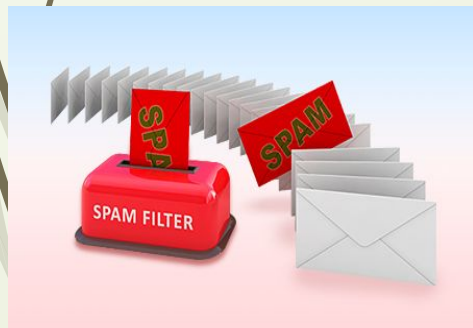


Universitatea POLITEHNICA
Licență 2019

PREZENTARE

APLICAȚIE DE FILTRARE ȘI CLASIFICARE DE E-MAILURI



STUDENT: DIGORI GHEORGHE
COORDONATOR ȘTIINȚIFIC: CONF.DR.ING. BOICEA ALEXANDRU



PREZENTAREA GENERALĂ A APLICAȚIEI

În cadrul acestui proiect de licență mi-am propus să dezvolt o aplicație web ce are ca scop filtrarea e-mailurilor în baza unor filtre implementate de mine care se configurează manual.

De asemenea, am implementat 2 clasificatoare ce au la bază Învățarea automată pentru a oferi utilizatorilor aplicației opțiunea de analizare și clasificare de email-uri în SPAM și HAM în baza conținutului acestora.

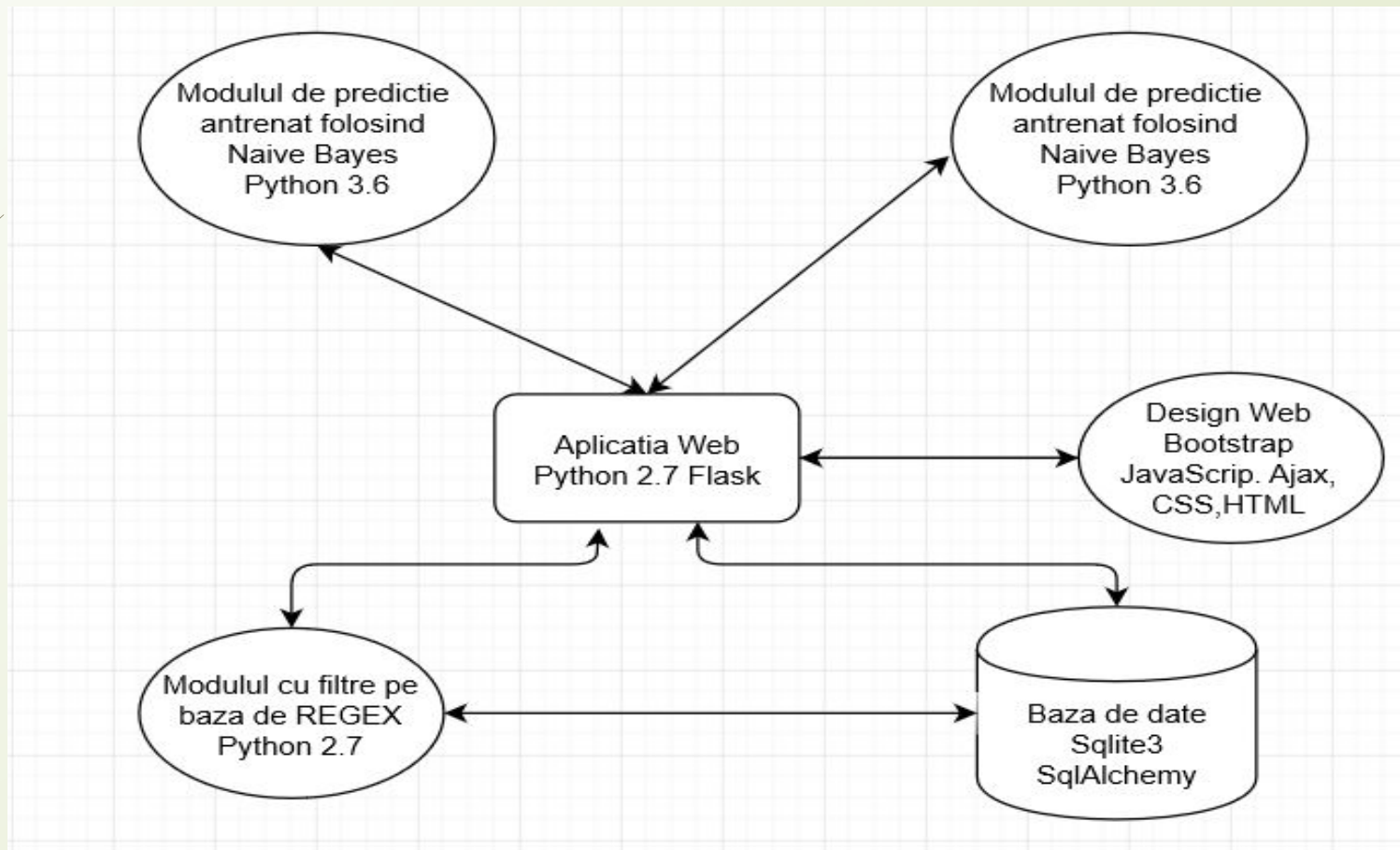


TEHNOLOGII FOLOSITE

Pentru dezvoltarea sistemului de filtrare si clasificare am folosit următoarele tehnologii:

- Pentru modulele de machine learning(Naïve BAYES si Tensorflow): Python 3.6, Tensorflow
- Pentru modulul de back-end server side: librăria Flask din Python 2.7
- Pentru modulul de front-end a aplicației web: Bootstrap 4 (include module de HTML CSS JQUERY si JAVASCRIPT)

