**HireRight Innovation Challenge**

**Intelligent Public Record Parser**

Interterm – Jan 2020

Contents

[Background: 3](#_Toc26354731)

[Challenge: 3](#_Toc26354732)

[Data set: 3](#_Toc26354733)

[Solution: 5](#_Toc26354734)

[Evaluation criteria: 5](#_Toc26354735)

[Training List 5](#_Toc26354736)

[Evaluation List 6](#_Toc26354737)

# Background

HireRight scrapes various public record web sites as part of criminal background screening to capture criminal records of candidates. Most of the sites have the key attributes representing criminal records, but they are often positioned on the website differently and with different vocabulary for the same attribute. This requires HireRight to build custom mapping and scraper for each public record website. For instance, the "Offense Date" could be represented as "offenseDate", "offensedt", "offDate" in different public record websites.

If there is an intelligent way to reliably parse the relevant attributes and validate the contents irrespective of the structure of the website, it will minimize or even eliminate manual parsing.

# Challenge

Build an intelligent parser that learns to extract key criminal record attributes from known set of public record websites such that resulting solution can automatically parse any new public record website without explicit mapping/training.

# Data set

HireRight will provide a list of websites for the teams to generate a training dataset. The dataset typically has the following attributes. Teams could write crawlers to generate the data set by searching for commonly used names. The below table provides the schema for the dataset.

Note: , and e

|  |  |  |
| --- | --- | --- |
| Field | Type | Comments |
| Scraped information | HTML/XML/Screenshot | Output |
| Case.Court Case ID | Text | Output |
| Case.First Name | Text | Output |
| Case.Middle Name | Text | Output |
| Case.Last Name | Text | Output |
| Case.Gender | Text | Output |
| Case.Date of Birth | Date | Output |
| Case.Jurisdiction Name | Text | Output |
| Case.Jurisdiction County | Text | Output |
| Case.Jurisdiction State | Text | Output |
| Charge.Offense Date | Date | Output |
| Charge.Arrest Date | Date | Output |
| Charge.File Date | Date | Output |
| Charge.Disposition Date | Date | Output |
| Charge.Charge Text (Offense Literal) | Text | Output |
| Charge.Charge Status | Text | Output |
| Charge.Disposition Text | Text | Output |
| Charge.Sentencing Notes | Text | Output |
| Charge.Prison Time | Number | Output |
| Charge.Jail Time | Number | Output |
| Charge.Probation Time | Number | Output |
| Charge.Fine(s) | Currency | Output |
|  |  |  |

# Solution

Once the dataset is generated, the teams could build Machine Learning models to extract the attributes defined in the schema, and the teams have an option to train from the HTML extract or screenshots taken from the websites. It is important to note that during the evaluation, a completely different set of public record websites will be used to evaluate the solution.

The solution should be repeatable, and all artifacts used, including source code, should be shared as part of the solution.

Preferred technologies: Git, Python, TensorFlow, Docker.

# Evaluation criteria

HireRight will provide a “Training list” – a list of public record websites to generate a training dataset for the challenge. An “Evaluation list” – a list of public record websites, which is a completely different set of websites will be shared at the time of evaluation for evaluating solutions from the teams. The solution that parses the criminal attributes from “Evaluation list” with no explicit training/mapping with the highest accuracy and consistency will be the winning solution.

# Training List

|  |  |
| --- | --- |
| **Jurisdiction** | **URL** |
| South Carolina | [https://www.sccourts.org/casesearch/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.sccourts.org_casesearch_&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=0I_UiiMW1v5cCHNtfmj2X2vjiQ0qt4ETE26MficHRCk&e=) |
| Wisconsin | [https://wcca.wicourts.gov/case.html](https://urldefense.proofpoint.com/v2/url?u=https-3A__wcca.wicourts.gov_case.html&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=bwkaO-SiBK0RgfX906G4ww1JCls-EymM5Ea2h8znF-U&e=) |
| Oklahoma | [https://www.oscn.net/dockets/search.aspx](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.oscn.net_dockets_search.aspx&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=aLy3FHprdIYEvxG4nSdvPOrukT0mZfg0bo_FaRw69B8&e=) |
| Various FL counties (Civitek) | [https://www.civitekflorida.com/ocrs/county/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.civitekflorida.com_ocrs_county_&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=-MzKTq2Rb63g4UEmYMO32cfBrNUtlx8OkaG2OME12QU&e=) |
| Arizona | [https://apps.supremecourt.az.gov/publicaccess/caselookup.aspx](https://urldefense.proofpoint.com/v2/url?u=https-3A__apps.supremecourt.az.gov_publicaccess_caselookup.aspx&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=t2wdUHLn8S3Jo5HxKubQ3K2zmv7eX_0lEFVlCCRG4ZE&e=) |
| Maricopa County, AZ (Superior court) | [https://www.superiorcourt.maricopa.gov/docket/CriminalCourtCases/caseSearch.asp](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.superiorcourt.maricopa.gov_docket_CriminalCourtCases_caseSearch.asp&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=uKle5BIHxGxuyynNydVrvnf5-_kahlkjTYnW0B7oq-o&e=) |
| Maricopa County, AZ (Justice courts) | [http://justicecourts.maricopa.gov/FindACase/casehistory.aspx](https://urldefense.proofpoint.com/v2/url?u=http-3A__justicecourts.maricopa.gov_FindACase_casehistory.aspx&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=NqwR_qseXEBnqyMmvzhAPk6W7kF-cy417sQ4-MVUJag&e=) |
| El Paso County, TX | [https://casesearch.epcounty.com/PublicAccess/default.aspx](https://urldefense.proofpoint.com/v2/url?u=https-3A__casesearch.epcounty.com_PublicAccess_default.aspx&d=DwMFAg&c=ZkecDCWBx1Gef1Z30rF7hcmiLU830RlabebNYewQ8lc&r=6naRknarsyNCYzLwnRWFyNxtirHpVTLDGuNIY6BUlcM&m=8QzxuOBcTp7PEBDsh22OJ2zo3-C95YczJAmk3SUO-io&s=ERrQuzoLMOe_eNMv1zwSAYRXhEYtgMB5-XK43KeGdKU&e=) |

# Evaluation List

Will be shared at the time of evaluation.