Berk Arıkan
437954
b.arikan@student.uw.edu.pl

1. Introduction

The website I chose to scrape is StackOverflow. It is commonly used website for asking questions related to programming. Every user can ask questions, answer other users' questions, vote them and tag it with different tags related to their questions.

In this project, we are scraping cumulative score, answer and views from tags which are taken from scraping tags out of questions tagged with "web-scraping". Our aim is to provide insight about tags. Further analysis can be done in order to gain more insight as the project proves.

2. Technical Details

Both three scrapers are logically doing the same thing. Questions at Stackoverflow are in *div* section with an ID of *"question-summary-*"* via regex. First, our scrapers find all sections with this id except for Scrapy spider which uses the class of *"s-post-summary js-post-summary"*. This initial phase scrapes tags in questions under "web-scraping" tag and keeps its score in a dictionary for further analysis purposes. Second phase starts with scraping other tags gathered at initial phase. Both phases are doing the same thing except for the first step where scraper gather tags.

3. Output

Scrapers output statistics of tags related to web-scraping. User can see which tag is getting more interaction from StackOverflow users. Also, performance of scrapers are given,too for user to compare performances. Scrapy does performance metrics internally; so no performance were tracked for Scrapy spider in this project and spider's performance can be seen as usual from the output of scrapy command.

Example output from BeautifulSoup4 scraper is below:

```
============
WEB-SCRAPING
Cumulative Score: -6    Cumulative Answers: 5    Cumulative View: 241
============
============
PYTHON
Cumulative Score: -4    Cumulative Answers: 2    Cumulative View: 102
============
============
JSON
Cumulative Score: -6    Cumulative Answers: 9    Cumulative View: 222
============
============
JAVASCRIPT
Cumulative Score: -2    Cumulative Answers: 4    Cumulative View: 142
============
============
R
Cumulative Score: -2    Cumulative Answers: 4    Cumulative View: 204
============
```

Performance of this bs4 scraper as seconds is: 5.8914403

4. Elementary Data Analysis

Scrapers scrape data from recent questions relative to the time when scrapers are run. We can use the example from previous section as a case for this project. We can clearly see JSON tag gets more interaction then others regardless of views in web-scraping tag.

With more in-depth analysis, scraping Stackoverflow can give someone an insight over trends in programming, people's interaction with different technology and the data can be piped into a machine learning algorithm in order to create a bot for recognizing different IT related problems and provide answers for.

5. Participants

The project is developed by the author of this paper, Berk Arıkan. Documents from respective Python modules were read and Stackoverflow questions are used for common programming problems.