Arhitectura aplicației SpamSherlock

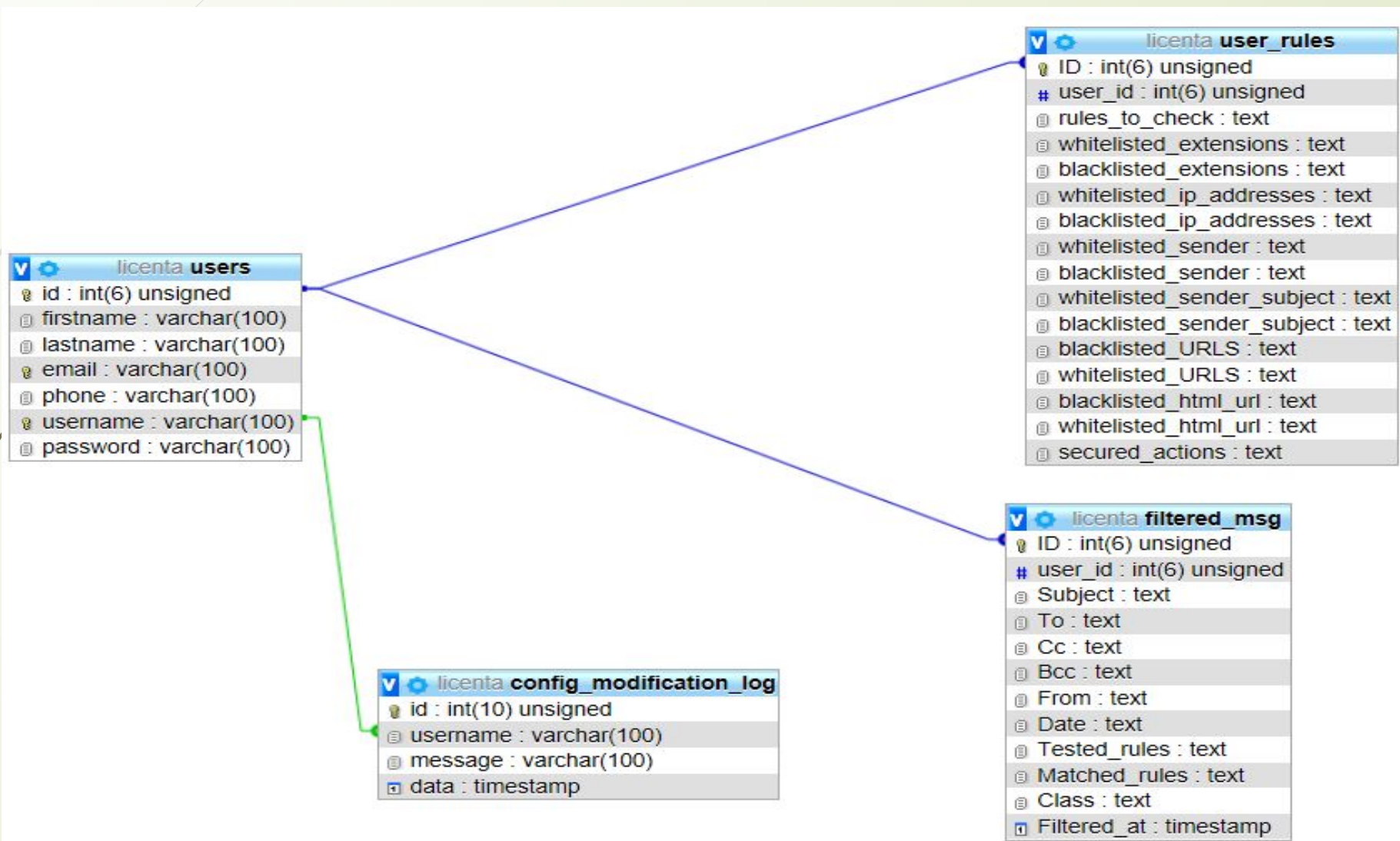




BAZA DE DATE

- Pentru stocarea persistentă a datelor folosesc SQLITE3
- Configurarea si crearea tabelelor este realizată folosind librăria de Python(ORM) – SQLALCHEMY
- Aplicația folosește 4 tabele interconectate:
 - Tabela în care se stochează utilizatorii
 - Tabela cu filtrele configurate pentru fiecare utilizator
 - Tabele cu rezultatele filtrării e-mailurilor.
 - Tabela cu logurile filtrării aplicației.

SCHEMA BAZEI DE DATE





PREZENTAREA APLICAȚIEI

Aplicația rulează pe portul 5000, utilizatorul se poate conecta prin intermediul unui browser folosind linkul: <http://localhost:5000/login>.

