Efficient Scientific Research with Scripts

Speaker
Thomas Kühn

Demo Day
17.9.2014
Picture by Nasa (public domain)
Reading

Writing

Organizing

Images from OpenClipart.org (Creative Commons by Steve Lambert)
Reading

Writing

Organizing

Images from OpenClipart.org (Creative Commons by Steve Lambert)
Panruby\(^1\)

- Focus on content rather than ideoms
- Concise ways to structure text
- Direct support for citations, figures, tables
- Transformation to arbitrary formats
- Template engine for (Multi-)markdown
- **One content source many output formats**

1) https://github.com/Eden-06/panruby
Organizing

Common Tasks

• Management of stored papers
  • Search text fragments in stored files
  • Look up BibTex for stored papers (pdfs)
• Conducting a literature survey
  • Look up BibTex for specific Publications from the web
  • Filtering large BibTex files
  • Downloading papers you previously referenced
Organizing

Common Tasks

- Management of stored papers
  - Search text fragments in stored files
  - Look up BibTex for stored papers (pdfs)
- Conducting a literature survey
  - Look up BibTex for specific Publications from the web
  - Filtering large BibTex files
  - Downloading papers you previously referenced
Organizing Common Tasks

- Management of stored papers
  - Search text fragments in stored files
  - Look up BibTex for stored papers (pdfs)

- Conducting a literature survey
  - Look up BibTex for specific Publications from the web
  - Filtering large BibTex files
  - Downloading papers you previously referenced
A Small Survey

- Q1: *What tools you use to organize your bibliography?*
- Q2: *What tools you use to organize stored papers?*
- 9 Answers named 5 different Tools

### Tools Named on Q1
- **JabRef**: 4 (33%)
- **Mendeley**: 3 (25%)
- **None**: 2 (17%)
- **Text Editor**: 2 (17%)
- **BibTool**: 1 (8%)

### Tools Named on Q2
- **File Browser**: 6 (50%)
- **Mendeley**: 4 (33%)
- **None**: 2 (17%)
Survey Results

- Basical only 4 Tools in Use
- Only few participants use special tools (i.e. Mendeley, JabRef)
- Most rely on the File Browser to manage papers

File Browser

- (Some) Support for search text fragments in stored pdf files

Mendeley

- Manages Papers and Reference
- Fully-searchable library
- Fetches BibTex entries for stored papers automatically

Which other tasks can be automated?

2) http://www.mendeley.com/
Automating the Organization Process

Picture by Nasa (public domain)
Management of Stored Papers
• Automated lookup of BibTex for stored papers

Conducting a Literature Survey
• Automated lookup specific publications from the web
• Automated filtering large BibTex files
• Automated downloading of referenced papers
A leather belt with a solid steel buckle and numerous metal compartments placed along its length, the utility belt is one of the foremost items in the Batman’s arsenal against crime. It must be noted that the Caped Crusader is continuously updating and improving the contents of the utility belt, so the accompanying illustration shows merely an average sampling of its contents.
What We Need?

- **getbibtex.rb**
  Fetches bibtex entries for stored papers
- **gsresearch.rb**
  Collects bibtex entries from Google Scholar
- **bibfilter.rb**
  Filters large BibTex files by various criteria
- **gsdownload.rb**
  Downloads all files referenced by a BibTex files
Never use these scripts in jurisdictions, which prohibit automated use of Google Scholar
- See Google's terms of Use
- Do not use these scripts to attack google services
- These tools are only for research purpose
- „I would pay for using a Google Scholar API“
Automated Management
• Find naming schema for stored publication
  <Full Name of First Author>_<Full Title>.pdf
  (e.g.: Charles W Bachman_Data Structure Diagrams.pdf)
• Keep all documents in one folder (e.g.: library/)
• Use author's last name for subfolder (e.g.: library/Bachman/)

Steps
1. Automated sorting of new files into subfolders
   $ ./mvtodir.sh
2. Generating the file list for getbibtex
   $ ./gettitles.sh > titles.txt
3. Initializing / Updating the bibliography
   $ ruby getbibtex.rb titles.txt bibliography.bib
      1>> bibliography.bib
Task

- Fetch all publications matching a query string
  
  **With:** ospp, workflow
  
  **Exact:** sebastian richly

- Sort out irrelevant publications
- Download pdf files for all relevant publications
- Collect statistics about survey process

Workflow

Fetch → Filter → Download
Task

- Fetch all publications matching a query string
  \textbf{With}: ospp, workflow
  \textbf{Exact}: sebastian richly
- Sort out irrelevant publications
- Download pdf files for all relevant publications
- Collect statistics about survey process

Workflow
**Task**

- Fetch all publications matching a query string
  
  **With:** ospp, workflow
  **Exact:** sebastian richly

- Sort out irrelevant publications
- Download pdf files for all relevant publications
- Collect statistics about survey process

**Workflow**

- **Fetch**
  - gsresearch
- **Filter**
  - raw/
  - autofilter
- **Download**
**Task**

