Lab 11A

Webscraping

[Optional]

Author: Ragnar Nohre, modified and moved to English Johannes Schmidt

This is the fourth of four optional labs. The optional labs can give together up to 10 *bonus points*. If all regular labs are passed by the general deadline, the accumulated bonus points will count as a bonus on the exam. The scale is that 10 bonus points correspond to 10% of the exam points. This *bonus rule* is only applicable to the first exam (December), and only if the exam is passed without the bonus points.

In order for this lab to qualify for bonus points, you must present and submit your solution before the first exam.

The 10 bonus points are distributed among the optional labs as follows:
Lab8A - 3 points
Lab9A - 1 point
Lab9B - 3 points
Lab11A - 3 points

In this lab you train your skills on webscraping, sqlite and regular expressions.

13.1 Task

On SVT's web-site you find TV-tables for different dates, for instance

```
https://www.svtplay.se/kanaler?date=2020-11-01
```

Write a python program that investigates the TV-tables of the near future and populates a database with relevant information on all `Rapport` emissions. You can also go for other emissions, frequent ones in November 2020 are:

```
270 Sändningsuppehåll   :)
138 Rapport
 79 Sportnytt
 74 Lokala nyheter
 60 Forum
 54 Tjena chavale
 49 Go'kväll
 46 Babblarna
 42 Kulturnyheterna
 41 Sverige idag
 41 Nyhetstecken
 39 Modern och dottern
 38 Alex och Carros vlogg
 34 Fåret Shaun
 31 UR Samtiden
 30 Pyjamashjältarna
 30 Ett fall för KLURO
 30 Bing
 30 Greta Gris
 30 Vargblod
```

You may decide yourself how the database looks like, i.e., which tables it shall contain. You shall only note that svt sends re-emissions, and you shall not store the same information in different places in the database.
The emission related data you shall retrieve/assemble should contain: `date, name, channel, startTime, subHeading, description`. Some example print out of some emissions of `Rapport` and `Babblarna` could look for instance like this:

```
...


date       : 2020-11-22
name       : Babblarna
channel    : ch-barnkanalen
startTime  : 05:15
subHeading : Avsnitt 4
description: En serie för de allra minsta med figurerna Babba, Bibbi, Bobbo, Dadda, Diddi och Doddo.

date       : 2020-11-22
name       : Babblarna
channel    : ch-barnkanalen
startTime  : 05:20
subHeading : Avsnitt 3
description: En serie för de allra minsta med figurerna Babba, Bibbi, Bobbo, Dadda, Diddi och Doddo.

...

date       : 2020-11-23
name       : Rapport
channel    : ch-svt1
startTime  : 22:25
subHeading :
description: Nyheter från Sverige och världen.

date       : 2020-11-23
name       : Rapport
channel    : ch-svt2
startTime  : 12:00
subHeading :
description: Sveriges största nyhetsprogram.

...
```

13.1.1 Hints

1. Fetch the web-sites with `urllib`, confer

   `https://docs.python.org/3/library/urllib.html`

2. Since the url's last part is a date-string, it should be possible to compute how the url for the upcoming days should look like.

3. Confer lab 10 and 11 for more information on webscraping and sqlite.

4. The information is found in a different format than in the html of lab 10. Investigate yourself and find out.

13.2 Examination

Submit your python program on Canvas and present it to your lab assistant. Be prepared to explain your code and database structure.