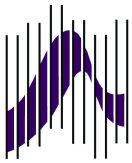
Hochschule
Ravensburg-Weingarten



Hochschule
Ravensburg-Weingarten
Technik | Wirtschaft | Sozialwesen

Erstellung eines Programms zur automatisierten Informationsbeschaffung von personenbezogenen Daten in Verbindung mit einem automatisierten Phishing-Mailgenerators

Bachelorarbeit

Social Engineering

im Studiengang Angewandte Informatik

an der Hochschule Ravensburg - Weingarten

von

Marco Lang Matr.-Nr.: 27416

Abgabedatum : 29. Januar 2019

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit mit dem Titel

Generierung eines personalisierten Mail-Generators

selbststandig angefertigt, nicht anderweitig zu Pruefungszwecken vorgelegt, keine anderen als die angegebenen Hilfsmittel benutzt und woertliche sowie sinngemaesse Zitate als solche gekennzeichnet habe.

Weingarten, 29. Januar 2019

Autor Name

_____

Inhaltsverzeichnis

Kurzfassung

Abstract

Im Rahmen dieser Abschlussarbeit wird gezeigt, wie eine automatisierte Suche nach personenbezogenen Daten im Internet aussehen kann und wie diese Daten für einen Phishing-Mail-Angriff verwendet werden können.

Danksagung

Vorwort

1 Einleitung

1.1 Motivation

Laut dem Bundeskriminalamt hat sich die Zahl der Cyberkriminalität mit einem klaren Trend nach oben entwickelt. [Bun18] Aus diesem Grund werden System immer sicherer und Firewalls immer noch besser. Das hat zu Folge, dass Angreifer oft auf Methoden ausweichen, bei denen der Mensch als Schwachstelle des Systems ausgenutzt wird. Daher ist eine häufig verwendetet Technik von Cyberkriminalität das E-Mail-Phishing.
In den neusten Fällen von Phishing-Mail-Attacken zeigt die Verbraucherzentrale Nordrhein-Westfalen, dass diese meist direkt an eine Person adressiert sind. Das heißt, in dieser Art von E-Mail, werden personenbezogene Daten verwendet. Ein Beispiel dafür, sind die gefälschten DSGVO-E-Mails. Hier wird die Zielperson im Namen der Sparkasse, persönlich mit Namen angesprochen. [NW18]
Solch ein Angriff benötigt im Voraus eine ausführliche Recherche über das Opfer. Als Informationsquelle für die Recherche können beliebig viele Quellen verwendet werden. Jedoch ist in der heutigen Zeit das Internet eine der meistgenutzten Informationsquellen. [All18]

1.2 Zielsetzung

Ziel dieser Arbeit ist es ein Programm zu entwickeln, welches automatisiert nach personenbezogenen Daten im Internet sucht und daraus eine Phishing-Mail generiert. Dabei soll der Fokus auf der automatisierten Informationsbeschaffung liegen.
Es sollen grundsätzlich zwei verschiedene Suchfunktionen mit diesem Programm möglich sein.

Ziel 1 *Informationen zu einer bestimmten Person im Internet suchen.*

Die erste Suchfunktion beinhaltet die Suche nach Informationen einer bestimmten Person. Dadurch können bereits bekannte Daten über die Person angegeben und somit die Suche verfeinert beziehungsweise verbessert werden. Hierbei ist es wichtig zu erkenne wann es sich um eine Information der gesuchten Person handelt.

Ziel 2 *Webseiten, die eine große Menge von personenbezogener Daten enthalten, auslesen und analysieren.*

Durch die zweite Suchfunktion soll eine große Menge an Daten gewonnen werden und dadurch ein weitläufiger Angriff zu simulieren.
Bei der zweiten Suchfunktion sollen nur bestimmte Webseiten vorgegeben werden, welche ausgelesen und analysiert werden sollen. Durch diese Funktion ist es möglich einen weitläufigen Phishing-Mail-Angriff zu simulieren.

Ziel 3 *E-Mail-Adressen aus den gewonnen Daten generieren.*

Durch die Zusammensetzung von Vorname, Name und Geburtsjahr und/oder Firma werden die E-Mail-Adressen generiert.

Ziel 4 *Phishing-Mail-Muster erstellt*

Abhängig von den nach gefundenen Informationen, soll mit Hilfe der Muster eine glaubhafte und sinnvolle Mail erstellt werden.

Ziel 5 *Phishing-Mail erzeugen.*

Mit der vorhandenen Information, der E-Mail-Adresse und einem passende Muster, soll eine Phishing-Mail erzeugt und versendet werden können.

1.3 Eigene Leistung

In dieser Arbeit wird ein Programm erstellt, welches personenbezogene Daten automatisiert aus dem Internet heraussucht und diese in potentielle Opferprofile ablegt. Die gewonnenen Informationen werden automatisiert in eine personalisierte Phishing-E-Mail eingebaut. Für einen höheren Erfolg werden E-Mail-Muster erstellt.

