2020

Digitalna forenzika

Forenzika e-mail saobraćaja

Ana Stojanović 1087

Sadržaj

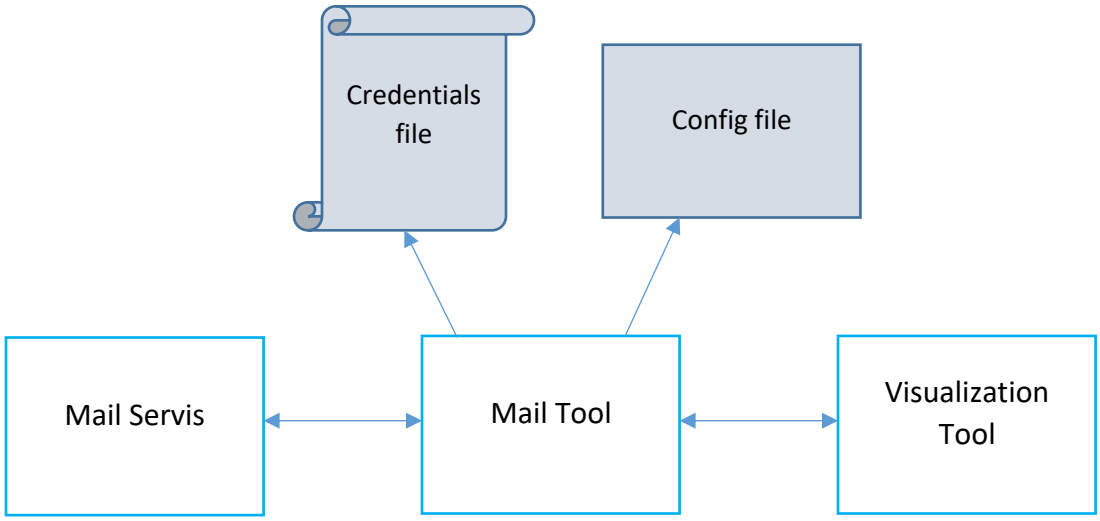
Pregled teme

U ovom radu će biti obrađen alat za obavljanje forenzike e-mail naloga. Ovaj alat je realizovan u programskom jeziku Python. E-mail nalogu se pristupa korišćenjem IMAP protokola i potrebno je unapred znati e-mail i šifru naloga kojem se pristupa.

Fukcionalnosti koje dati alat implementira su:

- Grafički prikaz broja poslatih mail-ova za određen vremenski period (Po satima, danima i mesecima)
- Grafički prikaz domena kojima dati mail nalog šalje poruke
- Grafički prikaz najkorišćenijih ključnih reči u telu i subjektu poslatih poruka
- Grafički prikaz e-mail kontakta sa kojima dati e-mail nalog komunicira (poslate i primljene poruke)

Arhitektura Sistema



Sistem sadrži tri glavne komponente: Mail Servis, Mail Tool i Visualization Tool. Mail Servis komponenta služi za pristup mail serveru i pribavljanje mailova, mail tool vrši obradu pribavljenih mailova i izvlači potrebne informacije, dok Visualization Tool služi za vizuelizaciju pribavljenih podataka.

Pokretanje koda

Za pokretanje sistema potreban je Python 3. Sve zavisnosti projekta su navedene u fajlu requirements.txt.

Prosleđivanje kredencijala sistemu

Kredencijali se smeštaju u fajl credentials.text, tako da je e-mail adresa na prvom, a šifra na drugom mestu. Email Tool dinamički čita kredencijale sa ove lokacije.

Konfiguracija sistema

Konfiguracioni fajl sistema se naziva config.py. U ovom fajlu se mogu podesiti mail server, mail port, kao i ime e-mail sandučića kojima se pristupa.

