 11	19	**6**
“Marijuana”	96	0
“Weed”	6	0

Παρατηρούμε πως η αναζήτηση με βάση την εικόνα προσέφερε στην συνολική αναζήτηση επιπλέον 6 αποτελέσματα, τα οποία δεν θα μπορούσαν να είχαν εντοπιστεί ακόμη και αν χρησιμοποιούσαμε συνώνυμα της λέξης-κλειδί, όπως για παράδειγμα είναι οι λέξεις “marijuana” και “weed”.

Η μηχανή αναζήτησης εικόνων κατάφερε να ανακτήσει 14% περισσότερα αποτελέσματα σε σχέση με την παραδοσιακή αναζήτηση κειμένου, όπως φαίνεται στο Σχήμα 9.

Σχήμα 9 – Κατανομή ανακτηθέντων αποτελεσμάτων

Όπως προκύπτει, η μηχανή RISE μπορεί να συνεισφέρει αρκετά (14%) στην ανάκτηση αποτελεσμάτων που δεν είχαν ανιχνευθεί από τις μηχανές αναζήτησης κειμένου. Αυτή η συμπεριφορά είναι αναμενόμενη, καθώς τα επιπλέον αποτελέσματα συμπεριλάμβαναν σελίδες που ήταν γραμμένες σε διαφορετική γλώσσα, χρησιμοποιούσαν διαφορετική «διάλεκτο», είτε απουσίαζε κείμενο περιγραφής.

Με τη χρήση συνδυαστικής αναζήτησης, δηλαδή αναζήτηση πολυμέσων και κειμένου, αναμένεται να υπάρχει μεγάλη αύξηση αποτελεσμάτων κακόβουλων ή/και παράνομων κρυμμένων υπηρεσιών, βοηθώντας περαιτέρω στον εντοπισμό, την κατηγοριοποίηση και την εκκίνηση ενεργειών για την εξάλειψη τους.

4.2.2 Αποτελεσματικότητα

Ως αποτελεσματικότητα της μηχανής αναζήτησης εικόνων ορίζουμε την δυνατότητά της να ανακτήσει σχετικά, στο ερώτημα, αποτελέσματα. Για το πείραμα αυτό, τροφοδοτήσαμε τον μηχανισμό ανίχνευσης της μηχανής με μία εικόνα και οκτώ παραλλαγές της. Όπως έχουμε αναφέρει στο αντίστοιχο κεφάλαιο, ο αλγόριθμος επεξεργασίας υπολογίζει την ομοιότητα δύο εικόνων ως την ευκλείδεια απόσταση μεταξύ των αντίστοιχων διανυσμάτων τους. Ταυτόχρονα, το κατώφλι ομοιότητας εικόνων έχει οριστεί πειραματικά στην δουλειά των Wong, Bern και Goldberg. Έτσι, για μεγάλες αποστάσεις μεταξύ των διανυσμάτων, ο βαθμός ομοιότητας είναι μηδέν.

Το είδος της παραλλαγής, καθώς και ο βαθμός της ομοιότητας που δόθηκε από τον αλγόριθμο, παρουσιάζονται στον Πίνακα 5.

Πίνακας 5 – Πλήθος αποτελεσμάτων ανά παραλλαγή

Είδος παραλλαγής	Βαθμός ομοιότητας
Αλλαγή απόχρωσης	100%
Απομάκρυνση χρωμάτων (ασπρόμαυρη εικόνα)	100%
Αλλαγή μεγέθους	95,23%
Προσθήκες μικρών αντικειμένων	73,01%
Αλλαγή λόγου (ratio)	60,31%
Μικρή περιστροφή εικόνας (< 90^o)	0%
Μεγάλη περιστροφή εικόνας (>180^o)	0%
Προσθήκες μεγάλων αντικειμένων	0%