Damit ein kompletter Ablauf eines Phishing-Mail-Angriffs simuliert werden kann, wird ein Algorithmus entwickelt, der aus den gewonnen Informationen eine E-Mail-Adresse generiert.

1.4 Aufbau der Arbeit

Die Arbeit gliedert sich in einem theoretischen und praktischen Teil auf. Der Theorie-Teil beginnt im zweiten Kapitel und beschreibt die Grundbegriffe im Bereich Social Engineering, Webtools, E-Mails und Programmiersprachen. Im nächsten Kapitel befindet sich die Anforderungsanalyse. Hier werden die Anforderungen an die Arbeit festgelegt. Darauf folgen die Lösungsvorschläge im Kapitel vier und die ausgewählte Lösung anhand den Anforderungen im Kapitel 5. Anschließend wird bei der Umsetzung auf den Praktischen Teil eingegangen.Am Ende befindet sich das Fazit, der Ausblick und der Anhang.

2 Grundlagen

2.1 Social Engineering

2.1.1 Definition

Die Definition von Social Engineering (SE) ist nicht eindeutig. Es gibt sehr verschiedene Ansichten von der Definition. Die Idee von Social Engineering ist, eine Ziel so zu manipulieren, damit das Ziel eine für den Angreifer bessere Entscheidung trifft. In dem Buch Social Engineering - The Art of Human Hacking, von Christopher Hadnagy, ist Social Engineering definiert als “social engineering is the act of manipulating a person to take an action that may or may not be in the “target’s“ best interest“ [Had11]. Die Definition in dem Buch von Kevin D. Mitnick lautet:“Social Engineering uses influence and persuasion to deceive people by convincing them that the social engineer is someone he is not, or by manipulation. As a result, the social engineer is able to take advantage of people to obtain information with or without the use of technology“ [Mit01].

2.1.2 SE im Alltag

SE wird Menschen von Geburt an beigebracht und begegnet einem beinahe jeden Tag. Schon ein Baby muss wissen wie es die Eltern manipulieren kann damit man Dinge wie Essen, Zuneigung, o.ä. bekommt. Darüber hinaus ist SE in vielen Berufen ein täglicher Bestandteil. Beispielsweise manipulieren Ärzte viele Patienten mit einer Placebo-Behandlung. Bei dieser Behandlung wird dem Patient ein wirkstoff-freies Medikament verschrieben. Nur durch die Manipulation des Patienten und den sogenannten Palzebo-Effekt können Erfolge erzielt werden.

2.1.3 SE in der Informationssicherheit

Im Bereich der Informationssicherheit spricht man von Social Engineering wenn man durch Manipulierung bzw. das Hacken von Menschen Passwörter, Zugänge zu Systemen oder vertrauliche Information bekommt. Die bekanntesten Angriffsmethoden sind Phishing, Pretexting, Baiting und Quad Pro Quo. Bei dieser Arbeit wird aber hauptsächlich auf das Thema Phishing eingegangen.

2.1.4 SE Angriffe

Aufbau eine SE-Angriffzykluses

Der Aufbau eines SE-Angriffes ist definiert in mehrere Phasen. Das wohl bekannteste Modell für einen Social Engineering-Angriffszyklus ist in dem Buch von Kevin D. Mitnicks - The art of deception: controlling the human element of security [Mit01] definiert. Dieser Zyklus besteht aus den 4 Phasen Research, Developing rapport and trust, Exploiting trust und Utilize information. In der Research-Phase geht es um die Informationsbeschaffung, bei der der Angreifer möglichst viel Informationen über das Ziel herausfindet. Die Developing rapport and trust Phase beschreibt den Aufbau für einen guten Kontakt, da der Angreifer ein leichteres Spiel hat wenn das Ziel dem Angreifer vertraut. Das nun erzeugte Vertrauen wird in der Exploitung trust Phase ausgenutzt. Hier will der Angreifer die eigentlich Information vom Opfer herausfinden. Dies geschieht einerseits durch bestimmtes nachfragen oder Manipulation. Utilize information ist die letzte Phase. Dort wird die gewonnene Information genutzt um das eigentliche Ziel des Angreifers zu erreichen.
Grundsätzlich werden bei einem Social Engineering Angriff menschliche Wünsche, Ängste und verbreitete Verhaltensmuster verwendet um ein Opfer zu manipulieren. [uDsiNe15]

Phishing

Das Wort Phishing wird von dem Wort “fishing“ abgeleitet, da die Angreifer nach Informationen fischen. Das “Ph“ kommt von “sophisticated“ und meint damit, dass die Angreifer ausgeklügelte Techniken verwenden um an Informationen heranzukommen. [Jam05]
Phishing ist ein Angriffsmethode, bei dem ein Angreifer glaubwürdige E-Mails versendet, um von einem Opfer Informationen zu erhalten. Die sogenannten E-Mails enthalten meist eine Aufforderung einen Link zu öffnen und sehen täuschend echt aus. Zum Beispiel könnten