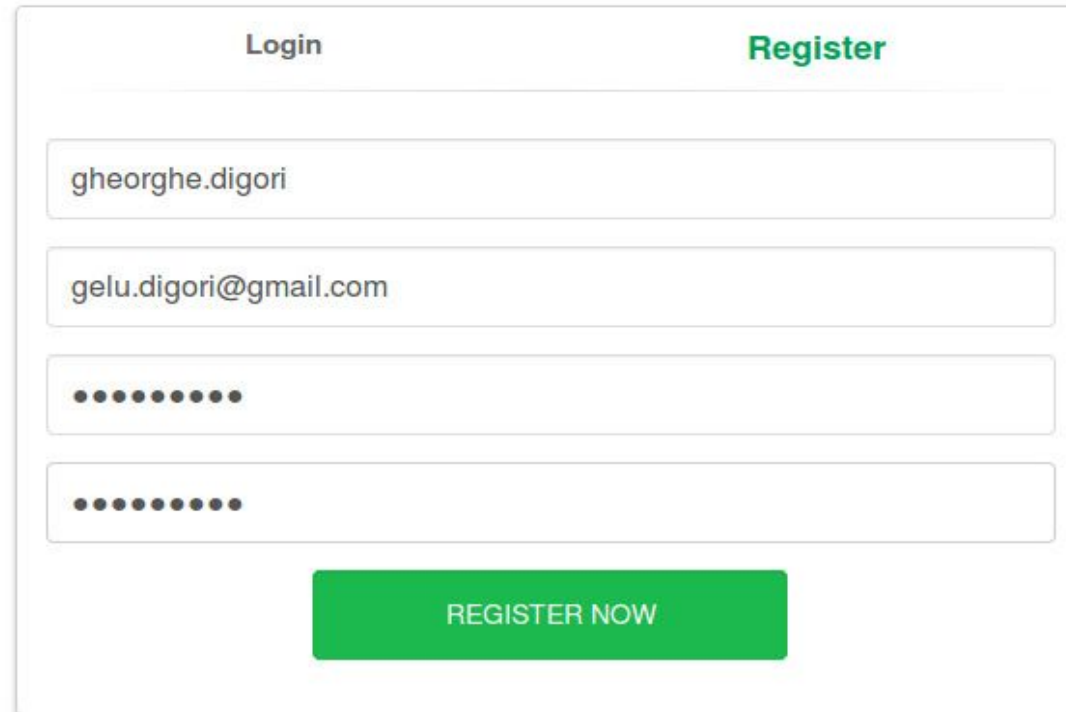
Utilizatorul are la dispoziție opțiunea de logare în aplicație pe baza unui cont existent sau poate crea un cont nou.



ÎNREGISTRARE + LOGARE

- Pentru înregistrarea unui cont nou, utilizatorul trebuie să folosească o adresă de email, un username - valide și o parolă care să îndeplinească criteriile specificate.
- Parolele user-ilor sunt criptate în baza de date din motive de securitate.
- Logarea in aplicație se realizează prin intermediul unor sesiuni securizate ce folosesc librăria Python FLASK-LOGIN

