EMAIL CRAWLER

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Niilo Production V.2017

MIT-License



**KUVAUS:**

Email Crawlerin tehtävänä on liikkua hämähäkin tavoin nettisivulta toiselle linkkejä pitkin ja kerätä kaikki löytämänsä sähköposti osoitteet. Ohjelma toimii samalla periaatteella kuin isojen hakukoneiden kuten Googlen ja Yahoon käyttämät crawlerit ja indexointi mönkijät.

Ohjelman käyttöön ottaminen vaatii hieman kikkailua.

LISÄINFOA: Lue README.txt

Aloitus osoite on pakko antaa http muodossa ja nettiyhteyden nopeus on suuri rajoittava tekijä. Muuten ohjelma on erittäin vakaa, eikä testauksessa löytynyt bugeja.

Ohjelmaa voisi parantaa muuntamalla se .exe muotoon, mutta siihen vaadittava urakka olisi turha tämän kokoiselle projektille. Myös tulostustiedoston valinta olisi ollut mahdollista, mutta siihenkin vaadittava aika olisi ollut liikaa tälle projektille.

**RATKAISUPERIAATE:**

Ohjelma perustuu kolmesta python tiedostosta ja niiden kavereista. Main.py on päätiedosto, jossa kursitaan kokoon multithreading ja GUI. Ohjelman GUI on toteutettu PyQt nimisellä kirjastolla. Kirjastolle on ominaista jäätyä, jos sen joutuu suorittamaan aikaa vievää funktiota, kuten nettisivujen selaamista. Tästä syystä on ohjelmaan rakennettu multithreading. Pythonin toteutuksen voi kuvitella lineaarisena polkuna ja jos polulle laittaa ison aikaa vievän funktion, ei sen takana olevia asioita voi toteuttaa, ennen kuin se on suoritettu. Siksi on kehitetty multithreading. Sen sijaan, että aikaa vievä funktio toteutettaisiin samalla polulla muun ohjelman kanssa, luodaan sille oma polku, josta vedetään lankoja (signaaleja) alkuperäiseen polkuun. Aikaa vievän funktion selvitettyä asioita lähettää se nämä infot lankoja pitkin alkuperäiseen ohjelmaan, joka tekee tiedoilla jotakin.

Ohjelma ottaa syötteenä linkin ja numeromäärän. Numeromäärä on läpikäytävät sivut ja linkki on aloitus sivu. Ohjelma tulostaa löytämänsä sähköpostiosoitteet Libs\mails.txt tiedostoon.

**OHJELMAN OSAT:**

Sen sijaan, että selittäisin yksittäiset moduulit, selitän mieluummin yksittäiset classit. Classin voi kuvitella moduulina, koska sellainen se tavallaan onkin. En aio selittää sen toiminta periaatetta tarkemmin. .py Tiedostoissa on yleensä vain yksi classi.

**MainWindow**(QtWidgets.QMainWindow, Ui\_MainWindow)

MainWindow on ohjelamn pää class, joka kasaa kaikki osat ja aloittaa suorituksen.

**def** \_\_init\_\_(self)

MainWindowin ensimmäinen funktio, rakentaa QUI rakennelman ja yhdistää nappien signaalit

**def main**(self)

Funktio, jonka submit napin painaminen käynnistää. Aloittaa threadin

**def closeEvent**(self, event)

Ylikirjoittaa eventin, joka tulee, kun painaa rastia ohjelmassa. Luo viestin, jossa kysytään, halutaanko varmasti poistua

**def activate**(self)

Aktivoi submit napin uusiksi, kun thread on valmis

**def fileOpen**(self)

Kutsutaan File Open nappia painamalla. Avaa mails.txt tiedoston OS perus tekstieditorissa

**def clearList**(self)

Tyhjentää ListWidgetin (sen johon linkit printantaan)

**def addLine**(self, string)

Lisää rivin ListWidgetiin. Ottaa argumenttina stringin, joka riviin kirjoitetaan. Tapahtuu ainakun threadistä tulee AddNewLine signaali

**def progress**(self, number)

Päivittää latauspalkin tilan 0-100 numerolla, joka annetaan parametrina number.

Kutsutaan signaalilla threadistä.

**if** \_\_name\_\_ == "\_\_main\_\_"

PyQt oma koodin pätkä, joka varsinaisesti printtaa GUIn. Ei kannata kiinnittää huomiota.

**GetThread**(QtCore.QThread)

Thread class, jonka sisällä threadaus tapahtuu

**def** \_\_init\_\_(self, string, number)

Funktio, joka alustaa threadin. Ottaa vastaan stringin ja intin, jotka alustetaan aloitus linkiksi ja maksimi sivumääräksi

**def** \_\_del\_\_(self)

Suoritetaan, kun thread on valmis. Sulkee threadin turvallisesti.

**def main**(self)

Threadin main osa, joka suorittaa itse osotteiden ja linkkien etsimisen, printtaamisen ja signaalien lähettämisen

**def run**(self)

PyQt Threadin oma funktio, joka suoritetaankun threadia kutsutaan. Aloittaa main osan

**Ui\_MainWindow**(object)

QtDesignerin luoma GUI class. EI SAA MUOKATA. Täältä kutsutaan kaikkia GUI elementtejä.

**Lisää infoa ohjelman käytöstä saa README.txt tiedostosta**