Mail servis

Mail servis služi kao omotač oko Python implementacije IMAP protokola. Omogućava da korisnik otvori konekciju ka mail serveru, autentikuje se, odabere sanduče, kao i da pribavi skup poruka za dati vremenski period. Na kraju interakcije, moguće je zatvoriti sanduče i konekciju ka serveru. Skup informacija koje se vraćaju za jednu poruku su:

- Pošiljalac, Primalac, CC, BCC
- Subject
- Datum slanja
- Tekstualno telo poruke

```
status, messages = self.__imap.search(None, f'(SINCE "{start}" BEFORE "{end}")')
message_info = []

for message_id in messages[0].split():
    status, data = imap.fetch(message_id, '(RFC822)')
    for response_part in data:
        if isinstance(response_part, tuple):
            message = message_from_bytes(response_part[1])
            subject = parse_subject(message['Subject'])
            sender = parse_email(message['From'])[0]
            recievers = parse_email(message['To'])
            cc = None
            if 'CC' in message:
                cc = parse_email(message['CC'])
            bcc = None
            if 'BCC' in message:
                bcc = parse_email(message['BCC'])
            date = __parse_email_datetime(message['Date'])
            body = _parse_message_body(message)
            dict = {'Sender': sender, 'Recievers': recievers, 'CC': cc, 'BCC': bcc, 'Date': date,
                    'Subject': subject, 'Text-Body': body}
            message_info.append(dict)
```

Isečak koda u Mail Servisu koji pribavlja i parsuje mail poruke

IMAP Protokol

IMAP je skraćenica za “Internet Message Access Protocol”. IMAP je protokol koji omogućava korisniku da pristupa i manipuliše elektronskim sandučićima na udaljenom mail serveru, ali ne omogućava slanje poruka. IMAP omogućava kreiranje, brisanje i preimenovanje sandučića, proveravanje poruka, trajno brisanje poruka, kao i selektivno preuzimanje delova poruka. U IMAP protokolu, porukama se pristupa korišćenjem identifikatora. Identifikatori mogu biti sekvencijalni ili jedinstveni. Jedinstveni identifikatori se dodeljuju svakoj poruci u sandučetu i moraju biti jedinstveni barem tokom jedne sesije između klijenta i servera (preporučljivo je da uvek budu

jedinstveni). Sekvencijalni identifikatori označavaju poziciju u sandučetu. Ukoliko se poruka obriše, sekvencijalni identifikatori se ponovo dodeljuju.

Da bi se omogućila komunikacija između klijenta i servera, potrebno je otvoriti komunikacioni kanal. Interakcija između klijenta i servera se sastoji od komandi i odgovora.

Komanda LIST

Komanda "LIST" se može iskoristiti da se pregleda skup dostupnih sandučića. Ova komanda je validna u autentikovanom stanju.

Komanda SEARCH

Komanda "SEARCH" se može koristiti za pribavljanje identifikatora mail poruka na osnovu određenih kriterijuma, tako da je dobijeni rezultat presek svih kriterujma. Server kao odgovor na komandu vraća sekvencijalne identifikatore poruka koje zadovoljavaju ovu komandu. Primeri kriterijuma koji se mogu koristiti za pretragu su "BEFORE" i "SINCE", koje se koriste za pretraživanje poruka u određenom vremenskom intervalu.

Komanda FETCH

Komanda fetch se koristi za pribavljanje pojedinačne e-mail poruke. Fetch komanda može pribaviti celu ili delove poruke, uključujući telo i zaglavlja. Ukoliko se specificira 'RFC-822', vraća se celokupna poruka.

