Dan's Web-crawler

The most fun web crawler on the web

A simple web crawler mainly targets fetching email addresses

Features

- Fetch mapping email by URL.
- Getting a list of emails.

Teach requirements

Support Python 3.X

Installation

```
pip install -r requirements.txt
```

Usage

You use the script by providing seed URL(s) and the number of threads you wish to use.

The number of active threads subjected to your system capabilities and bandwidth connection.

Examples

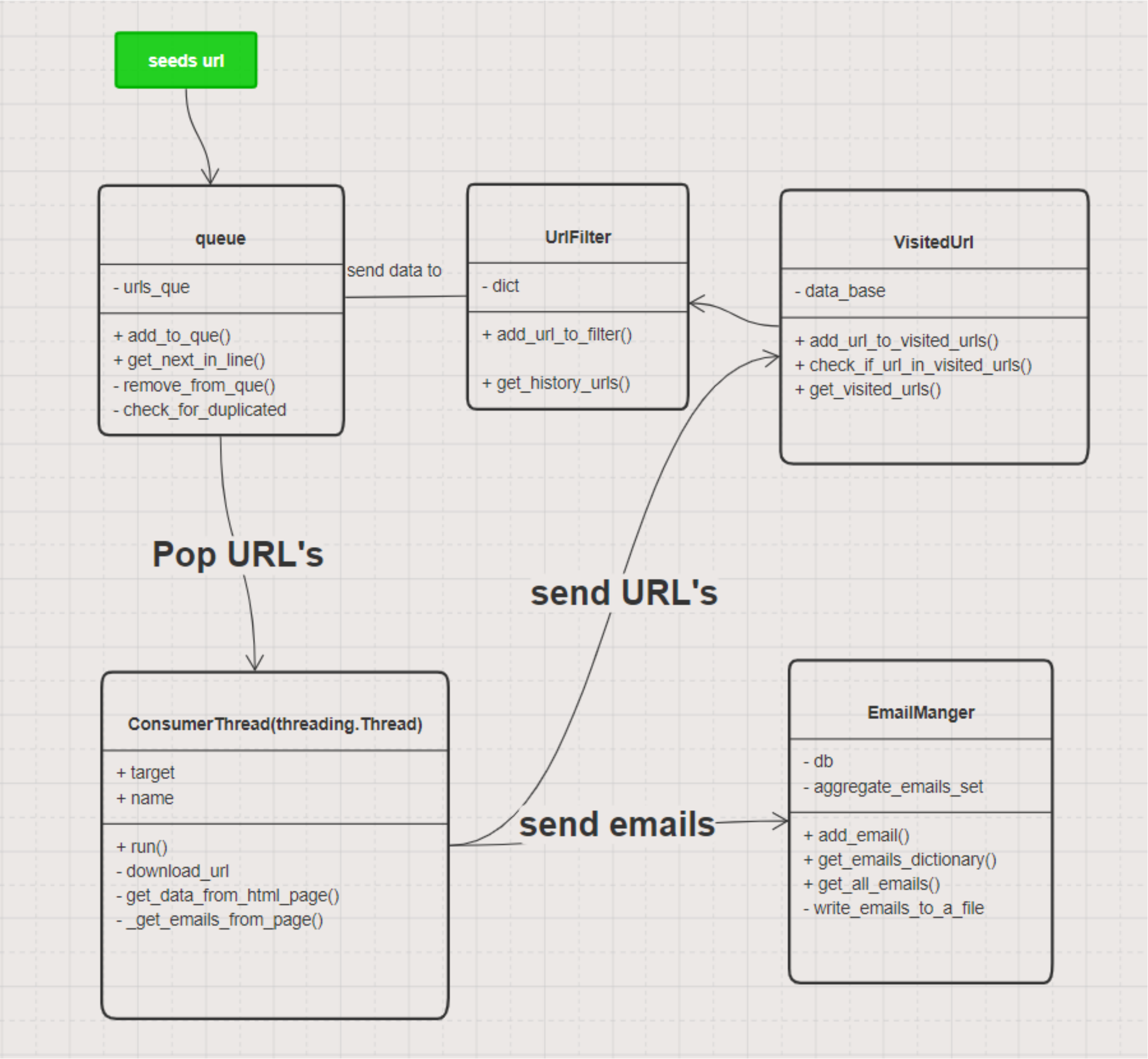
After setting the threads number, duration time and providing an example url .

```
SEEDS_URL = ['https://www.imdb.com/',
'https://www.tel-aviv.gov.il/Residents/Transportation/Pages/Appeal.aspx']
NUM_OF_THREADS = 10
TIME_DURATION_IN_SECOUNDS = 2000
```

