xml miner Documentation

Release 0.0.5

Chao Li

Oct 24, 2019

Contents:

1 XML/TRXML Selector **1**
- 1.1 Description . . . 1
- 1.2 Status . . . 1
- 1.3 Requirements . . . 1
- 1.4 Installation . . . 1
- 1.5 Usage . . . 1
- 1.6 Development . . . 3
- 1.7 selector and output details: . . . 3

2 Installation **5**
- 2.1 Stable release . . . 5
- 2.2 From sources . . . 5

3 Credits **7**
- 3.1 Development Lead . . . 7
- 3.2 Contributors . . . 7

4 0.0.5 (2019-10-14) **9**

5 0.0.4 (2019-09-11) **11**

6 0.0.3 (2019-08-11) **13**

7 0.0.2 (2019-08-09) **15**

8 0.0.1 (2019-08-09) **17**

9 0.0.0 (2019-08-06) **19**

10 Indices and tables **21**

CHAPTER 1

XML/TRXML Selector

1.1 Description

This package provides two scripts: `mine-xml` and `mine-trxml`.

`mine-xml` selects tags from xml/mxml files, and save the selected values to file.

`mine-trxml` selects fields from trxml/mtrxml files, and save the selected values to file.

1.2 Status

1.3 Requirements

Python 3.6+

1.4 Installation

```
pip install xml-selector
```

1.5 Usage

1.5.1 Use xml selector script

The xml selector supports:

- one or more tagnames:
- selector could be one tagname `name`
- or comma separated tagnames `langskill,compskill,softskills`
- multiple sources:
- e.g. select from xml dir, xml files, mxml file, or directly from annotation server

examples:

```
#select from xml directory
mine-xml --source tests/xmls/ --selector name --output_file name.tsv
mine-xml --source tests/xmls/ --selector langskill,compskill,softskill --output_file␣
↪skill.tsv --with_field_name

#select from xml file or mxml file
mine-xml --source tests/sample.mxml --selector experience --output_file experience.tsv

#select directly from annotation server
mine-xml --source localhost:50249 --selector name --output_file name.tsv --query "set␣
↪Data2018"
```

1.5.2 Use trxml selector script

The trxml selector supports:

- one or more selectors:
- selector can be one field: `name.0.name`
- or comma separated fields: `name.0.name,address.0.address`
- single or multi item:
- can select field from one item, e.g. `experienceitem.3.experience`
- or select field value of all item, e.g. `experienceitem.experience` (or `experienceitem.*.experience`)
- multiple sources:
- e.g. select from trxml dir, trxml files, or mtrxml file

examples:

```
# one selector, single item
mine-trxml --source tests/trxmls/ --selector name.0.name --output_file name.tsv

# one selector, multiple item
mine-trxml --source tests/sample.mxml --selector experienceitem.experience --output_
↪file experience.tsv

# more selectors, single item
```

(continues on next page)

(continued from previous page)

```
mine-trxml --source tests/trxmls/ --selector name.0.name,address.0.address,phone.0.
↪phone --output_file personal.tsv

# more selectors, multiple item
mine-trxml --source tests/sample.mxml  --itemgroup experienceitem --fields experience,
↪experiencedate --output_file experience.tsv
mine-trxml --source tests/sample.mxml  --selector experienceitem.*.experience,
↪experienceitem.*.experiencedate --output_file experience.tsv
mine-trxml --source tests/sample.mxml  --selector experienceitem.experience,
↪experienceitem.experiencedate --output_file experience.tsv
```

1.6 Development

To install package and its dependencies, run the following from project root directory:

```
python setup.py install
```

To work the code and develop the package, run the following from project root directory:

```
python setup.py develop
```

To run unit tests, execute the following from the project root directory:

```
python setup.py test
```

1.7 selector and output details:

- mine-xml:

  input: documents, selector(s), output

  output:

  - default (parameter `with_field_name` not set): `filename, field_value`

  e.g. select all names with selector `name`

filename	value
xxxx	Chao Li

  - parameter `with_field_name` set: `filename, field_value, field_name`

  e.g. select skills with selector `compskill,langskill,otherskill`

filename	value	field
xxxx	java	compskill
xxxx	dutch	langskill

- mine-trxml
  - input:
  - documents, selector(s), output,

- documents, itemgroup, fields, output
- single selector:
- single item (`name.0.name`): filename field

filename	name.0.name
xxxx	Chao Li

- multi items (`skill.*.skill`): filename item_index field

filename	item_index	field
xxxx	0	java
xxxx	1	dutch

- multiple selectors
- single item: filename, field1, field2 ...

each selector points to a field of a specific item with a digital index, e.g. `name.0.lastname,name.0.firstname,address.0.country`

filename	name.0.lastname	name.0.firstname	address.0.country
xxxx	Li	Chao	China
xxxx	Lee	Richard	USA

- multi items: filename, item_index, field1, field2 ...

each selector points to a field from all items in an itemgroup, e.g. `skill.skill,skill.type,skill.date`

filename	skill	skill	type	date
xxxx	0	java	compskill	2001-2005
xxxx	1	dutch	langskill	2002-

CHAPTER 2

Installation

2.1 Stable release

To install xml-miner, run this command in your terminal:

```
$ pip install xml-miner
```

This is the preferred method to install xml-miner, as it will always install the most recent stable release.

If you don't have pip installed, this Python installation guide can guide you through the process.

2.2 From sources

The sources for xml-miner can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/tilaboy/xml-minder
```

Or download the tarball:

```
$ curl  -OL https://github.com/tilaboy/xml-miner/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```

CHAPTER 3

Credits

3.1 Development Lead

- Chao Li <chaoli.job@google.com>

3.2 Contributors

- Chao Li

CHAPTER 4

0.0.5 (2019-10-14)

- bug fix: ElementTree xpath find will return a None if value is an empty string, restore to empty string

CHAPTER 5

0.0.4 (2019-09-11)

- bug fix: reading always use utf8, and not continue reading if failed on encoding of one document

CHAPTER 6

0.0.3 (2019-08-11)

- expand miner.py module to generate matched phrases per doc

CHAPTER 7

0.0.2 (2019-08-09)

- added support for CI

CHAPTER 8

0.0.1 (2019-08-09)

- make two script: mine-xml and mine-trxml

CHAPTER 9

0.0.0 (2019-08-06)

- Add the first version of the mine_xml and mine_trxml

CHAPTER 10

Indices and tables

- genindex
- modindex
- search