Company reconciliation Task

# Requirements

* The task runs in an app written in Flask[[1]](#footnote-1) (Python microframework) that runs in an Heroku Dyno[[2]](#footnote-2).
* Workers are recruited from Amazon Mechanical Turk[[3]](#footnote-3).
* The interaction with the mTurk API takes place by IPython Notebooks[[4]](#footnote-4).
* Companies to be resolved has to be stored in a MongoDB instance (this example uses MLab).

Several guides and tutorial are available to use such technologies: <https://stackabuse.com/deploying-a-flask-application-to-heroku/>

<https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/Welcome.html>

## Environment variable

The service requires the following environment variable:

|  |  |
| --- | --- |
| **Name** | **Description** |
| **URL\_MONGO** | Standard MongoDB URI[[5]](#footnote-5) of the DB where companies are stored. |
| **IAM\_USER\_ACCESS\_KEY** | Amazon ITM Access Key |
| **IAM\_USER\_SECRET\_KEY** | Amazon ITM Secret Key |
| **MTURK\_REGION\_NAME** | us-east-1 |
| **URL\_MTURK\_SANDBOX** | https://mturk-requester-sandbox.us-east-1.amazonaws.com |
| **URL\_MTURK\_PRODUCTION** | https://mturk-requester.us-east-1.amazonaws.com |

To know more about IAM have a look at this: <https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html>

## Input Data and Golden Standard

The companies that require to be resolved has to be stored in a database named **companies** Mongo DB instance in this way:

|  |
| --- |
| {  "\_id": {  "$oid": "5e4bfc63df8ef924f306bd06"  },  "ocid": "ocds-0c46vo-0133-011071-2019",  "name": "Groupe hospitalier de La Rochelle",  "opencorporates\_id": "/companies/fr/200047835",  "country\_code": "fr",  "full\_address": "rue du Dr Schweitzer La Rochelle 17019",  "street\_address": "rue du Dr Schweitzer",  "locality": "La Rochelle",  "postcode": "17019",  "country": "France",  "g": "7200047835" } |

The attribute **g** has to be present for each company. Companies that are golden standard (the ones for which we already have the *opencorporate id* and we want to use as indicator of the reliability of the worker) need to have the g attribute starting with the number 7, followed by the *opencorporate id* of the company. For example, the *Groupe hospitalier de La Rochelle* is a golden standard company, since the attribute g begins with 7, and after that the expected *opencorporate id*. All the others, not-golden standard companies, have to have the attribute **g** starting by a number different than 7, followed by a few of random digits.

## Scripts

The folder *ipython* contains two *notebooks, with some examples* to use the *reconciliation service*:

* *HITs-Launcher.ipynb* - Is a script to setup and launch new HITs, based on the list of required companies, the qualification to filter out low-quality workers, and other params, such as, HIT title, description, and amount of payment for resolution.
* *Results Viewer.ipynb* - This script downloads the results of one or more HITS and store them in a Pandas Dataframe[[6]](#footnote-6) for further analysis.

## 

1. <https://palletsprojects.com/p/flask/> [↑](#footnote-ref-1)
2. <https://www.heroku.com/dynos> [↑](#footnote-ref-2)
3. <https://www.mturk.com/> [↑](#footnote-ref-3)
4. <https://ipython.org/> [↑](#footnote-ref-4)
5. <https://docs.mlab.com/connecting/#connect-string> [↑](#footnote-ref-5)
6. <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html> [↑](#footnote-ref-6)