Scraping DSL

Project that is using textX and Arrpeggio to define and interpret DSL in order to generate and run Scrapy project for scraping various types of items and storing it in the MongoDB.

The project currently supports scraping top 250 movies from imdb, councils at otvoreniparlament.rs , and all the products from gigatron and winwin website.

# Prerequisites:

**1.** Python 2.7+

**2.** Python modules installed:

Arpeggio 1.9.0

Jinja2 2.10

pip 10.0.1

pydot 1.2.4

pymongo 3.7.1

pyparsing 2.2.0

Scrapy 1.5.0

textX 1.7.1

**3.** MongoDB database installed (with default settings)

# Item to be scrapped:

For each item there is:

1. Definition of product with properties and css selectors for each property under product\_name\_type.rbt file.

2. Configuration file (.py file) for properties that include more complex manipulation of the tag content under product\_name.py file.

# DSL example:

Main DSL query is defined under product\_name\_program.rbt file.

Example (movie\_program.rbt):

find Movie where cast = 'Morgan Freeman'

All the examples are under dsl/language folder. (together with type definition and configuration files)

# How to run:

*\*Code expects all the files regarding the grammar definition, type definition and configuration and the main DSL query to be under dsl/language folder.*

In dsl/generate.py file you define the scraping item:

# DEFINE THE TYPE OF THE SCRAPING ITEM

type = "movie"

main(type, False)

*\*Currently supported types are: movie, poslanik and most of the products on winwin and gigatron (laptop, mobilni\_telefon, frizider, televizor etc)*

Run generate.py and check the output folder. The scrapy is automatically run.

# Project structure:

* Templates are located at dsl/templates folder
* Generator is in the main dsl folder (with .py extension)
* Language files are in the dsl/language folder.
* Generated Scrapy project is in the dsl/output folder.

# License:

MIT