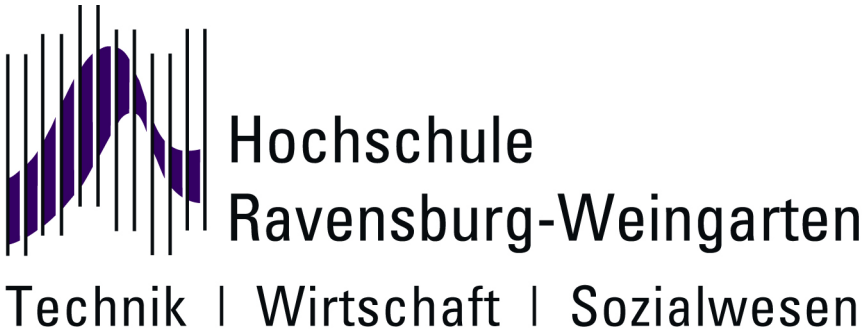


Bild 2.1: Logo der HS – oder nicht?

der Angreifer ein Layout von Amazon verwenden und Sie auffordern, den Link zu öffnen, wegen einem Authentifizierungsproblem. Nachdem Sie auf den Link geklickt haben müssen Sie sich anmelden. Hier könnten die Angreifer Ihre Anmeldedaten abgreifen, nachdem Sie sie eingeben haben. Sobald Sie die Anmeldedaten haben könnten Sie mit der Meldung :“Hoppla, ein Fehler ist aufgetreten, melden Sie sich bitte neu an!“ auf die originale Seite weitergeleitet werden. Durch diesen Vorgang hätten die Angreifer ihre Anmeldedaten bekommen.
Für diese Methode benötigt der Angreifer nicht nur Social Engineering Fähigkeiten sonder auch technische. [CH15]

Spear-Phishing

Spear-Phishing ist eine Unterkategorie des normalen E-Mail-Phishings. Der Unterschied besteht darin, dass anstatt einer anonymen E-Mail, eine Mail an ein ausgewähltes Opfer gesendet wird. In einer Spear-Phishing-E-Mail wird ein Opfer beispielsweise mit einem Namen angesprochen oder es sind E-Mails mit Inhalten die das Opfer interessieren könnten. Aus diesem Grund benötigt man hier Zeit für die Informationsbeschaffung. Dennoch ist der Erfolg hier vielversprechender als beim normalen E-Mail-Phishing. 91% der Advanced Persistent Threat (APT) Angriffe auf Firmen beginnen mit einer Spear-Phishing-E-Mail. Die Schadsoftware wir meisten als Remote Access Trojans (RATs) in einem Zip-Datei überliefert. [Cal13]

2.2 Webtools

2.2.1 Web Scraping

Definition

In der Theorie bedeutet web scraping die Informationsbeschaffung im Internet mit unterschiedlichsten Mitteln. [Mit15]

Funktionsweise

Meist wird dies mit einem automatisierten Programm realisiert, das Daten von einem Webserver anfragt, bekommt, analysiert und auswertet. In der Praxis gibt es ein großes Feld von Programmiertechniken und Einsatzmöglichkeiten. Mit Hilfe von web scraping ist es möglich große Datenmengen zu erfassen und zu verarbeiten. [Mit15]

2.2.2 Web Crawling

Definition

Beim Web Crawling werden Webinhalte geladen und nach Hyperlinks durchsucht. Diesen wird wieder gefolgt und der Prozess beginnt von vorne. Das ist die Grundfunktion einer Suchmaschine. [Mit15] Web Crawlers are computer programs which traverse the hypertext structure in the Web. There are two categories of Web Crawler such as: Internal and External Web Crawler. Internal Crawler crawls through internal pages of the Website which are returned by external crawler. External Crawler crawls through unknown Website. http://ijarcet.org/wp-content/uploads/IJARCET-VOL-1-ISSUE-8-287-293.pdf

https://patentimages.storage.googleapis.com/9c/d7/74/6c0126bb79bb3b/US7065483.pdf

Funktionsweise

Die Funktionsweise besteht darin, dass in den meisten Fällen ein automatisiertes Programm (Web Crawler) erstellt wird. Der Web Crawler lädt Webinhalte herunter und durchsucht diese nach Hyperlinks bzw. URLs. Den gefundenen Hyperlinks werden wieder gefolgt, um neue Webseiten mit weiteren URLs zu laden. So hangelt sich ein Web Crawler von Link zu Link durch das Internet. [Mit15]

2.3 Personenbezogene Daten

2.3.1 Definition

Laut DSGVO sind personenbezogene Daten "alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person") beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind;" [DSG]

2.4 Textanalyse

2.4.1 Stoppwörter

Als Stoppwörter werden Wörter bezeichnet, die sehr oft auftreten und keinen großen Informationsgewinn mit sich bringen. Beispiele dafür sind *und*, *weil*, *der* oder *als*. [Sla]

2.4.2 Stemming

Stemming is a method of word standardization used to match some morphologically related words. The stemming algorithm is a computational process that gathers all words that share

the same stem and have some semantic relation. [EAD09] vielfaches komprimiert. Stammformreduktion bei der Wörter auf ihren Stammform zurückgeführt werden und dadurch eine weitere Reduktion der Anzahl an Wörter im Text durchgeführt werden kann.

