Apache Distill Documentation

Release 1.0

Michelle Beard

Jun 21, 2016

CONTENTS

1 User's Guide **1**
- 1.1 Installing Distill . . . 1
- 1.2 Quickstart . . . 2

2 API Reference **3**
- 2.1 API . . . 3
- 2.2 distill package . . . 3

3 Additional Notes **11**
- 3.1 Distill Changelog . . . 11
- 3.2 License . . . 11

4 Indices and tables **15**

Python Module Index **17**

Index **19**

CHAPTER

ONE

USER'S GUIDE

Entire documentation of Distill lives here. Provides step by step instructions to install and deploy Distill in a development and production environment. Including is a API reference portal and quickstart quide.

1.1 Installing Distill

The first step to getting going is installing Distill. Distill is a python project, so it can be installed like any other python library. Several Operating Systems (Mac OS X, Major Versions of Linux/BSD) have Python pre-installed, so you should just have to run:

```
$ easy_install distill
```

Advanced users can install this in a virtualenv if they wish.

When the package is installed via `easy_install` or `pip` this function will be bound to the `distill` executable in the Python installation's `bin` directory (on Windows - the `Scripts` directory).

1.1.1 Development and Testing

To build the source code and run all unit tests:

```
$ python setup.py develop test
```

To start up a local web server, running on localhost:8090:

```
$ dev
```

1.1.2 Distill Deployment

I will describe a setup with nginx as a web server on Ubuntu. A web server cannot communicate directly with a Flask application such as Distill, that's why gunicorn will be used to act as a medium between the web server and Distill. Gunicorn is like an application web server that will be running behind nginx, and it is WSGI compatible. It can communicate with applications that support WSGI – Flask, Django, etc.

Install requirements:

```
$ sudo apt-get update
$ sudo apt-get install -y python python-pip nginx gunicorn
```

Create a directory to store the project:

```
$ sudo mkdir /home/pubic_html && cd /home/public_html
```

Download the project from the GitHub repository and copy the application to the home/www directory:

```
$ git clone https://github.com/draperlaboratory/distill.git /home/public_html
```

Install Distill's requirements:

```
$ pip install -r requirements.txt
```

Distill has provided an nginx configuration file located in `distill/deploy/nginx.conf`.

Gunicorn will use port 8000 and handle the incoming HTTP requests.

Restart nginx to load the configuration changes:

```
$ sudo /etc/init.d/nginx restart
```

Run gunicorn on port 8000:

```
$ gunicorn --workers 4 --bind unix:distill.sock -m 007 deploy/run_server:app
```

Start a new browser instance and navigate to:

```
$ http://server_address
```

1.1.3 Building Documentation

To build Distill's documentation:

```
$ python setup.py build_sphinx
```

Or:

```
# Inside top-level docs/ directory.
$ make html
```

This should build the documentation in your shell, and output HTML. At then end, it should say something about documents being ready in _build/html. You can now open them in your browser by typing:

```
$ open _build/html/index.html
```

1.2 Quickstart

1.2.1 Usage

Using `curl`:

```
$ curl -XGET 'http://localhost:8090/app/register' -d '{
        "application_name" : "my_app",
        "version" : "0.1",
        "application_description" : "my test app"
}'
```

CHAPTER

TWO

API REFERENCE

2.1 API

2.2 distill package

2.2.1 Subpackages

distill.algorithms package

Subpackages

distill.algorithms.graphs package

Subpackages

distill.algorithms.graphs.tests package

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.algorithms.stats package

Subpackages

distill.algorithms.stats.tests package

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.algorithms.tests package

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

Module contents

distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.deploy package

Submodules

distill.deploy.run_server module

distill: Development Server scripts.

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

`distill.deploy.run_server.`**`dev_server`**`()`

Module contents

distill.models package

Subpackages

distill.models.tests package

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

Submodules

distill.models.stout module

distill: This package contains a flask app RESTful api for distill

This flask app exposes some restful api endpoints for querying User-ALE. Very similar to Lucene syntax for basic query operations.

class distill.models.stout.**Stout**
Bases: object

static ingest()

class distill.models.stout.**StoutDoc**(*meta=None*, ***kwargs*)
Bases: elasticsearch_dsl.document.DocType

get_model_obj()

save(**args*, ***kwargs*)

classmethod sync(*stout*)

distill.models.stout.**parse**()

distill.models.userale module

class distill.models.userale.**UserAle**
Bases: object

Generic class supporting basic CRUD operations

static create(*app*)

static delete(*app*)

static denoise(*app*, *doc_type='parsed'*, *save=False*)

static getApps()
Fetch all the registered applications for an Elasticsearch instance.

Note: Privated indexes starting with a period are not included in the result set.

Returns A

Return type dict

static getStatus()
Fetch the status of an Elasticsearch instance.

Returns True/False if connection to Elasticsearch instance has been established.