Παρατηρούμε ότι ο αλγόριθμος υπολογισμού ομοιότητας που χρησιμοποιήθηκε έχει μεγάλη ανοχή σε παραλλαγές που αφορούν το χρώμα και το μέγεθος. Αυτό είναι αναμενόμενο, καθώς ο αλγόριθμος στα πρώτα βήματά του αφαιρεί το χρώμα από τις εικόνες και τις μετατρέπει στο ασπρόμαυρο ισοδύναμό τους. Επίσης η μέθοδος δειγματοληψίας της εικόνας πριν τον κατακερματισμό της, επιτρέπει στον αλγόριθμο να μην επηρεάζεται από αλλαγές μεγέθους και προσθήκες μικρών αντικειμένων, όπως για παράδειγμα μια λέξη.

Αντίθετα, η περιστροφή της εικόνας ή η προσθήκη ενός μεγάλου αντικειμένου (περισσότερο από το 50% της επιφάνειάς της) την καθιστά μη αναγνωρίσιμη. Τέτοιου είδους παραλλαγές δημιουργούν μεγάλες αλλαγές στην έξοδο του αλγορίθμου κατακερματισμού, με αποτέλεσμα η απόσταση της νέας υπογραφής να είναι πολύ μεγάλη από την αρχική εικόνα. Όταν η απόσταση είναι πολύ μεγάλη, οι εικόνες θεωρούνται μη-όμοιες, και η μηχανή δεν επιστρέφει κανένα αποτέλεσμα.

Σχετικά με την αποτελεσματικότητα, βλέπουμε ότι η μηχανή RISE διαπρέπει σε εφαρμογές που ο στόχος είναι να εντοπιστούν πανομοιότυπες εικόνες, όπως για παράδειγμα εφαρμογές ψηφιακής εγκληματολογίας, πνευματικών δικαιωμάτων κ.ο.κ. Για την κάλυψη μεγαλύτερου εύρους εφαρμογών και χρήσεων της μηχανής, η επέκτασή της είναι δυνατή, όπως περιγράφεται στο «Κεφάλαιο 6 – Μελλοντικές Προεκτάσεις».

4.2.3 Σύγκριση με Google Image Search

Σε αυτή τη δοκιμή έγινε μια σύγκριση της μηχανής RISE με την μηχανή αντίστροφης αναζήτησης εικόνων της Google. Ο σκοπός της σύγκρισης είναι να δείξουμε πώς η μηχανή RISE μπορεί να συνεισφέρει όταν χρησιμοποιείται σε συνδυασμό με άλλα εργαλεία και μηχανές αναζήτησης.

Στην πρώτη φάση της δοκιμής, δόθηκε ως ερώτημα η Εικόνα 11 στην μηχανή αναζήτησης εικόνων Google Image Search (GIS). Η μηχανή ανέκτησε περίπου 670 μοναδικά αποτελέσματα. Στον δεύτερο κύκλο αναζήτησης, δόθηκε ως ερώτημα μια παραλλαγή της Εικόνας 11, η οποία είχε υποστεί περιστροφή κατά 10 μοίρες. Η μηχανή ανέκτησε 460 αποτελέσματα. Η ανάλυση των αποτελεσμάτων σε σχέση με την μηχανή RISE παρουσιάζεται στον ακόλουθο πίνακα:

Πίνακας 6 - Ανάλυση αποτελεσμάτων

Αποτελέσματα	Εικόνα 11	Περ/μένη Εικόνα 11	Νέα domains
GIS Clearnet	670	460	673
RISE Clearnet	19	0	0
GIS Tor	0	0	0
RISE Tor	5	0	5