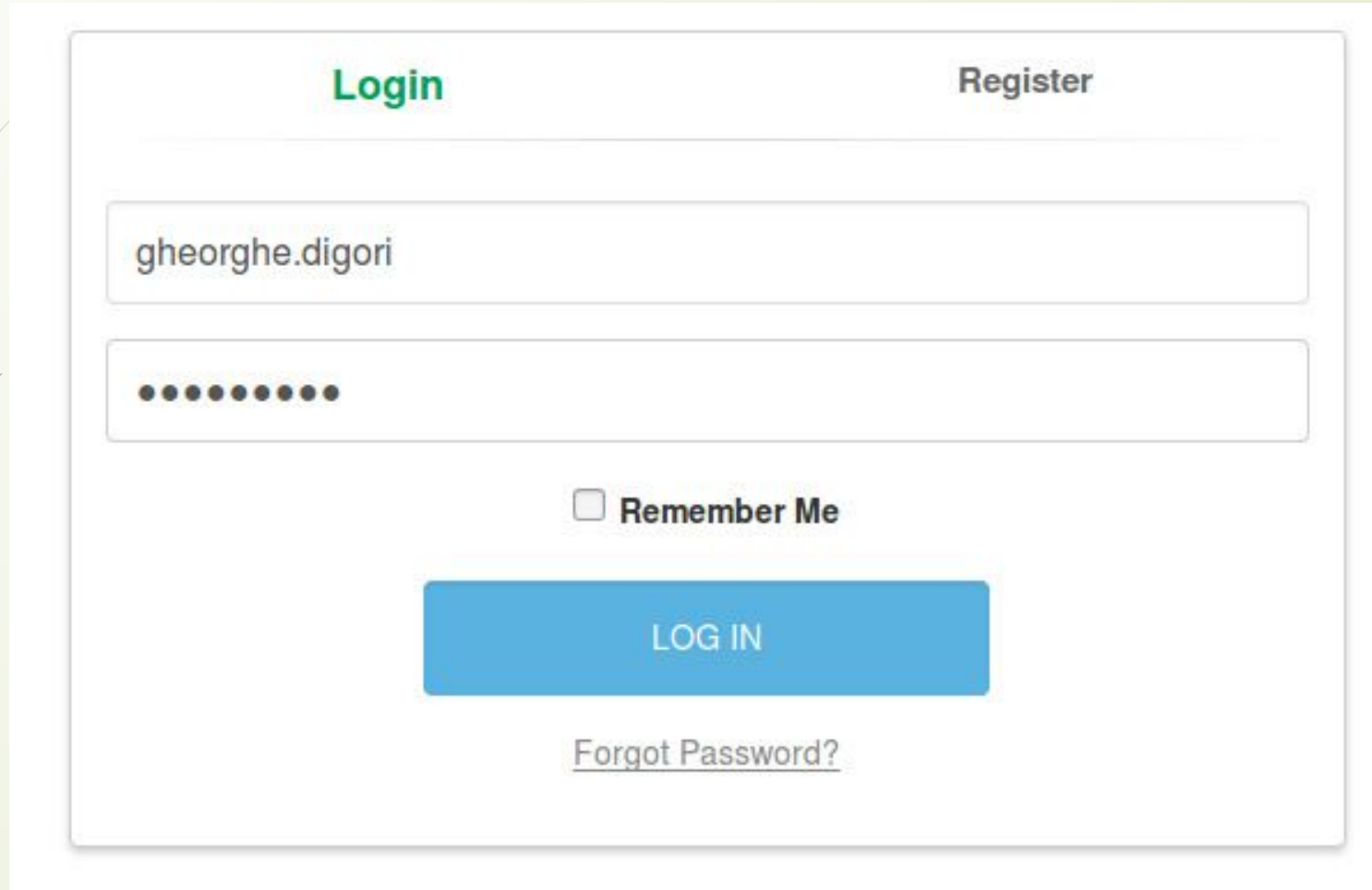
Înregistrare cont nou



The registration form is a white rectangular box with a thin grey border. At the top, it has two tabs: 'Login' and 'Register'. The 'Register' tab is highlighted in green. Below the tabs are four input fields: a username field containing 'gheorghe.digori', an email field containing 'gelu.digori@gmail.com', and two password fields, each containing ten dots. At the bottom of the form is a green button with the text 'REGISTER NOW' in white capital letters.

1. Usernames must be less than 20 characters
2. Email address must be valid
3. Passwords must match email password and contain the following:
 - Minimum of 8 characters.
 - Contain a number.
 - Contain upper case character.
 - Contain lower case character.
 - Contain special character.

Logare



The image shows a login form with a white background and a subtle shadow. At the top, there are two tabs: 'Login' in green text and 'Register' in grey text. Below the tabs are two input fields. The first field contains the text 'gheorghe.digori'. The second field contains ten black dots, representing a password. Below the password field is a checkbox labeled 'Remember Me'. At the bottom of the form is a blue button with the text 'LOG IN' in white. Below the button is a link that says 'Forgot Password?'.

Login Register

gheorghe.digori

.....

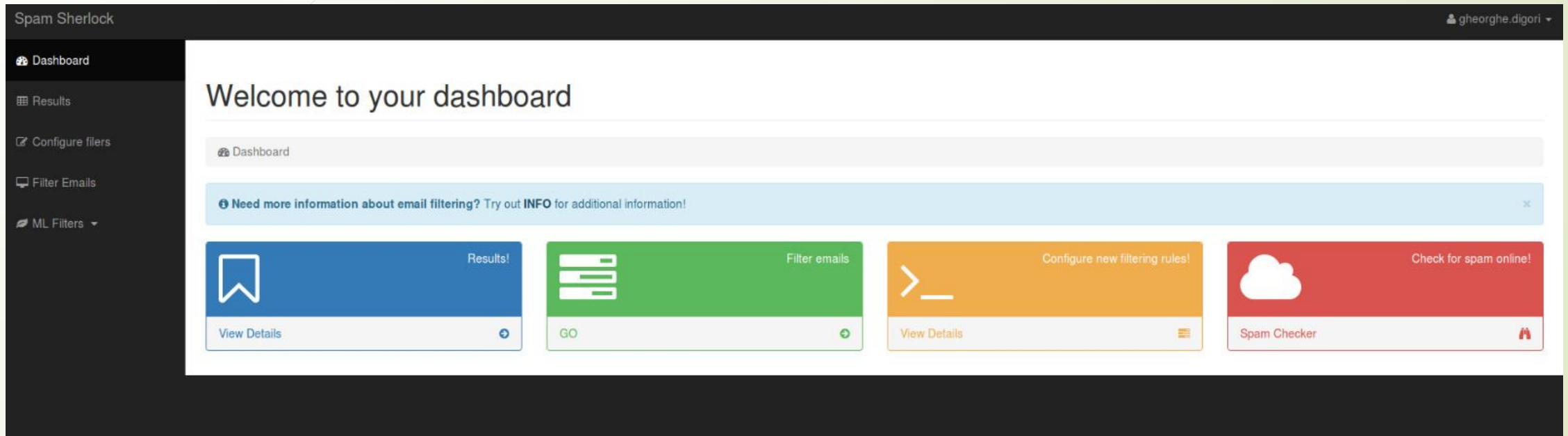
☐ Remember Me

LOG IN

[Forgot Password?](#)