Return type bool

static read(*app*)

static select(*app*, *app_type=None*, *params=None*)

static update(*app*)

class distill.models.userale.**UserAleParsedDoc**(*meta=None*, ***kwargs*)
Bases: elasticsearch_dsl.document.DocType

save(***kwargs*)

distill.models.userale.**get_all_fields**(*app*, *app_type=None*)

distill.models.userale.**get_cluster_status**(*app*)

distill.models.userale.**merge_dicts**(*lst*)

distill.models.userale.**parse_mappings**(*app*, *app_type=None*)

distill.models.userale.**parse_query_parameters**(*indx*, *app_type=None*, *request_args={}*)

Module contents

distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.tests package

Submodules

distill.tests.basic_test module

distill: Test module.

Meant for use with py.test. Write each test as a function named test_<something>. Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.tests.basic_test.**test_example**()

distill.tests.distill_test module

distill.tests.distill_test.**test_example**()

Module contents

distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Copyright 2016, The Charles Stark Draper Laboratory, Inc. Licensed under Apache Software License

distill.utils package

Subpackages

distill.utils.tests package

Module contents distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

Module contents

distill: tests module.

Meant for use with py.test. Organize tests into files, each named xxx_test.py Read more here: http://pytest.org/

2.2.2 Submodules

2.2.3 Distill RESTful API

distill.app.**create**(*app_id*)
Registers an application in Distill.

```
$ curl -XPOST https://localhost:8000/xdata_v3
```

Parameters **app_id** – Application name

Returns Newly created application's status as JSON blob

distill.app.**delete**(*app_id*)
Deletes an application permentantly from Distill

```
$ curl -XDELETE https://localhost:8000/xdata_v3
```

Parameters **app_id** – Application name

Returns Boolean response message as JSON blob

distill.app.**denoise**(*app_id*)
Bootstrap script to cleanup the raw logs. A document type called "parsed" will be stored with new log created unless specified in the request. Have option to save parsed results back to data store. These parsed logs can be intergrated with STOUT results by running the stout bootstrap script.

```
$ curl -XGET https://localhost:8000/denoise/xdata_v3?save=true&type=parsed
```

Parameters **app_id** – Application name

Returns JSON blob of status

distill.app.**index**()
Show Distill version information, connection status, and all registered applications.

```
$ curl -XGET https://localhost:8000

{
        "author" : "Michelle Beard",
        "email" : "mbeard@draper.com",
        "name": "Distill",
```

```
            "status" : true,
            "version" : "1.0",
            "apps" : {
                    "xdata_v3" : {
                            testing: 205,
                            parsed: 500,
                    },
                    "test_app" : {
                            logs: 500,
                            parsed: 100,
                    }
            }
    }
```

Returns Distill's status information as JSON blob

distill.app.**merge_stout**()
Bootstrap script to aggregate user ale logs to stout master answer table This will save the merged results back to ES instance at new index stout OR denoise data first, then merge with the stout index... If STOUT is enabled, the select method expects a stout index to exist or otherwise it will return an error message.

```
    $ curl -XGET https://locahost:8000/stout/xdata_v3
```

Returns Status message

distill.app.**page_not_found**(*error*)
Generic Error Message

distill.app.**search**(*app_id*, *app_type*)
Search against an application on various fields.

```
    $ curl -XGET https://[hostname]:[port]/app_name/select?q=session_id:A1234&size=100&scroll=false&
```

Parameters

- **app_id** – Application name
- **app_type** – Optional document type to filter against
- **q** – Main search query
- **size** – Maximum number of documents to return in request
- **scroll** – Scroll id if the number of documents exceeds 10,000
- **fl** – List of fields to restrict the result set

Returns JSON blob of result set

distill.app.**stat**(*app_id*, *app_type*)

> **Warning:** Not implemented/available

Generic histogram counts for a single registered application filtered optionally by document type.

```
    $ curl -XGET https://localhost:8000/xdata_v3/testing/?elem=signup&event=click
```

Parameters

- **app_id** – Application name
- **app_type** – Application type

Returns JSON blob of result set

distill.app.**status**(*app_id*)
Presents meta information about an registered application, including field names and document types.

```
$ curl -XGET https://localhost:8000/status/xdata_v3
```

Parameters **app_id** – Application name

Returns Registered applications meta data as JSON blob

distill.app.**update**(*app_id*)
Renames a specific application

```
$ curl -XPOST https://localhost:8000/update/xdata_v3?name="xdata_v4"
```

Parameters **app_id** – Application name

Returns Boolean response message as JSON blob

2.2.4 Distill Exceptions

distill: This package contains a flask app RESTful api for distill

This flask app exposes some restful api endpoints for querying User-ALE. Very similar to Lucene syntax for basic query operations.

exception distill.exceptions.**ValidationError**
Bases: exceptions.Exception

2.2.5 Distill Validation Library

distill: This package contains a flask app RESTful api for distill

This flask app exposes some restful api endpoints for querying User-ALE. Very similar to Lucene syntax for basic query operations.

distill.validation.**str2bool**(*v*)

distill.validation.**validate_request**(*q*)

2.2.6 Module contents

Distill: This package contains a flask app RESTful api for distill

CHAPTER

THREE

ADDITIONAL NOTES

Design notes, legal information and changelong are here.