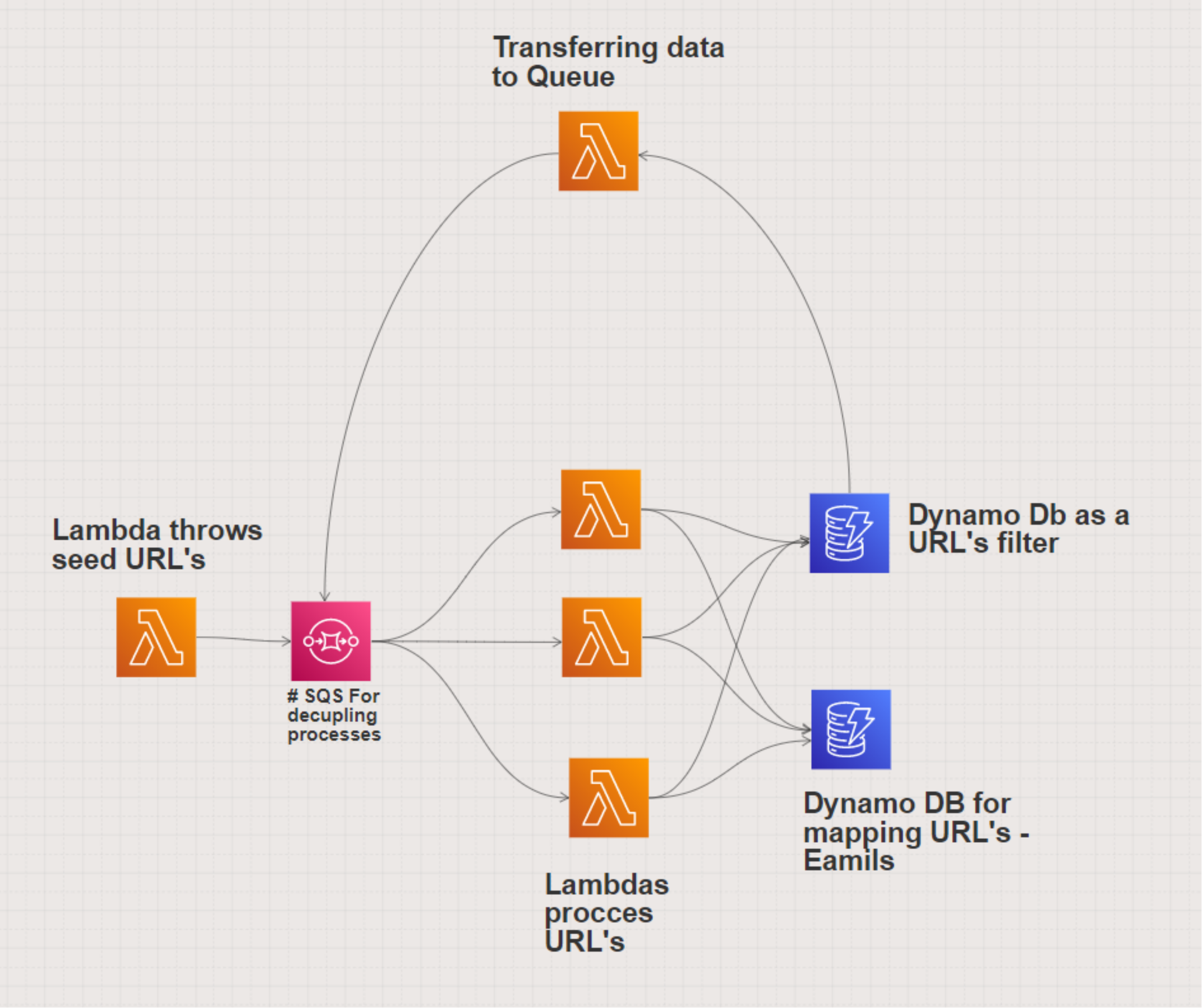
Run the following:

```
python WebCrawler.py



Output:

Json file as follows

{
    "leshem.of@gmail.com": 0,
    "periodicalsba@mail.tel-aviv.gov.il": 0,
    "IMeir@JewishLA.org": 0,
    "info@nana-pirsum.co.il": 0,
    "ily@hagalsheli.co.il": 0,
    "liat19931@gmail.com": 0,
    "al_yesodi@tel-aviv.gov.il": 0,
    "compostlv@gmail.com": 0,
    "noa@ayalaronel.com": 0,
    "vaadatzakaut@tel-aviv.gov.il": 0,
    "r-dan@bna.org.il": 0,
    "himelfarb_i@mail.tel-aviv.gov.il": 0,
    "milgat100@gmail.com": 0,
    "Ra_vaada@mail.tel-aviv.gov.il": 0,
    "pninae@moag.gov.il": 0,
    "treiger_s@mail.tel-aviv.gov.il": 0,
    "julian@krembo.org.il": 0,
    "meretz@ezra.org.il": 0,
    "moked@tel-aviv.gov.il": 0,
    "nevo_m@mail.tel-aviv.gov.il": 0,
    "Gordon_a@tel-aviv.gov.il": 0,
    "tlv.h.gov@gmail.com": 0
}
```

UML’s diagram for current app:



Serverless scheme for future Infinite scaling:


Transferring data
to Queue
Lambda throws
seed URL's
# SQS For
decupling
processes
Dynamo Db as a
URL's filter
Dynamo DB for
mapping URL's -
Eamils
Lambdas
procces
URL's