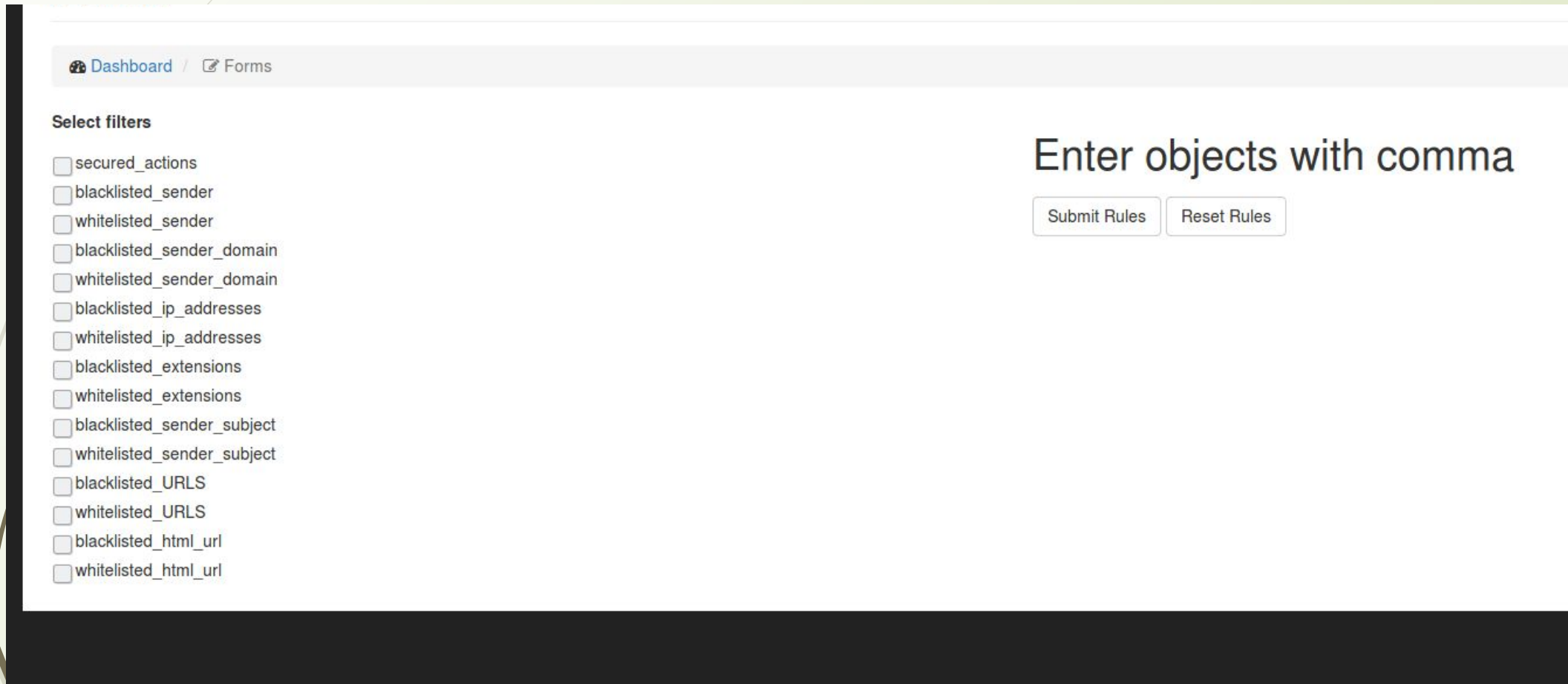
După autentificarea cu succes în aplicație utilizatorul este întâmpinat de un dashboard personalizat.

DASHBOARD



- Utilizatorului îi sunt puse la dispoziție 2 metode de conectare la celelalte pagini – NavBar-ul din stânga precum și tag-urile din dashboard.

PAGINA DE CONFIGURARE A FILTRELOR PERSONALIZATE





Dashboard / Forms

Select filters

- ☐ secured_actions
- ☐ blacklisted_sender
- ☐ whitelisted_sender
- ☐ blacklisted_sender_domain
- ☐ whitelisted_sender_domain
- ☐ blacklisted_ip_addresses
- ☐ whitelisted_ip_addresses
- ☐ blacklisted_extensions
- ☐ whitelisted_extensions
- ☐ blacklisted_sender_subject
- ☐ whitelisted_sender_subject
- ☐ blacklisted_URLS
- ☐ whitelisted_URLS
- ☐ blacklisted_html_url
- ☐ whitelisted_html_url

Enter objects with comma

Submit Rules Reset Rules

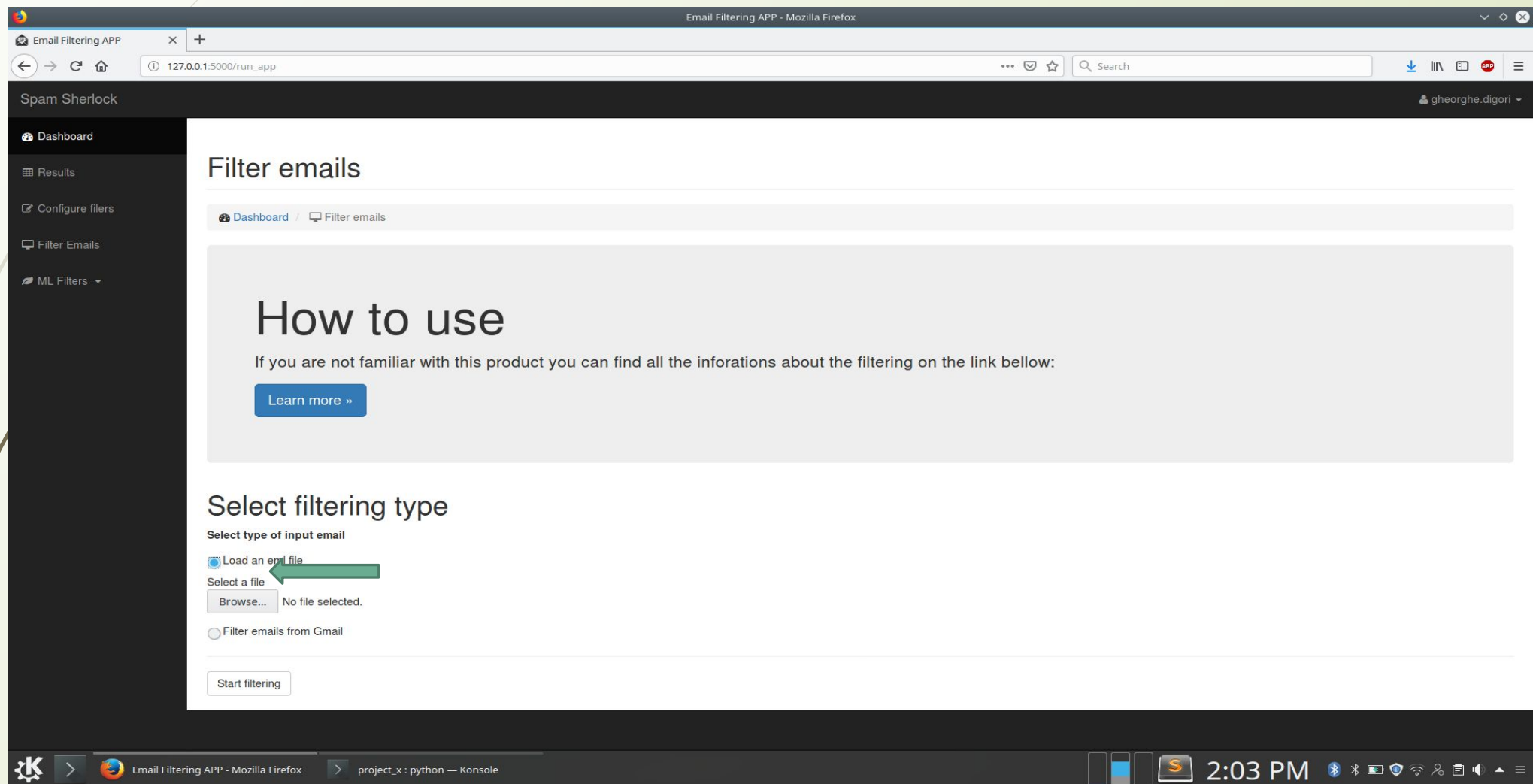
- 
- 
- În cadrul acestei pagini utilizatorul îi sunt puse la dispoziție un set de 15 filtre configurabile pe baza de REGEX, majoritatea de tip whitelist/blacklist.
 - Exista si 4 filtre default(care nu pot fi configurate de utilizator)
 - Filtrele default scanează e-mailul pentru GTUBE-uri executabile în html-uri și pentru apartenența la social media(FACEBOOK, TWETER, TINDER, SNAPCHAT, LINKEDIN și altele).
 - Filtrele scaneaza mesajele atât la nivel de conținut, cât și de header-e.