3 Problemspezifikation

Persönliche Daten sind im Internet oft frei zugänglich. Das heißt, dass unterschiedlichste Webseiten persönliche Information von Menschen öffentlich bereitstellen. Die bekanntesten Webseiten sind wahrscheinlich die Social Media Seiten wie Twitter, Facebook und Instagram. Allerdings wird auch auf anderen Webseiten personenbezogene Daten in großen Mengen bereitgestellt. Ein Beispiel dafür ist das Fußballportal “www.fupa.net“. Diese Art von Webseiten sind perfekte Informationsquelle für Phisher.
Im Bereich von Social Engineering Angriffen wird diese Information oft genutzt um ein Opfer zu täuschen oder manipulieren.
Dass hier beschriebene Problem zeigt, dass der Zugang für persönliche Information durch das Internet für die Öffentlichkeit einfacher gemacht wird. Es soll mit einem kritisch Blick darauf gezeigt werden, mit welchem Aufwand, personenbezogene Daten aus dem Internet herausgelesen, analysiert und für einen Phishing-Mail-Angriff verwendet werden kann.

4 Anforderungsanalyse und Priorisierung

4.1 Anforderungsanalyse

Die im Kapitel 1.2 definierten Ziele sollen mit den folgenden Anforderungen gewährleistet werden.

4.1.1 Anforderung an das Programm bzw. an die Programmiersprache

Es sollte eine möglichst einfache und sehr mächtige Skriptsprache verwendet werden, mit der Automatisierungen gut möglich sind. Eine Oberfläche kann vorhanden sein ist aber kein muss. Eine Eingabe über die Konsole oder über die GUI soll möglich sein.

4.1.2 Anforderung an die Informationsbeschaffung

Die Anforderungen an die Informationsbeschaffung von personenbezogenen Daten lässt sich in zwei Teile gliedern. Erstens in die Informationsbeschaffung von bestimmten bzw. ausgewählten Personen und zweitens die Informationsbeschaffung von einer großen Menge unbestimmter Personen.

Informationsbeschaffung von ausgewählten Personen

Bei dieser Informationsbeschaffung soll eine Suchfunktion entwickelt werden, welche Informationen zu einer angegeben Person sucht. Dies soll mit Hilfe eines Web-Crawlers und mit einem Web-Scraper umgesetzt werden. Das zu entwickelnde Tool soll bekannte Daten wie Vorname, Nachname, Geburtsjahr, Ort, Benutzernamen von Social Media Webseiten, usw.

über eine Konsolen-Abfrage einlesen können. Die Herausforderung besteht darin, zu erkennen, wann und ob es sich um die Information der gesuchten Person handelt.

Informationsbeschaffung von unbestimmten Personen

Es soll eine Prototyp-Suchfunktion entwickelt werden, die eine komplette Website durchsucht. Dabei sollen möglichst viele Informationen von vielen Personen herausgefunden werden. Jedoch sind diese Personen dem Programm-Anwender unbekannt. Die Informationen werden aus Webseiten mit einer großen Anzahl von Mitgliedern herausgelesen. Bei dieser Suchfunktion soll es möglich sein, aus vorgegebenen Webseiten eine auszuwählen und diese anschließend auszulesen und zu analysieren.

4.1.3 Anforderung an die Datenverwaltung/-speicherung

Ausgelesene Daten sollen vor dem speichern formatiert und klassifiziert werden, damit die Daten später korrekt in die Phishing-Mails eingesetzt werden können. Die Schwierigkeit besteht darin, zu erkennen, um welche Art von Information es sich handelt. Beispielsweise um ein Hobby oder Beruf. Zusätzlich sollen die Daten in einer gut übersichtlichen Struktur gespeichert werden und müssen beliebig erweiterbar sein.

4.1.4 Anforderung an die Generierung der E-Mail-Adressen

Da nicht zu jeder Suche eine E-Mail-Adresse im Internet gefunden werden kann, muss die E-Mail-Adresse aus den vorhandenen Informationen generiert werden. Es soll eine größere Anzahl von möglichen E-Mail-Adressen erzeugt werden. Durch den Pool an E-Mail-Adressen soll die Wahrscheinlichkeit erhöht werden, dass die richtige E-Mail-Adresse dabei ist. Des Weiteren kann die E-Mail-Adresse auf Verfügbarkeit und Gültigkeit geprüft werden.