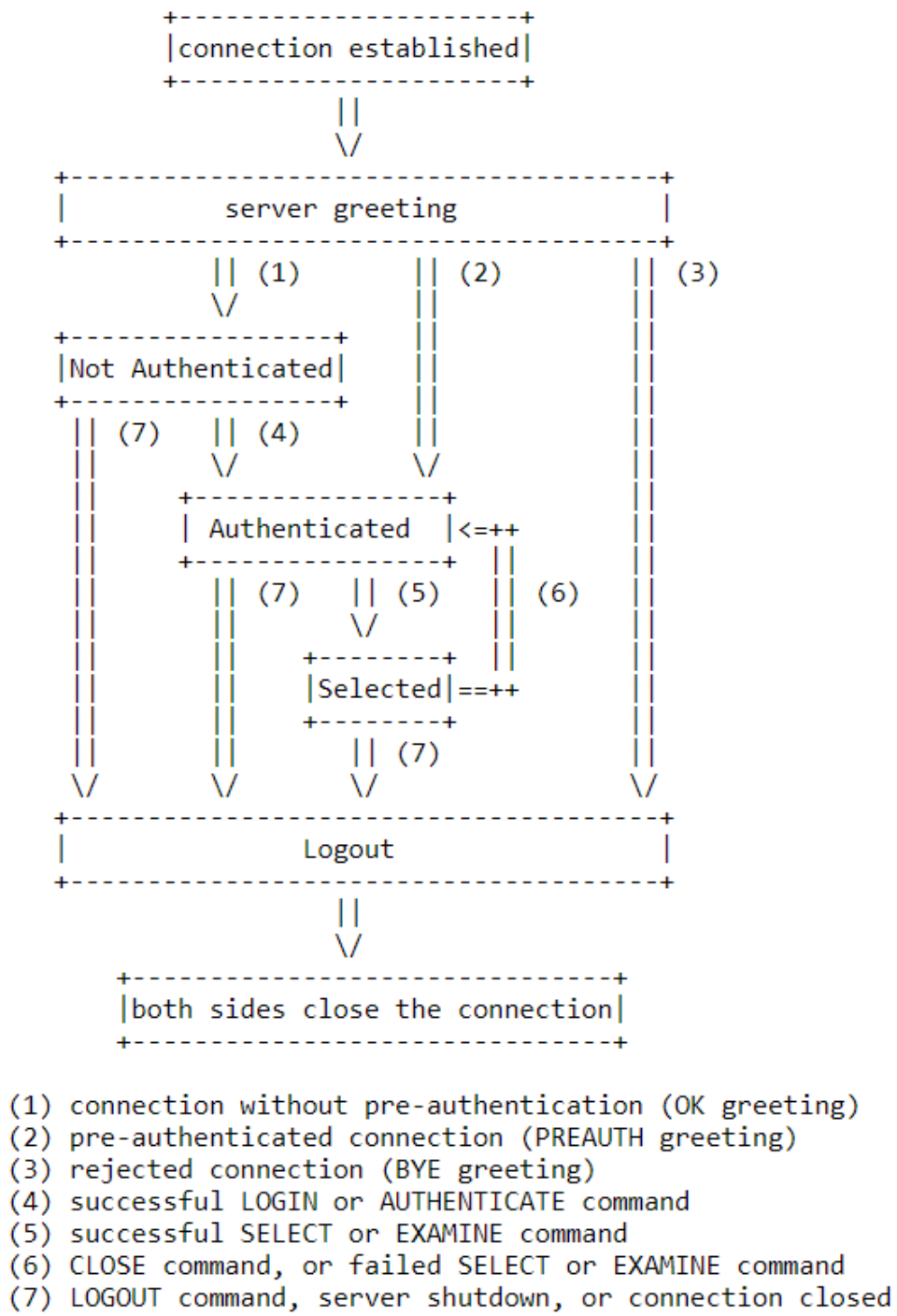
Autentikacija kod IMAP protokola

Postoji grupa stanja u kojima može da se nalazi klijent-server konekcija. Inicijalno stanje se uspostavlja prilikom konekcije između klijenta i servera. Stanja u kojima konekcija može da se nađe su:

- Not Autenticated (Nije autentikovan)
- Authenticated (Autentikovan)
- Selected (Selektovan)
- Logout