Παρατηρούμε ότι η μηχανή αναζήτησης της Google έχει σαφώς υπεροχή στο πλήθος των ανακτηθέντων αποτελεσμάτων σε σχέση με την RISE. Το πολύ μικρό πλήθος των αποτελεσμάτων της RISE οφείλεται παράλληλα και στο γεγονός της μικρής διάρκειας της διαδικασίας indexing στην οποία έχει εκτεθεί, στα πλαίσια αυτής της εργασίας. Παράλληλα, είναι αξιοσημείωτο ότι η GIS επέστρεψε αποτελέσματα σε περίπου 10 γλώσσες, αναιρώντας εν μέρει την αρχική μας υπόθεση ότι η αναζήτηση με βάση το κείμενο βρίσκει εμπόδιο στην ποικιλία γλωσσών. Παραμένει όμως ανοιχτό το πρόβλημα της χρήσης «ειδικών» διαλέκτων (αργκό), στο οποίο η μηχανή RISE προσφέρει λύση.

Παρ’ όλο που η RISE επέστρεψε σημαντικά λιγότερα αποτελέσματα, κατάφερε να συνεισφέρει στο σύνολο σε ποσοστό 0.73%, προσφέροντας 5 καινούρια domains. Τα domains αυτά ανήκουν στο Darknet, στο οποίο η μηχανή GIS δεν δραστηριοποιείται.

5. ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Στην εργασία αυτή περιγράφηκε μια μηχανή αναζήτησης πολυμέσων ως εργαλείο για τον εντοπισμό κρυμμένων υπηρεσιών στο Tor. Ως απόδειξη της λειτουργίας της ιδέας, υλοποιήθηκε μια μηχανή που ανιχνεύει πολυμέσα που έχουν την μορφή εικόνας.

Η μηχανή μπορεί να βρει πολλές εφαρμογές, ως πλατφόρμα έρευνας και ανάπτυξης νέων αλγορίθμων στην αναζήτηση στο Διαδίκτυο, καθώς η εκτέλεση και αξιολόγησή τους μπορεί να γίνει γρήγορα και εύκολα, μέσω της modular αρχιτεκτονικής. Η κυρίως μονάδα επεξεργασίας σελίδων μπορεί να αναπτυχτεί σε Python και να αντικαταστήσει την υπάρχουσα μονάδα.

Μια ακόμα προέκταση θα μπορούσε να είναι ο εμπλουτισμός της μηχανής αναζήτησης με επεξεργαστές πολυμέσων άλλων τύπων, όπως είναι το βίντεο, ήχος κλπ.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Bullying	Εκφοβισμός
Cache	Προσωρινή Αποθήκευση
Exit Node	Κόμβος Εξόδου
Framework	Πλαίσια
Hidden Service	Κρυμμένη Υπηρεσία
Hidden Service Descriptor	Περιγραφέας Κρυμμένης Υπηρεσίας
Hosting Provider	Πάροχος Ιστοτόπου
Introduction Points	Σημεία Εισαγωγής
IP address	Διεύθυνση του Πρωτοκόλλου Διαδικτύου
Keyword	Λέξη-κλειδί
Link	Σύνδεσμος
Log	Ημερολόγιο συμβάντων
Malware	Κακόβουλο Λογισμικό
Metadata	Μεταδεδομένα
Modular	Δομοστοιχειωτός
Modular	Επεκτάσιμο
Perceptual Hashing	Αντιληπτικός Κατακερματισμός
Perfect Forward Secrecy	Τέλεια Μυστικότητα προς τα Εμπρός
Predictive Analytics	Προγνωστική Ανάλυση
Privacy	Ιδιωτικότητα
Proxy	Εξυπηρετητής Προώθησης
Query	Ερώτημα
Relay Server	Εξυπηρετητής Αναμετάδοσης
Replication	Αναπαραγωγή
Request	Κλήση
Scalable	Κλιμακώσιμο
Server	Εξυπηρετητής
Thread	Νήμα
Website	Ιστοσελίδα
Bullying	Εκφοβισμός
Cache	Προσωρινή Αποθήκευση