4.1.5 Anforderung an die E-Mail-Muster

E-Mail-Muster sollen erstellt werden und so klassifiziert sein, dass für jedes gefundene Opferprofil ein passendes Muster vorhanden ist. Des Weiteren soll der E-Mail-Text mit den eingesetzten Informationen Sinn ergeben und eine korrekte Grammatik beinhalten.

4.1.6 Anforderung an die Erstellung der Phishing-Mail

Die Phishing-Mails sollen automatisiert erstellt werden. Die Auswahl des richtigen E-Mail-Musters zu der gewonnenen Opferinformation soll ebenfalls automatisiert ablaufen.

4.1.7 Unter anderem soll die Arbeit Antworten auf folgende Fragen finden

- Mit welchem Aufwand ist eine Phishing-Mail-Angriff verbunden?
- Ist es möglich ein Personenprofil zu erstellen, bei dem ausschließlich korrekte Informationen vorhanden sind?
-

4.2 Priorisierung

Die Tabelle 4.1 zeigt die Priorisierung der Anforderungen.

Tabelle 4.1: Priorisierung der Anforderungen

Anforderung	Priorisierung (A-C)
Informationsbeschaffung von ausgewählten Personen	*A*
Informationsbeschaffung von vielen ubekannten Personen	*A*
E-Mail-Muster erstellen	*A*
Phsishing-Mail erzeugen	*B*
Datenverwaltung/-speicherung	*B*

5 Lösungsideen

Für die Umsetzung der im Kapitel 1.2 definierten Ziele, werden folgende Lösungsideen vorgeschlagen.

5.1 Programmiersprache/ GUI

Für die Auswahl der Programmiersprache gibt es viele Auswahlmöglichkeiten. Dennoch wird in dieser Abschlussarbeit die Sprach Python verwendet, da sie die Nötigen Eigenschaften mit sich bringt.
Für die Eingabe von Suchdaten, besteht für beide Informationsbeschaffungen die Möglichkeit eine Grafische-Bedienoberfläche oder eine Konsolen-Eingabe zu verwenden.

5.2 Informationsbeschaffung einer ausgewählten Person

5.2.1 Wie sieht die Suche nach einer Person im Internet aus?

Die Suche nach einer Person im Internet kann durch mehrere Ansätze erfolgen. Die nachstehenden Ansätze unterscheiden sich in der Art Suche und in dem Umgang der eingeben Daten.

Die Art der Personensuche wird anhand den eingegebenen Daten angepasst

Abhängig von der Anzahl und Art der Daten, die von dem Programm-Anwender eingeben wurden, wird die Art und Reihenfolge der Suche variiert. Die nachfolgenden Fälle sollen diesen Ansatz verdeutlichen.

Im Fall, dass der Vorname, Nachname und Wohnort der gesuchten Person eingegeben wird, kann mit der Hilfe von herkömmlichen Suchmaschinen wie Google, Bing und DuckDuckGo nach Information gesucht werden. Die von den Suchmaschinen vorgeschlagenen Seiten werden anschließend analysiert, interpretiert und gespeichert. Dadurch können weitere Informationen gewonnen werden. Falls Benutzernamen von anderen Webseiten wie Instagram, Facebook oder ähnliches vorgeschlagen werden, kann somit die Suche mit diesen Daten speziell auf den entsprechenden Seiten erweitert werden.

Ein weiterer Fall beschreibt das Szenario, wenn ein Benutzername von der gesuchten Person in das Programm eingegeben wird. Hierbei handelt es sich um einen Benutzernamen von Webseiten wie Facebook, Instagram, usw.
Zuallererst, kann die entsprechende Webseite nach Informationen zu dem angegebenen Benutzername durchsucht werden. Dadurch können zusätzliche Daten herausgefunden werden, die bei der weiteren Suche von Vorteil wären.
Nachdem die Webseite nach dem Nutzernamen durchsucht und ausgewertet wurde, kann nun mit herkömmlichen Suchmaschinen die Suche erweitert werden.

Es wird unabhängig von den eingegebenen Daten direkt mit einer Suchmaschine nach der Person gesucht

Bei diesem Lösungsansatz werden ausschließlich die herkömmlichen Suchmaschinen verwendet. Die Funktion der Suche besteht darin, dass das Programm den vorgeschlagenen Links der Suchmaschinen folgt, wobei die eingegebenen Daten die Art der Suche nicht beeinflussen.

Nur ausgewählte Webseiten werden nach einer Person durchsucht

Unabhängig von den eingegebenen Daten, werden verschiedene Webseiten durchsucht. Allerdings ohne die Verwendung einer Suchmaschine. Vorschläge für die ausgewählten Webseiten sind Facebook, FuPa, Instagram, Xing, LinkedIn und Twitter.

5.2.2 Wann handelt es sich um die gleiche Person?

Bei jeder einzelnen Suchvariante, besteht die Herausforderung darin, zu erkennen, wann es sich um die gesuchte Person handelt. Durch die große Anzahl an verfügbaren Informationen im Internet, besteht eine hohe Wahrscheinlichkeit, dass Personen mit exakt den gleichen

Daten gefunden werden. Um dieses Problem zu umgehen, werden folgende Lösungsideen vorgeschlagen.