- Fetch all publications matching a query string
  
  *With*: ospp, workflow
  
  *Exact*: sebastian richly

- Sort out irrelevant publications
- Download pdf files for all relevant publications
- Collect statistics about survey process

**Workflow**

1. Fetch
   - gsresearch
   - raw/

2. Filter
   - autofilter
   - filter/

3. Download
   - gsdownload
   - downloads/
Automatic Querying

- Defining a search query
  - Exact, With, Any, and Without
  - Time span (from year to year)
- Directly supported by gsresearch

Steps
1. Test your query with Google Scholar\(^3\) (advanced search)
2. Change the gsresearch.sh accordingly
3. Run the script with
   \[ $ ./gsresearch.sh \]
4. Be patient, very patient

\(^3\) https://scholar.google.com
Automatic Filtering

- Further filter the initial dataset
- Using bibfilter to select items by
  - document class, publisher, citation count, ...
- Two automatic filtering steps in autofilter
  1. Select items by publisher
     ACM, IEEE, Springer, ScienceDirect
  2. Filter items with low impact
     Citation Count < Log( Age )

Human Filtering

- Check the title of the paper and (abstract, content)
  $ mkdir filter_human
  $ for f in `ls filter_rel/`; do
     ruby bibfilter.rb 'filter_rel/$f'
     > 'filter_human/$f'
  done
**Automatic Download**

- Download final set of relevant
- Access files via the publisher's site
- Support for the big four: *ACM, IEEE, Springer, ScienceDirect*
- Extensible towards other publishers
- Downloaded files are referenced within bibtex items

**Steps**

1. Run the script with
   
   `$ ./gsdownload.sh`

2. Be patient

3. Rerun
   
   `$ ./autofilter.sh$`
Collecting Statistics

- Crucial to explain selection method of survey
- Generated automatically by autofilter
- Stored as csv files in states_ */ folder

Workflow

Fetch

Filter

Download

gsresearch

raw/

autofilter

filter/

gsddownload

downloads/
Collecting Statistics
• Crucial to explain selection method of survey
• Generated automatically by **autofilter**
• Stored as csv files in **states_*/** folder

**Workflow**

```
Fetch          Filter          Download
gsresearch     raw/autofilter  gsdownload
downloads/autofilter/stats/
```
Example

- Query for publications from 2008 to 2014
  
  With: ospp, workflow
  
  Exact: sebastian richly

- Initial dataset: 9 entries
- Automatic Filter: 4 entries
Example

- Query for publications from 2008 to 2014
  - *With*: ospp, workflow
  - *Exact*: sebastian richly
- Initial dataset: 9 entries
- Automatic Filter: 4 entries
- Human Filter: 3 entries
- Download: 3 pdf files
A leather belt with a solid steel buckle and numerous metal compartments placed along its length, the utility belt is one of the foremost items in the Batman’s arsenal against crime.

It must be noted that the Caped Crusader is continuously updating and improving the contents of the utility belt, so the accompanying illustration shows merely an average sampling of its contents.

**TWO-WAY RADIO (CONCEALED IN BELT BUCKLE)**

**BATARANG FITS INTO COMPARTMENT BEHIND BELT**

**BATLINE IS DRAWN OUT OF BELT LINING FROM SPRING-LOADED REEL**

---

**BATMAN'S UTILITY BELT**

---

**PRODUCTS:**
- Infrared Flashlight
- Smoke Capsules
- Fingerprint Dusting Kit
- Miniature Camera
- Lockpick Tools
- Tear Gas Pellets
- Microprocessor
- Power Source
- Microcassette Recorder
- Batline Reel
- Laser Torch
- Plastic Explosive Grenades
- Breathing Apparatus
GitHub
- **bibfilter** (https://github.com/Eden-06/bibfilter) contains the `bibfilter.rb` script as independent tool
- **gsresearch** (https://github.com/Eden-06/gsresearch) contains the various Ruby scripts
  - `getbibtex.rb`,
  - `gsresearch.rb`, and
  - `gsdownload.rb`
Additionally, all the presented shell scripts
Automated Tasks

- Automated BibTex lookup for stored papers
- Automated BibTex lookup for specific Publications from the web
- Automated filtering of large BibTex files
- Automated download of papers referenced by a BibTex file
- Semi-automatic literature survey
Writing Papers

“Ghost in the Shell” by Production I.G ALL RIGHTS RESERVED
Writing Papers

Now Automated
- Overview on Paper generators
  - SCIgen
  - Mathgen
  - ...
- Automating idea generation
- Random topic generator
- Random canvas generator
- Predefined Structure

3) http://pdos.csail.mit.edu/scigen/
4) http://thatsmathematics.com/mathgen/

“Ghost in the Shell” by Production I.G ALL RIGHTS RESERVED
Efficient Scientific Research with MultiMarkdown

Automation (By Randell Munroe from xkcd.com)