|  |
| --- |
| Photo of building columns from a perspective looking up, with blue sky above |

[Address Book]

[Notes]

[krishna.sadula@gmail.com]

CONTENTS

[INTRODUCTION 0](#_Toc17737)

[MODULES/FILTERS 0](#_Toc6968)

[ACTIVE DOMAINS 0](#_Toc17433)

[EMAIL SERVICE AVAILABLITY 1](#_Toc5080)

[SIGN-IN CANDIDATES 1](#_Toc9769)

# INTRODUCTION

In brief, from the given list of web URLs (approximately 5million), need to get the contacts which are associated with the known e-mail id which is already subscribed to that mail server.

Need to apply various filters across all the given mail domains. Provided file as input to initial level filter will be transformed to condensed file, again which will be input to next level

# MODULES/FILTERS

* Filter for active domains
* Email Service Availability
* Sign-in filter
* Scraping sign-in pages and finding similarities
* Logging in to URLs with the available/scraped data, then download contacts
* Running scripts

## ACTIVE DOMAINS

The primary input (available resource) is raw domains list. For example:

deloitte.com 13387

webmail.co.za 13355

columbia.edu 13236

philips.com 13083

cfl.rr.com 13032

shell.com 12985

virginia.edu 12968

zonnet.nl 12936

aliyun.com 12920

linkedin.com 12527

mchsi.com 12448

inwind.it 12334

its.jnj.com 12277

club-internet.fr 12201

bu.edu 12071

Short comings in finding active domains:

* Response from each individual URL
* Redirection
* http vs https
* Redirected site with https or https
* Browser/protocol specific security instructions

## EMAIL SERVICE AVAILABLITY

Each active URL may not have e-mail service. If website contains mail service, it might have following prefix to the login page. These are observations from various URLs, this list is obviously to grow further.

* [https://www.mail.xxx.com](https://www.mail.com/)
* [http://www.mail.xxx.com](http://www.mail.xxx.com/)
* [http://www.webmail.xxx.com](http://www.webmail.xxx.com/)
* [https://webmail.xxx.com](https://webmail.xxx.com/)
* [http://webmail.xxx.com](http://webmail.xxx.com/)
* mail.xxx.com
* webmail.xxx.com
* connect.xxx.com
* Smtp.xxx.com
* Outbound.xxx.com
* Outgoing.xxx.com

.com might be replaced with .it/.edu/.ac.in/ country specific domain/ etc,…

## SIGN-IN CANDIDATES

URLs may offer e-mail service, but all of such URLs will not be candidates.

Characteristics of suitable candidates:

* Contact list
* No Captcha
* No Locality requirements

## SCRAPING SIGN-IN PAGES

Need to find common fields among the candidate URLs