Die Art der Suche wird anhand den eingegebenen Daten angepasst

Diese Lösung entspricht dem Ansatz 5.2.1. Die Suche kann dadurch verfeinert werden und die Anzahl der fehlerhaften Vorschläge wird geringer. Dadurch wird die Wahrscheinlichkeit höher, dass es sich um die richtige Person handelt.

Bei keiner perfekten Übereinstimmung wird die Suche erweitert

Hier kann die Suche erweitert werden, indem auf soziale und berufliche Verbindungen der Zielperson eingegangen wird. Das heißt, dass bekannte Kontakte der gesuchten Person ebenfalls durchsucht werden. In diesem Fall könnten Facebook-Freunden, FuPa-Teammitglieder, Instagram-Follower oder LinkedIn/Xing-Kontakte als Kontaktquelle dienen.

Profilbilder können verglichen werden

Durch die Google Bildersuche, ist es möglich, anstatt einem Suchbegriff ein Bild zu verwenden und nach diesem zu suchen. Dabei kann ein zu suchendes Bild selbst hochgeladen oder ein URL angegeben werden. Bei dem Ergebnis kann es sich um ein ähnliches Bild oder eine Webseite, die das Bild enthält, handeln.
Des Weiteren kann eine Bilderkennungssoftware verwendet werden um gleiche Personen zu identifizieren.

Die Personensuche mit Hilfe von korrekten Suchbefehlen verfeinern

In dem Buch “Open Source Intelligence Techniques“ [Baz18], werden Suchbefehle für bekannte Suchmaschinen aufgezeigt, mit denen die Suche verbessert und verfeinert werden kann. Dies bedeutet, bei einer Personensuche ist es mit den richtigen Suchbefehlen möglich, die Anzahl der Vorschläge zu verringern. Ein Beispiel in dem Buch von Michael Bazzell zeigt, wie es funktioniert von 8770 Vorschlägen auf lediglich neun Vorschläge zu reduzieren. [Baz18]Dadurch wird auch bei dieser Lösungsidee die Wahrscheinlichkeit erhöht, dass es sich um die gesuchte Person handelt.

5.2.3 Wie erkennt das Programm wenn es sich um wichtige Informationen handelt?

Für die Suche einer ausgewählten Person können verschiedenste Arten von Webseiten vorgeschlagen werden. Aus diesem Grund muss das Programm eine gewisse Intelligenz beweisen um die wichtigsten Daten aus einer Seite herauszufiltern. Dabei ist es nicht möglich eine *Hartkodierung* zu verwenden und bestimmte Bereiche einer Webseite auszulesen. Die Grundidee zur Lösung diese Problems ist die Analyse des vorliegenden Textes durch verschiedenste Methoden.

Automated Keyword Extraction

Eine Methode zur Textanalyse ist die automatisierte Schlüsselwort-Gewinnung. Hierbei wird die HTML-Seite zu einem verwendbaren Text umgewandelt, wobei die meisten Sonderzeichen herausgefiltert werden. Sonderzeichen wie "." und "@" werden dabei nicht herausgefiltert, da sie für die E-Mail-Erkennung wichtig sind.
Die Anzahl der im Text befindenden Wörter werden anschließend mit Hilfe der sogenannten *Stoppwörter* und *Stammformreduktion* um ein großen sehr großen Teil reduziert.
Im darauffolgenden Schritt wird eine Liste der potentiellen Schlüsselwörter erstellt, welche nach der Häufigkeit des Vorkommens sortiert sind. Zusätzlich können weitere Listen erstellt werden, die N-Gramme des Textes enthalten.
Für die Erkennung wichtiger Schlüsselwörter werden Datenbanken bzw. Wortsammlungen erstellt, welche die zu suchenden Schlüsselwörter beinhalten. Mit diesen Datenbanken kann nun die Liste mit den bereits verarbeiteten Wörter verglichen werden. Die Datenbanken können mit Hilfe von bekannter Listen im Internet befüllt werden. Beispiele hierfür sind eine aktuelle Liste aller Hochschulen in Deutschland, Berufsbezeichnungen, Studiengänge, Hobbys, Städte und Gemeinden.

Textanalyse indem nach Schlüsselwörtern gesucht wird

Es kann ein Algorithmus entwickelt werden, der nach Schlüsselwörtern in einer Webseite sucht.

Textanalyse mit Hilfe Machine Learning

In der Theorie ist es möglich, ein Neuronales Netz mit den Begriffen zu trainieren und eine Kategorisierung durchzuführen. Dabei entsteht ein Netz, welches selbst entscheidet in welche Kategorie ein Wort fällt. Das Wort “Fußball“ müsste dadurch in die Kategorie Hobby eingeordnet werden.

