2016-06-09
kallesoderlund@gmail.com
bengtson.asa@gmail.com

Project Documentation

Requirement Specification

Explanation of the web crawlers:

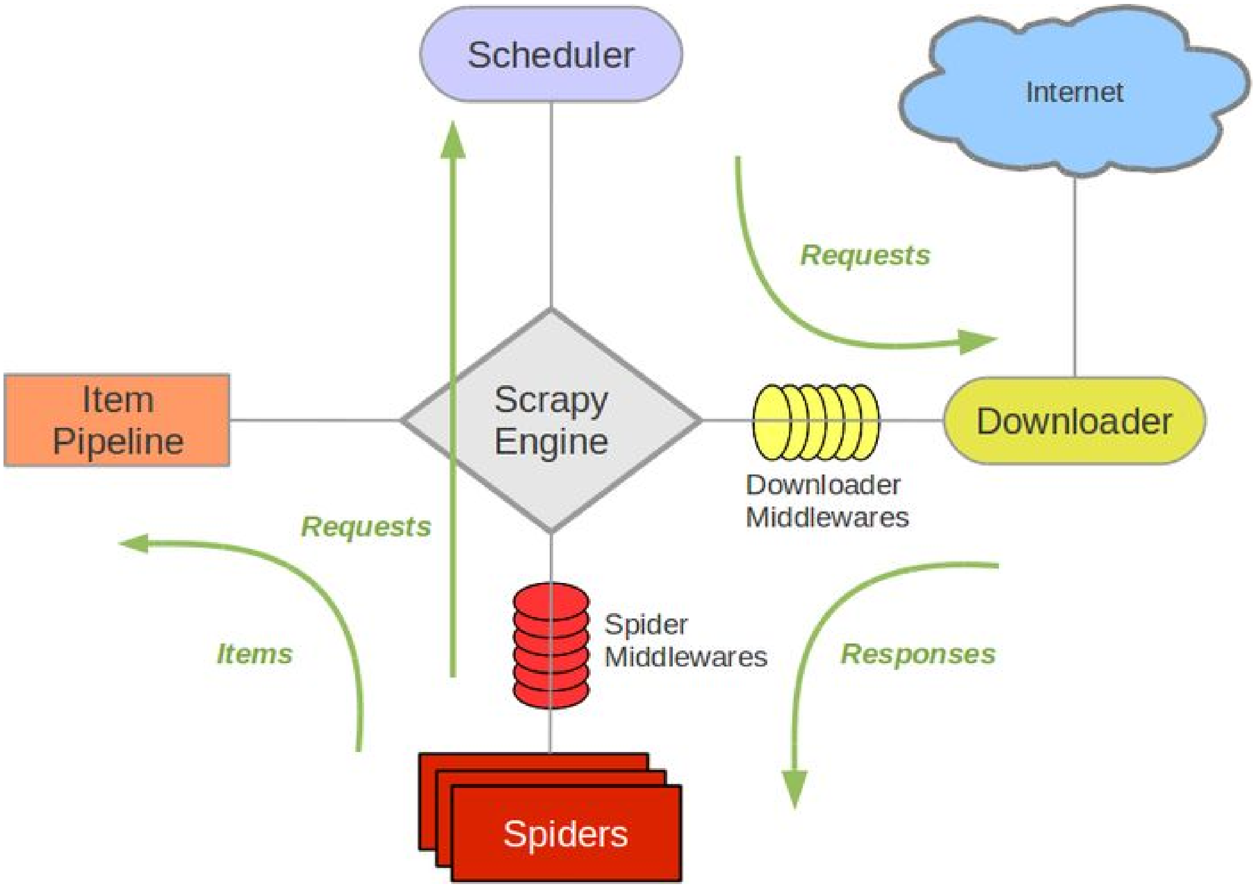
1) When the web crawler is about to collect information, it starts by connecting to the specific URL which it is programmed to visit. From this URL, it collects the source code of the website. The web crawler goes through the source code in order to find specific HTML tags. It can find tags either by the tag's unique name or by its placement in the source code. The information it finds is saved in specific objects, in this case as different events. When all the sought out information is collected and saved as attributes for each event, the sorted information is sent to the so called "pipeline".
2) The pipeline collects the objects with its sorted data. Then, it is being processed some more in order to give all attributes a generic form, for example the same date format. Also, the information is going through a process in which every event is given some other attributes which are not available to extract from the websites, for example "type of event" and "keywords".
3) When all the data has been gone through, and is in the desired format, the object is sent to a database. When an object is saved in the database the web crawler starts over with the next object. Objects which are saved in the database will later be shown in an event calendar on a local website.

How to run the spiders

- "runSpiders.py" is a Python script which run all the web crawlers at the same time. This script contains all eight web crawlers. In order to run this script, the user must type "python runSpiders.py" in the right path in the command prompt. This program can be scheduled to run automatically through Windows' "Task Manager".

2016-06-09
kallesoderlund@gmail.com
bengtson.asa@gmail.com

Architecture and Design Documentation



Scrapy

Scrapy is the web crawling software. It is an open source platform where users can configure it to extract information from various online sources. Scrapy is based on Python and requires Python 2.7 to run properly.

The program consists of a set of folders sorted into a certain structure. The project folder, in this case "afevent", contains several files and an additional folder. In "afevent" there is a few files which are necessary for the program to run properly. "Items.py" is a file where the attributes of each event are given. These attributes are what Scrapy will give every item it extracts. If an attribute isn't given a value, it will stay blank.
http://doc.scrapy.org/en/latest/topics/items.html

"Settings.py" is the settings folder. It contains information about which ports Scrapy will use, and what database server it will connect to. The settings are to a large extent default configurations, apart from database names and name of the bot itself. Please refer to the Scrapy documentation for further information.
http://doc.scrapy.org/en/latest/topics/settings.html

"Pipelines.py" is the pipeline of the entire project. Even though a project may consist of several spiders, each of them will be passed through the same pipeline. In this file, additional attributes, such as type of event and keywords, are given to each item.
http://doc.scrapy.org/en/latest/topics/item-pipeline.html

"Keyfords_final.json" and "type.json" are two json files that contains a list each. Keywords_final.py contains keywords which will be searched to give an item a certain keyword when an item passes through the pipeline. The same thing goes for type.json, but for type of event. Both of these files can be altered to append additional keywords for any kind of item.

Finally there is a file called "runSpiders.py". This is a script that will run automatically after being set up. It contains all the spiders of the project, which will run on command when the user runs the file through the command prompt (by writing "python runSpiders.py" when in the project folder).

```
scrapy.cfg
afevent/
    __init__.py
    items.py
    pipelines.py
    Settings.py
    runSpiders.py
    Keywords_final.json
    type.json
    spiders/
        __init__.py
        spider1.py
        spider2.py
        ...
```

2016-06-09
kallesoderlund@gmail.com
bengtson.asa@gmail.com

Spiders

There are several spiders in this project. Each one of the functions similarly, and the following will be explained below. Please use Google Chrome as a web browser to gain access to source code and Xpaths of web sites. Please refer to the Scrapy documentation and code comments for additional information. http://doc.scrapy.org/en/latest/topics/spiders.html

Steps to take in order to get the web crawlers running:

1. Install Python 2.7
    a. Install Scrapy (http://doc.scrapy.org/en/latest/intro/install.html)
    b. Set up runSpiders.py to run automatically in Windows' "Task Manager"
2. Install MongoDB
3. Install Xampp
    a. Set up your local website