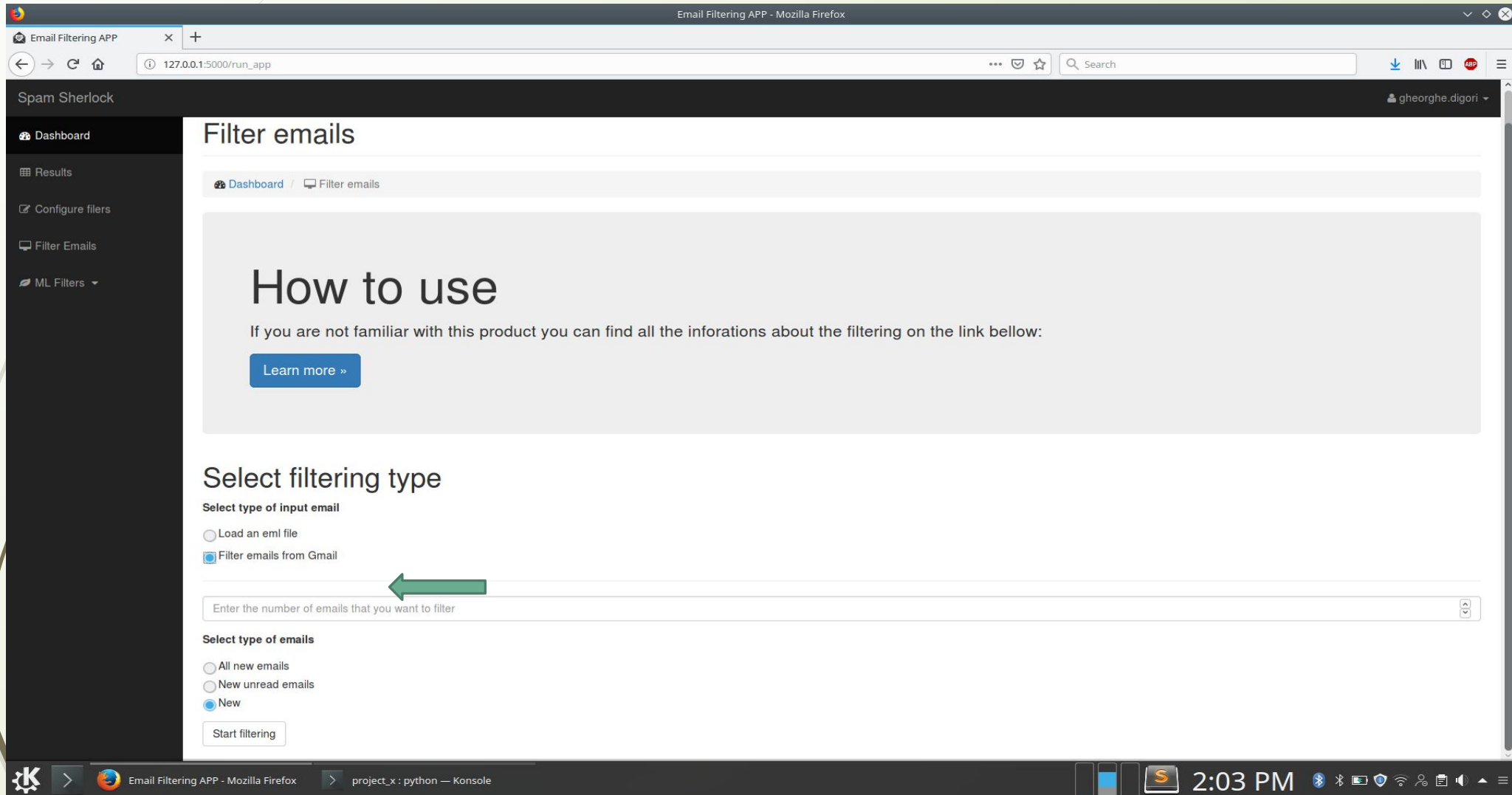
PAGINA “FILTER E-MAILS”