Mit Hilfe von NLTK Rake den Text interpretieren

Rake hat die Aufgabe, einen Text mit vielen Wörtern auf eine geringe Anzahl von Schlüsselwörter zu reduzieren. Dadurch kann möglicherweise der Inhalt des Textes verstanden werden ohne ihn komplett gelesen zu haben.

5.2.4 Speicherung der gewonnenen Daten

Die gewonnenen Daten können in einem beliebig erweiterbaren Personen-Objekt erstellt werden. Erweiterungen von bekannten Kontakten sind ebenfalls möglich.

5.3 Informationsbeschaffung von einer großen Menge unbestimmter Personen

Webseiten mit großen Menge von Daten, ausgenommen von den bekannten Social Media Seiten, sind das Fußballportal FuPa, Xing und LinkedIn.

5.3.1 Informationsgewinnung durch Hartkodierung

Diese Suchfunktion wird *hartkodiert* und benötigt dadurch keine Textanalyse, da der Aufbau der Webseite im voraus bekannt ist. Das bedeutet, dass das Programm genau weiß wo welche Information auf einer Webseite steht. Beispielsweise befindet sich das Geburtsjahr einer Person, auf der Seite von dem Fußballportal “FuPa“, immer an der gleichen Position einer Tabelle. Dies bringt den Vorteil mit sich, dass der Text nicht analysiert werden muss und das Programm genau weiß, was mit diesen Daten gemacht werden muss.

5.3.2 Speicherung der gewonnenen Daten

Für die Speicherung von vielen unbekannten Personen-Daten kann eine SQL-Datenbank erstellt werden.
Als Alternative kann eine Datei angelegt werden, bei der alle Daten zu allen Personen gut strukturiert gespeichert werden können. Eine Möglichkeit dafür ist das Dateiformat *CSV* oder *TXT*.

5.4 Generierung der E-Mail-Adressen

Es kann das opensource tool von intelligencetechniques mit Hilfe eines automatisierten Webbrowsers verwendet werden. Algorithmus entwickeln, der alle möglichen Mail-Adressen aus den Daten Vorname, Nachname, Geburtsjahr und den bekanntesten Mail-Providern erzeugt.

5.5 Erstellung der E-Mail-Muster

Die Muster können in zwei große Kategorien unterteilt werden. Es gibt einen privaten und geschäftlichen Teil. Der private Teil hat weiter Unterteilungen wie Familie, Hobby/Interessen.

5.6 Erzeugung der Phishing-Mail

6 Auswahl der Lösung anhand den Anforderungen

6.1 Programmiersprache/ GUI

Python

6.2 Informationsbeschaffung von bestimmten/ausgewählten Personen

6.3 Informationsbeschaffung von einer großen Menge unbestimmter Personen

Einfachheitshalber wird CSV verwendet.

6.4 Generierung der E-Mail-Adressen

6.5 Erstellung der E-Mail-Muster

6.6 Erzeugung der Phishing-Mail

7 Umsetzung

7.1 Textanalyse mit Hilfe von Python NTLK

Mit dem *Natural Language Toolkit* ist es möglich, den vorhandenen Webseitentext zu analysieren. Zu Beginn können sogenannte “stopwords“ aus dem vorgegebenen Text herausgefiltert werden. Stopwords sind Wörter die sehr oft auftreten und keinen großen Informationsgewinn mit sich bringen. Beispiele dafür sind ist, ein, einer, usw. Dadurch verringert sich die Anzahl der gesamten Wörter im Text um einen sehr großen Teil. Anschließend können Funktionen wie das Zählen des Vorkommens einzelner Wörter angewendet werden, um einen Überblick von dem Text zu bekommen. Des Weiteren kann der Text in Fragmente zerlegt werden um weitere Informationen über den Inhalt zu erlangen. Abschließend kann eine Liste der analysierten Wörter bzw. Fragmente erstellt werden.
Für die Erkennung wichtiger Schlüsselwörter Es wäre denkbar, Datenbanken bzw. Wortsammlungen zu erstellen, welche die zu suchenden Schlüsselwörter beinhalten. Mit diesen Datenbanken kann nun die Liste mit den bereits verarbeiteten Wörter verglichen werden. Die Datenbanken können mit Hilfe von bekannter Listen im Internet befüllt werden. Beispiele hierfür sind eine aktuelle Liste aller Hochschulen in Deutschland, Berufsbezeichnungen, Studiengänge, Hobbys, Städte und Gemeinden, etc..

7.2 Informationsbeschaffung von der Website www.fupa.net

7.2.1 Erstellung eines Web Crawlers

Anforderung

Der Web Crawler soll die komplette Webseite www.fupa.net durchgehen und Links mit Spielerinformationen speichern. Die Funktionsweise des Web Crawlers besteht darin, dass das Programm auf der Startseite von Fupa.net beginnt nach links zu suchen und diesen folgt.