U neautentikovanom stanju, klijent izdaje "LOGIN" komandu i prosleđuje validne kredencijale (username i password) pre nego što su dalje komande dozvoljene. Klijent prelazi u autentikovano stanje izdavanjem "AUTHENTICATE" komande. U autentikovanom stanju, klijent mora da

odabere sanduče pre nego što može da izdaje komande koje rade sa porukama (“SELECT” komanda). U selektovanom stanju, klijent je spreman da pristupa porukama u sandučetu. Ovo stanje se može prekinuti „CLOSE“ komandom (u tom slučaju se konekcija vraća nazad u autentikovano stanje). Da bi se u potpunosti prekinula konekcija, potrebno je izdati “LOGOUT” komandu.



Tok prilikom IMAP autentikacije. Izvor slike: (1)

Mail Tool

Mail tool je komponenta koja se bavi forenzičkom obradom e-mail poruka.

Izračunavanje broja poslatih poruka po mesecima, danima i satima

Moguće je prebrojati poslate poruke na mesečnom nivou za datu godinu, na dnevnom nivou za dati mesec, kao i na a nivou sati za dati dan.

```
def count_sent_messages_hourly(self, day):
    period_start = datetime(day.year, day.month, day.day)
    period_end = period_start + relativedelta(days=+1)
    messages = self.__get_sent_messages(period_start, period_end)

    time_dictionary = self.__generate_time_dictionary(period_start, period_end, "Hourly")
    for message in messages:
        key = message['Date'].strftime("%Y-%m-%d %H:00")
        time_dictionary[key] += 1
    return time_dictionary
```

Isečak koda koji prikazuje pribavljanje poruka po satima

Izračunavanje broja poslatih poruka po domenima

Moguće je naći broj poruka koji je poslat određenim domenima za dati vremenski period. Korišćenjem regularnih izraza je moguće izdvojiti domen iz mail adrese.

```
def count_sent_messages_by_domain(self, period_start, period_end):
messages = self.__get_sent_messages(period_start, period_end)
domain_dict = defaultdict(int)
for message in messages:
    recievers = message['Recievers']
    if "BCC" in message and message['BCC']:
        recievers.extend(message['BCC'])
    if "CC" in message and message['CC']:
        recievers.extend(message['CC'])
    domains = self.__parse_domains(recievers)
    for domain in domains:
        domain_dict[domain] += 1
sorted_dict = self.__sort_dictionary_by_value(domain_dict)
return sorted_dict
```

Isečak koda koji prikazuje brojanje poslatih poruka po domenu

Izračunavanje najćešće korišćenih ključnih reči

Funkcija koja računa najkorišćenije ključne reči u poslatim porukama je data na slici. Pribavljaju se 'Subject' i 'Body' delovi poruke i iz njih se korišćenjem regularnih izraza uklanjaju svi linkovi i tagovi oblika <tag>. Zatim se koristi "word_tokenize" funkcija iz Python nltk (natural language toolkit) paketa za pribavljanje tokena.

```
def count_most_used_keywords(self, period_start, period_end):
    messages = self.__get_sent_messages(period_start, period_end)
    token_dict = defaultdict(int)
    for message in messages:
        text_body = message['Text-Body']
        subject = message['Subject']
        text_to_process = None
        if text_body and subject:
            text_to_process = text_body + " " + subject
        elif text_body:
            text_to_process = text_body
        else:
            text_to_process = subject

        if text_to_process:
            text = re.sub('(https|http){1}:\/\/[^\s]+\.[\w\d]+\/{0,1}', '', text_to_process, flags=re.MULTILINE)  # remove links
            text = re.sub('<.*>', '', text, flags=re.MULTILINE)
            tokens = word_tokenize(text)
            tokens = list(filter(lambda token: token not in string.punctuation and token not in ['``', '\'\'', '\"\"', '...', ], tokens))
            for token in tokens:
                token_dict[token] += 1

    sorted_dict = self.__sort_dictionary_by_value(token_dict)
    return sorted_dict
```

Isečak iz koda koji prikazuje izdvajanje najkorišćenijih ključnih reči

Izračunavanje jačine veze između kontakta

Računanje jačine veze između kontakta podrazumeva nalaženje jačine veze između poznatog mail naloga i kontakta za dati vremenski period. Kontaktom se smatra svaki nalog s kojim je poznat mail nalog interagovao na bilo koji način (sent, recieved, cc, bcc).