- Această pagină pune la dispoziția utilizatorului 2 opțiuni de încărcare de date (e-mailuri) și anume încărcarea dintru-un folder local și extragerea automată de e-mailuri pe baza contului de user din GMAIL printr-o sesiune de IMAP.
- User-ul poate specifica numărul de email-uri care se doresc a fi extrase precum și tipul acestor email-uri (NEW(necitite încă), ALL(toate), etc.)

PAGINA DE FILTER E-MAIL - LOCAL



PAGINA “FILTER EMAILS”, EMAIL-URI DIN GMAIL





PAGINA DE REZULTATE

- În urma filtrării e-mailurilor rezultatele sunt stocate într-o tabelă persistentă.
- User-ul poate accesa pagina de rezultate pentru a vedea rezultatele filtrării e-mailurilor sale sub o forma tabelară.
- De asemenea, acesta poate șterge înregistrări direct din tabelă având la dispoziție un buton de delete și un buton de “delete all” cu ajutorul căruia poate șterge toate înregistrările din tabela sa de rezultate.

REZULTATE

Email Filtering APP - Mozilla Firefox

Email Filtering APP

127.0.0.1:5000/tables

Spam Sherlock

gheorghe.digori

Dashboard

Results

Configure filters

Filter Emails

ML Filters

Filtering Results

Delete all rows

ID	Client ID	Subject	To	Cc	Bcc	From	Date	Tested_rules	Matched_rules	Class	Filtered_at	Action
2	1	When your fridge or freezer just won't chill out	gelu.digori@gmail.com	None	None	=?UTF-8?B?U2VhcnMgSG9tZSBtZXJ2aWNlcw==?=<shs@em.searshomeservices.com>	Mon, 6 May 2019 13:47:23 -0700	['check_string', 'check_social_media', 'check_html_script', 'check_redirect_options', 'check_blacklisted_sender', 'check_whitelisted_sender', 'check_blacklisted_domains', 'check_whitelisted_domains', 'check_blacklisted_ip_addresses', 'check_whitelisted_ip_addresses', 'check_blacklisted_extensions', 'check_whitelisted_extensions', 'check_blacklisted_sender_subject', 'check_whitelisted_sender_subject', 'check_blacklisted_urls', 'check_whitelisted_urls', 'check_blacklisted_html_url', 'check_whitelisted_html_url']	[u'check_redirect_options']	social-media [False]	2019-05-07 09:23:24.973961	Delete
5	1	Data Engineer (Python) at Bitdefender and 9 other jobs for you.	Digori Gheorghe <gelu.digori@gmail.com>	None	None	LinkedIn <jobs-listings@linkedin.com>	Tue, 7 May 2019 03:33:39 +0000 (UTC)	['check_string', 'check_social_media', 'check_html_script', 'check_redirect_options', 'check_blacklisted_sender', 'check_whitelisted_sender', 'check_blacklisted_domains', 'check_whitelisted_domains', 'check_blacklisted_ip_addresses', 'check_whitelisted_ip_addresses', 'check_blacklisted_extensions', 'check_whitelisted_extensions', 'check_blacklisted_sender_subject', 'check_whitelisted_sender_subject', 'check_blacklisted_urls', 'check_whitelisted_urls', 'check_blacklisted_html_url', 'check_whitelisted_html_url']	['check_social_media', 'check_blacklisted_domains', 'check_whitelisted_sender_subject']	social-media [linkedin]	2019-05-07 09:23:24.973961	Delete

Email Filtering APP - Mozilla Firefox

project_x: python — Konsole

9:24 AM

Exemplu filtrare e-mailuri

Email filters

[Dashboard](#) / [Email filters](#)

Select filters

- ☒ secured_actions
- ☒ blacklisted_sender
- ☒ whitelisted_sender
- ☒ blacklisted_sender_domain
- ☒ whitelisted_sender_domain
- ☒ blacklisted_ip_addresses
- ☒ whitelisted_ip_addresses
- ☐ blacklisted_extensions
- ☐ whitelisted_extensions
- ☐ blacklisted_sender_subject
- ☐ whitelisted_sender_subject
- ☐ blacklisted_URLS
- ☐ whitelisted_URLS
- ☐ blacklisted_html_url
- ☐ whitelisted_html_url

Enter objects with comma

Enter secured actions

pay click \$\$\$ shop order money free million spam super win winner sale fantastic cheap won last

Enter blacklisted senders (names)

info contact root no-reply webmaster

Enter whitelisted senders (names)

Gheorghe Gelu Mihai Alexandru Vasile Andreea Ioana Maria Gelu Anca Alexandru

Enter blacklisted domains

uber.com value.sears.com e2.emag.ro elephant.ro

Enter whitelisted domains

gmail.com facebook.com linkedin.com

Enter blacklisted ip addresses

52.2.92.94 166.78.71.40 168.245.71.148 167.89.42.142 12.130.139.38

Enter whitelisted ip addresses

125.115.130.57 115.207.38.84 54.240.6.1 89.34.107.93 64.186.170.56

Submit Rules

Reset Rules

Filter emails

 [Dashboard](#) /  Filter emails

How to use

If you are not familiar with this product you can find all the information about the filtering on the link bellow:

