Report for Final Project Data Checkpoint

by Antong Zhou

The final project code has been upload to `https://github.com/AT-kiriko/UMSI-507-final-proj/`.

Data Sources Description

We finally decide to crawl two kinds of pages on `scpwiki.com`:

- Pages for a particular SCP object (item).

  Example: `http://www.scpwiki.com/scp-002`

  Data Fields to crawl:

  - Index of this SCP item. e.g., `SCP-002`;
  - Object Class of this SCP item. e.g., `Euclid`;
  - Rate of this SCP article. e.g., `1464`;
  - Number of comments this SCP article has got. e.g., `110`;
  - Some tags for this SCP article. e.g., `alive, euclid, featured, scp, structure, transifiguration`.

- Pages for every thousand SCP objects.

  Example: `http://www.scpwiki.com/scp-series`

  We crawl these pages basically because each SCP article has a title, which however is not contained in the article page but in the series-index page.

  Data Fields to crawl:

  - Titles of SCP articles. e.g., `The ‘‘Living’’ Room`.

Two screenshots of these two kinds of pages have been added to the project repo, with contents we gonna to crawl marked out.

`scpwiki.com` now has more than 5000 SCP items pages so there are 6 series-index pages in total. In priciple the submitted code is able to crawl all these 5000+ pages, but for illustration purpose we only crawled regular ones in the first 200 SCP items (from `SCP-001` to `SCP-200`), by regular we mean pages have the same layouts as the screenshotted sample pages, there are few outliers like `SCP-001` and `SCP-139` omitted in our list.

These pages will be fetched using techniques (basically cached HTTP GET requests) we learned from Homework 6 and Project 2, these part of codes are mainly placed in `fetch_utils.py` as some utility functions.

Database Description

The crawled data will be used to build a relational database containing two tables:

Column Name	Type
Id	Integer (Primary key)
Title	Text
URL	Text
ObjectClass	Text
Comments	Integer
Rating	Integer

Table 1. Layout for table `items`.

Column Name	Type
Id	Integer (Primary key)
ItemId	Integer
TagName	Text

Table 2. Layout for table `itemtags`.

As an illustration, the database built from the forementioned ∼200 data records will also be uploaded.

Interaction and Presentation Plans

We decided to provide a CLI interface similar to the one we implemented in Project 3, where the user is allowed to type in a few keyword, which will be further translate into some SQL query. The query result will be prettified and then output on the screen. There is an optional keyword to indicate whether to show a barplot using libraries like `plot.ly` and `matplotlib`.