U prvoj fazi izračunavanja, pribavljaju se svi poslati i primljeni mailovi za dati period i izvlače se potrebne informacije za svakog kontakta:

- First-Contact (Prvi kontakt) – Datum u okviru odabranog vremenskog perioda kada su kontakt i nalog prvi put interagovali
- Last-Contact (Poslednji kontakt) – Datum u okviru odabranog vremenskog perioda kada su kontakt i nalog poslednji put interagovali
- From-Me (Od mene) – Broj poruka koje je kontakt primio direktno od poznatog naloga
- From-Me-CC (Od mene CC) – Broj poruka koje je kontakt primio od poznatog naloga u CC polju
- From-Me-BCC (Od mene BCC) – Broj poruka koje je kontakt primio od poznatog naloga u BCC polju
- To-Me (Od mene) – Broj poruka koje je poznat nalog primio direktno od kontakta
- To-Me-CC (Od mene CC) – Broj poruka koje je nalog primio od kontakta u CC polju

- To-Me-BCC (Od mene BCC) – Broj poruka koje je nalog primio od kontakta u BCC polju
- To-Me-Groups (Od mene grupe) – Ukoliko se e-mail adresa poznatog naloga ne nalazi ni u jednom od gore pomenutih polja, poruka se svrstava kao masovna tj. grupa i smatra se da je dobijena koliko se e-mail adresa poznatog naloga ne nalazi ni u jednom od gore pomenutih polja, poruka se svrstava kao masovna tj. grupa i smatra se da je dobijena zbog članstva u nekoj grupi

U drugoj fazi izračunavanja, računaju se parametri veze[1]:

- Ukupan broj primljenih poruka – Suma primljenih poruka u to, cc, bcc i groups poljima
  - num_{recv}
- Ukupan broj poslatih poruka – Suma poslatih poruka u from, cc i bcc poljima
  - num_{sent}
- Skorašnjost veze – Broj dana proteklih od poslednjeg kontakta do današnjeg dana
  - $recen = now - LastContact$
- Dužina veze
  - $len = LastContact - FirstContact$
- Frekvencija slanja poruka – Predstavlja koliko često poznat mail nalog šalje poruke kontaktu
  - $F_{sentfreq} = \frac{num_{sent}}{len}$
- Frekvencija primanja poruka – Predstavlja koliko često kontakt šalje poruke poznatom nalogu
  - $F_{recvfreq} = \frac{num_{recv}}{len}$
- Odnos primljenih poruka i ukupnog broja poruka
  - $F_{to} = \frac{num_{to}}{num_{total}}$
- Odnos poslatih poruka i ukupnog broja poruka
  - $F_{from} = \frac{num_{from}}{num_{total}}$
- Odnos broja poslatih/primljenih sekundarnih poruka (CC, BCC, Groups) i ukupnog broja poruka
  - $F_{sec} = \frac{num_{sec}}{num_{total}}$
- Reciprocitet veze – Prikazuje odnos između poslatih i primljenih poruka
  - $Recip = 1 - \frac{|num_{recv} - num_{sent}|}{num_{total}}$

[1] Deo parametara veze preuzet iz (2)

U trećoj fazi se normalizuju parametri $F_{sentfreq}, F_{recvfreq}$ i $recen$.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Normalizacija

Nakon toga se računa težina veze po sledećoj formuli[2]:

$$w_1\left(F_{recip} + F_{from} + F_{sentfreq}\right) + w_2\left(F_{to} + F_{recvfreq}\right) + w_3(recen + F_{sec}),$$