[Learn more »](#)

Select filtering type

Select type of input email

- ☐ Load an eml file
- ☒ Filter emails from Gmail

5

Select type of emails

- ☐ All new emails
- ☒ New unread emails
- ☐ New

[Start filtering](#)

ID	Client ID	Subject	To	Cc	Bcc	From	Date	Tested_rules	Matched_rules	Class	Filtered_at	Action
26	1	Vlad Nicolae sent you a new message	Digori Gheorghe <gelu.digori@gmail.com>	None	None	Vlad Nicolae Dumitrescu via LinkedIn <messaging-digest-noreply@linkedin.com>	Mon, 1 Jul 2019 16:33:45 +0000 (UTC)	['check_string', 'check_social_media', 'check_html_script', 'check_redirect_options', 'check_blacklisted_sender', 'check_whitelisted_sender', 'check_blacklisted_domains', 'check_whitelisted_domains', 'check_blacklisted_ip_addresses', 'check_whitelisted_ip_addresses', 'check_blacklisted_extensions', 'check_whitelisted_extensions', 'check_blacklisted_sender_subject', 'check_whitelisted_sender_subject', 'check_blacklisted_urls', 'check_whitelisted_urls', 'check_blacklisted_html_url', 'check_whitelisted_html_url']	['check_social_media', 'check_whitelisted_domains', 'check_whitelisted_urls', 'check_whitelisted_html_url']	social-media [linkedin]	2019-07-02 00:14:01.285335	Delete
27	1	Celebrate Canada Day with drone videos 🇨🇦 🚁 WATCH TOP DRONE VIDEOS	<gelu.digori@gmail.com>	None	None	AirVuz - Daily Dose of Drones <Info@AirVuz.com>	Mon, 1 Jul 2019 18:15:44 +0000	['check_string', 'check_social_media', 'check_html_script', 'check_redirect_options', 'check_blacklisted_sender', 'check_whitelisted_sender', 'check_blacklisted_domains', 'check_whitelisted_domains', 'check_blacklisted_ip_addresses', 'check_whitelisted_ip_addresses', 'check_blacklisted_extensions', 'check_whitelisted_extensions', 'check_blacklisted_sender_subject', 'check_whitelisted_sender_subject', 'check_blacklisted_urls', 'check_whitelisted_urls', 'check_blacklisted_html_url', 'check_whitelisted_html_url']	[]	social-media [False]	2019-07-02 00:14:01.285335	Delete
28	1	Fii pe fază! În curând încep Zilele elefant.ro! 🐘	gelu.digori@gmail.com	None	None	"elefant.ro" <contact@elefant.ro>	Mon, 01 Jul 2019 16:08:15 +0000	['check_string', 'check_social_media', 'check_html_script', 'check_redirect_options', 'check_blacklisted_sender',	[u'check_redirect_options', 'check_blacklisted_domains', 'check_blacklisted_urls', 'check_blacklisted_html_url']	social-media [False]	2019-07-02 00:14:01.285335	Delete

CLASIFICATORUL NAÏVE BAYES

- În cadrul acestei aplicații am antrenat și un model de predicție ce are la bază învățarea automată folosind clasificatorul Naive Bayes Multinomial.
- Modelul a fost antrenat pe un set de 72 mii e-mailuri(spam și ham) în limba engleză.
- Modelul antrenat cu NAÏVE BAYES are o precizie de 96,9%.
- Utilizatorul are la dispoziție 2 metode de interacțiune cu acest model:
 1. Inserare e-mail în casuța de text
 2. Încărcare fișier dintr-un folder local.

PAGINA DE PREDICȚIE FOLOSIND NAÏVE BAYES – EXEMPLU HAM

[Dashboard](#) / [Forms](#)

Spam-Ham Email Classifier

Email Body:

Your Email Content

Or choose an email from your computer

Browse...

No file selected.

Classify

"Hello Andrew, We should have a meeting at 17 pm. See you!"

This Email looks like a ['ham']

Exemplu spam

Naive Bayes predictions

[Dashboard](#) / [Naive Bayes Email Classifier](#)

Spam-Ham Email Classifier

Email Body:

Your Email Content

Or choose an email from your computer

Browse...

No file selected.

Classify

"Subject: your prescription is ready . . oxwq s f e low cost prescription medications soma , ultram , adipex , vicodin many more prescribed online and shipped overnight to your door !! one of our us licensed physicians will write an fda approved prescription for you and ship your order overnight via a us licensed pharmacy direct to your doorstep fast and secure !! click here ! no thanks , please take me off your list ogrg z lqlokeolnq lnu"

This Email looks like a ['spam']



PAGINA DE PREZICERE CU UN MODEL DE TENSORFLOW

- Am implementat și o altă opțiune de prezicere folosind un model TENSORFLOW bazată pe antrenarea unei rețele neuronale.
- Acest model folosește algoritmul BAG OF WORDS și prezintă o predicție Gaussiană pe baza unei regresii logistice.
- Modelul a fost învățat cu un dataset public de 5600 de e-mailuri.

Învățare model Tensorflow



Exemplu predicție Spam – Clasificatorul Tensorflow

Tensorflow Classifier

[Dashboard](#) / [Tensorflow Classifier](#)

Spam-Ham Tensorflow Email Classifier

Email Body:

Your Email Content

Or choose an email from your computer

Browse...

No file selected.

Classify

"thats the way u feel. Thats the way its got a b"

This Email looks like a spam



THANK YOU!