Probleme

1. Python hat einen verkürzten und erkennbaren Standard http-Header. Dieser wird von vielen Administratoren geblockt und mit der Fehlermeldung 451 erkennbar gemacht. 451 for legal reason
2. Honeypots gewollt oder ungewollt, hier Kalender darstellung mit links zu neuen Jahren die eine sehr hohe bis überhaupt keine Begrenzung haben.
3. Rekursion erreicht schnell die Maximale tiefe von 1500.
4. Zu langsamer Algorithmus

Lösungen

1. http-Header selber konfigurieren
2. Links mit möglichen Honeypots nicht beachten
3. Stack Klasse schreiben damit keine Rekursion benötigt wird
4. Algorithmus anpassen auf fupa-Webseite

7.3 Datenverwaltung und Speicherung

7.3.1 Speicherung von Personendaten in CSV oder mySQL

8 Hauptteil

8.1 Hauptteil

8.1.1

9 Hauptteil

9.1 Hauptteil

9.1.1

10 Hauptteil

10.1 Hauptteil

10.1.1

11 Hauptteil

11.1 Hauptteil

11.1.1

12 Schlussbemerkungen und Ausblick

A Ein Kapitel des Anhangs

Glossar

Active Directory

Active Directory ist in einem Windows Server 2000, Windows Server 2003, oder Windows Server 2008-Netzwerk der Verzeichnisdienst, der die zentrale Organisation und Verwaltung aller Netzwerkressourcen erlaubt. Es ermöglicht den Benutzern über eine einzige zentrale Anmeldung den Zugriff auf alle Ressourcen und den Administratoren die zentral organisierte Verwaltung, transparent von der Netzwerktopologie und den eingesetzten Netzwerkprotokollen. Das dafür benötigte Betriebssystem ist entweder Windows Server 2000, Windows Server 2003, oder Windows Server 2008, welches auf dem zentralen Domänencontroller installiert wird. Dieser hält alle Daten des Active Directory vor, wie z.B. Benutzernamen und Kennwörter. 3

Glossareintrag

Erweiterte Informationen zum einem Wort oder einer Abkürzung, ähnlich einem Eintrag im Duden. 3

AbkÃ¼rzungsverzeichnis

AD Active Directory 3

Symbolverzeichnis

π	Die Kreiszahl. 3

Literatur

[All18] Allensbach, IfD: *Meistgenutzte Informationsquellen der Bevoelkerung in Deutschland im Jahr 2018.* https://de.statista.com/statistik/daten/studie/171257/umfrage/normalerweise-genutzte-quelle-fuer-informationen/, 2018. Abrufdatum: 18.01.2019.

[Baz18] Bazzell, Michael: *Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information.* CreateSpace Independent Publishing Platform, USA, 6th , 2018.

[Bun18] Bundeskriminalamt: *Polizeilich erfasste Fälle von Cyberkriminalität im engeren Sinne* in Deutschland von 2004 bis 2017.* https://de.statista.com/statistik/daten/studie/295265/umfrage/polizeilich-erfasste-faelle-von-cyberkriminalität-im-engeren-sinne-in-deu 2018. Abrufdatum: 29.10.2018.

[Cal13] Caldwell, Tracey: *Spear-phishing: how to spot and mitigate the menace.* Computer Fraud & Security, 2013(1):11–16, 2013.

[CH15] Christopher Hadnagy, Michele Fincher: *Phishing Dark Waters: The Offensive and Defensive Sides of Malicious E-mails.* 2015.

[DSG] DSGVO: *Art. 4 DSGVO Begriffsbestimmungen.* https://dsgvo-gesetz.de/art-4-dsgvo/. Abrufdatum: 09.01.2019.

[EAD09] Eldesouki, Mohamed I, W Arafa K Darwish: *Stemming techniques of Arabic language: Comparative study from the information retrieval perspective.* The Egyptian Computer Journal, 36(1):30–49, 2009.

[Had11] Hadnagy, Christopher: *Social Engineering: The Art of Human Hacking.* 2011.

[Jam05] James, Lance: *Phshing Exposed: Uncover Secrets from the Dark Side.* 2005.

[Mit01] Mitnick, Kevin D.: *The art of deception:controlling the human elemnet of security.* 2001.

[Mit15] Mitchell, Ryan: *Web Scraping with Python: Collecting Data from the Modern Web.* 2015.

[NW18] Nordrhein-Westfalen, Verbraucherzentrale: *Phishing-Radar: Aktuelle Warnungen*. https://www.verbraucherzentrale.nrw/wissen/digitale-welt/phishingradar/phishingradar-aktuelle-warnungen-6059, 2018. Abrufdatum: 29.10.2018.

[Sla] Slavin, Tim: *Stop Words*. https://www.kidscodecs.com/stop-words/. Abrufdatum: 29.01.2019.

[uDsiNe15] Netz e.V., DATEV und Deutschland sicher im: *Verhaltensregeln zum Thema "Social Engineering"*. 2015.