3.1 Distill Changelog

Here you can see the full list of changes between each Distill release.

3.1.1 Version 1.0

- Initial version.

3.2 License

Distill is licensed under Apache License.

The full license text can be found below (distill-license).

3.2.1 Authors

Distill is written and maintained by Michelle Beard and various contributors:

Development Lead

- Michelle Beard <mbeard@draper.com>

Additional Staff

- Laura Mariano <lmariano@draper.com>
- Dr. Joshua Poore <jpoore@draper.com>
- Clay Gimenez <cgimenez@draper.com>
- Steven York <syork@draper.com>

3.2.2 Distill License

Apache License

Version 2.0, January 2004

http://www.apache.org/licenses/

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

   "License" shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

   "Licensor" shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

   "Legal Entity" shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition, "control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

   "You" (or "Your") shall mean an individual or Legal Entity exercising permissions granted by this License.

   "Source" form shall mean the preferred form for making modifications, including but not limited to software source code, documentation source, and configuration files.

   "Object" form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

   "Work" shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

   "Derivative Works" shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship. For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

   "Contribution" shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, "submitted" means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution."

   "Contributor" shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

   (a) You must give any other recipients of the Work or Derivative Works a copy of this License; and

   (b) You must cause any modified files to carry prominent notices stating that You changed the files; and

   (c) You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and

   (d) If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

   You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

7. Disclaimer of Warranty. Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in accepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

APPENDIX: How to apply the Apache License to your work.

> To apply the Apache License to your work, attach the following boilerplate notice, with the fields enclosed by brackets "{}" replaced with your own identifying information. (Don't include the brackets!) The text should be enclosed in the appropriate comment syntax for the file format. We also recommend that a file or class name and description of purpose be included on the same "printed page" as the copyright notice for easier identification within third-party archives.

CHAPTER

FOUR

INDICES AND TABLES

- genindex
- modindex
- search

d

distill.algorithms, 4
distill.algorithms.graphs, 3
distill.algorithms.graphs.tests, 3
distill.algorithms.stats, 4
distill.algorithms.stats.tests, 4
distill.algorithms.tests, 4
distill.deploy, 4
distill.deploy.run_server, 4
distill.models, 6
distill.models.stout, 5
distill.models.tests, 4
distill.models.userale, 5
distill.tests, 6
distill.tests.basic_test, 6
distill.tests.distill_test, 6
distill.utils, 7
distill.utils.tests, 7

C

create() (distill.models.userale.UserAle static method), 5
create() (in module distill.app), 7

D

delete() (distill.models.userale.UserAle static method), 5
delete() (in module distill.app), 7
denoise() (distill.models.userale.UserAle static method), 5
denoise() (in module distill.app), 7
dev_server() (in module distill.deploy.run_server), 4
distill (module), 9
distill.algorithms (module), 4
distill.algorithms.graphs (module), 3
distill.algorithms.graphs.tests (module), 3
distill.algorithms.stats (module), 4
distill.algorithms.stats.tests (module), 4
distill.algorithms.tests (module), 4
distill.app (module), 7
distill.deploy (module), 4
distill.deploy.run_server (module), 4
distill.exceptions (module), 9
distill.models (module), 6
distill.models.stout (module), 5
distill.models.tests (module), 4
distill.models.userale (module), 5
distill.tests (module), 6
distill.tests.basic_test (module), 6
distill.tests.distill_test (module), 6
distill.utils (module), 7
distill.utils.tests (module), 7
distill.validation (module), 9

G

get_all_fields() (in module distill.models.userale), 6
get_cluster_status() (in module distill.models.userale), 6
get_model_obj() (distill.models.stout.StoutDoc method), 5
getApps() (distill.models.userale.UserAle static method), 5
getStatus() (distill.models.userale.UserAle static method), 5

I

index() (in module distill.app), 7
ingest() (distill.models.stout.Stout static method), 5

M

merge_dicts() (in module distill.models.userale), 6
merge_stout() (in module distill.app), 8

P

page_not_found() (in module distill.app), 8
parse() (in module distill.models.stout), 5
parse_mappings() (in module distill.models.userale), 6
parse_query_parameters() (in module distill.models.userale), 6

R

read() (distill.models.userale.UserAle static method), 5

S

save() (distill.models.stout.StoutDoc method), 5
save() (distill.models.userale.UserAleParsedDoc method), 6
search() (in module distill.app), 8
select() (distill.models.userale.UserAle static method), 5
stat() (in module distill.app), 8
status() (in module distill.app), 9
Stout (class in distill.models.stout), 5
StoutDoc (class in distill.models.stout), 5
str2bool() (in module distill.validation), 9
sync() (distill.models.stout.StoutDoc class method), 5

T

test_example() (in module distill.tests.basic_test), 6
test_example() (in module distill.tests.distill_test), 6

U

update() (distill.models.userale.UserAle static method), 5
update() (in module distill.app), 9
UserAle (class in distill.models.userale), 5
UserAleParsedDoc (class in distill.models.userale), 5

V

validate_request() (in module distill.validation), 9
ValidationError, 9