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

ISP	Internet Service Provider
HTML	HyperText Markup Language
SOCKS	Socket Secure
TLD	Top Level Domain
WWW	World Wide Web
URL	Uniform Resource Locator
URI	Uniform Resource Identifier
FTP	File Transfer Protocol
HTTP	HyperText Transfer Protocol
REST	Representational State Transfer
API	Application Programming Interface
SPA	Single Page Application
JSON	JavaScript Object Notation

ΠΑΡΑΡΤΗΜΑ

Ρύθμιση rise-spider ώστε να χρησιμοποιεί τον proxy Polipo

Middleware.py

```
from scrapy.conf import settings


class ProxyMiddleware(object):
    def process_request(self, request, spider):
        request.meta['proxy'] = settings.get('HTTP_PROXY')
```

Settings.py

```
BOT_NAME = 'rise'

SPIDER_MODULES = ['rise.spiders']

NEWSPIDER_MODULE = 'rise.spiders'

HTTP_PROXY = 'http://127.0.0.1:8123'

ROBOTSTXT_OBEY = True

CONCURRENT_REQUESTS = 100

COOKIES_ENABLED = False

DOWNLOADER_MIDDLEWARES = {
     'rise.middlewares.ProxyMiddleware': 400,

'scrapy.contrib.downloadermiddleware.useragent.UserAgentMiddlewa
re': None
}

REACTOR_THREADPOOL_MAXSIZE = 20

LOG_LEVEL='INFO'

RETRY_ENABLED=False

REDIRECT_ENABLED=False
```

ΑΝΑΦΟΡΕΣ

[1] "About Tor," [Online]. Available: https://www.torproject.org/about/torusers.html.en.

[2] S. Pederson, "Understanding the Deep Web in 10 minutes," Brightplanet white paper, Sioux Falls, 2013.

[3] R. Dingledine, N. Mathewson and P. Syverson, "Tor: The Second-Generation Onion Router," Usenix Security, 2004.

[4] Houses of Parliament, “The Darknet and Online Anonymity”, PostNote, March 2015.

[5] E. Klinger and D. Starkweather, "pHash," 2014. [Online]. Available: http://www.phash.org/.

[6] H. Chi Wong, M. Bern and D. Goldberg, "An image signature for any kind of image," in Image Processing, 2002.

[7] S. Berchtold, D. A. Keim, H. Kriegel, “The X-tree: An Index Structure for High-Dimensional Data”, Proceedings of the 22nd VLDB Conference, 1996.

[8] J. Chroboczek, "Polipo - a caching web proxy," 2003. [Online]. Available: https://www.irif.fr/~jch/software/polipo/.

[9] A. Chowdhury and I. Soboroff, "Automatic Evaluation of World Wide Web Search Services," SIGIR, pp. 421 - 422, 2002.

[10] F. Breitinger, H. Liu, C. Winter, H. Baier, A. Rybalchenko and M. Steinebach, "Towards a Process Model for Hashing Functions in Digital Forensics," Digital Forensics and Cyber Crime, pp. 170-186, 2013.

[11] "Reporters Sans Frontieres," [Online]. Available: https://rsf.org/.

[12] D. Neumann, K.R. Gegenfurtner, “Image Retrieval and Perceptual Similarity”, ACM Transactions on Applied Perception, Vol. 3, No. 1, January 2006, pp 31-47.

[13] A. Paro, “Elasticsearch Cookbook”, Second Edition, PACKT publishing, 2015.

[14] O. Kononenko, O. Baysal, R. Holmes, M. W. Godfrey, “Mining Modern Repositories with Elasticsearch”, MSR 2014 Proceedings of the 11th Working Conference on Mining Software Repositories, pp 328-331.

[15] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, J. A. Halderman, “A Search Engine Backed by Internet-Wide Scanning”, CCS ’15 Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp 542 – 553.

[16] “National Strategy for Child Exploitation Prevention and Interdiction”, [Online]. Available: https://www.justice.gov/psc/national-strategy-child-exploitation-prevention-and-interdiction