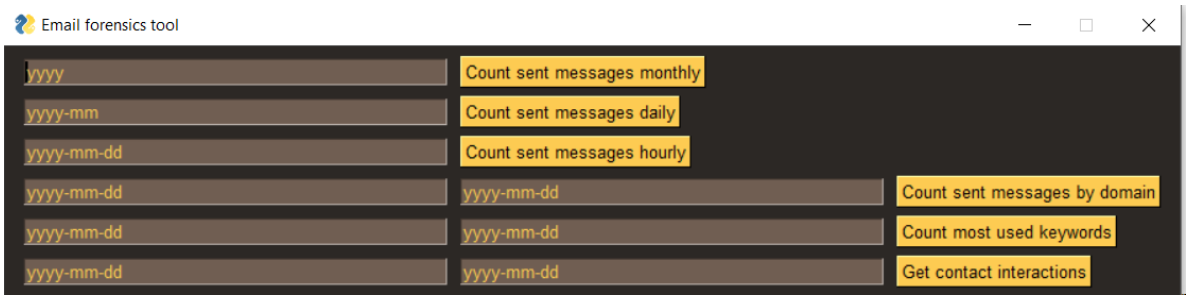
gde su $w_1 = 1, w_2 = 0.5, w_3 = 0.3$ težinski faktori.

Za veću težinu veze se smatra da je veća jačina kontakta između poznatog mail naloga i kontakta koji se proučavaju.

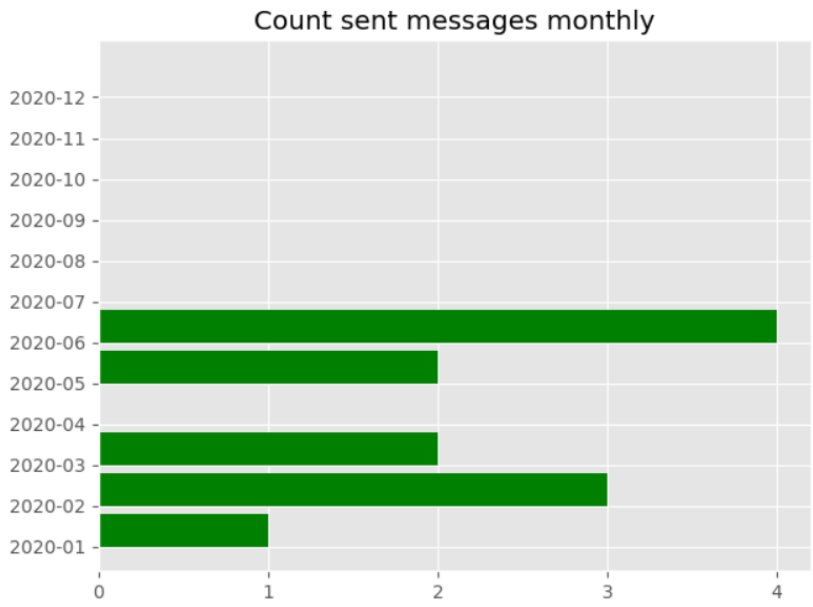
[2] Deo formule preuzet iz (2)

Visualisation Tool

Visualisation tool je komponenta koja se koristi za vizuelizaciju statistike za poznat mail nalog. Implementiran je korisnički interfejs korišćenjem PySimpleGui biblioteke, dok se grafikoni crtaju korišćenjem matplotlib biblioteke. Pregled statistike se ograničava na prvih trideset rezultata.



Izgled grafičkog interfejsa aplikacije



Izgled matplotlib grafikona

Reference

1. RFC 3501. https://tools.ietf.org/html/rfc3501 - IMAP Protocol

2. Tansel Özyer, Jon Rokne, Gerhard Wagner, Arno H.P. Reuser. *The Influence of Technology on Social Network Analysis and Mining.* Springer-Verlag/Wien : Springer, 2013.