CAST HIGHLIGHT

Keyword Analyzer
GDPR & Beyond

Getting Started Guide

CAST

Few words about GDPR

According to Wikipedia: *“The General Data Protection Regulation (GDPR) is a regulation in EU law on data protection and privacy for all individuals within the European Union and the European Economic Area. It also addresses the export of personal data outside the EU and EEA areas. The GDPR aims primarily to give control to citizens and residents over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU.”*

Concretely speaking, it means that organizations will have to know if their applications read, process or store Personally Identifiable Information (PII) data of EU-based users, in order to set up the appropriate actions (register the application, declare a Data Processor, qualify the nature and purpose of data collection, modify the application behavior to ask users whether they consent to share their data or not, etc.).

Keyword Detection thanks to CAST Highlight

According to the GPDR regulation, organizations now have to know if their applications are processing PII data. This is something quite obvious and easy to determine if an application is connected with a central database which holds tables or columns like “first_name”, “email_address”, “social_security_number”, etc. **But in Software, nothing is obvious anymore**.

First, applications and databases don’t necessarily have a 1:1 ratio. You may have a few central databases that are accessed by hundreds of apps. Then, GPDR is not only about identifying the databases you have and putting them in the GPDR process. This verification also needs to be done at the application level.
Secondly, applications can manipulate PII **data without any database**. As today API, JSON, web and micro services are the norm, meaning that a piece of source code can read, process and share data with other components without having a clue about the database that initially stored it. A small script cooked by your HR department read LinkedIn’s API to hire the best profiles? There is a risk that it manipulates PII data, at least names, location and profile pictures.

Fortunately, developers love code they can easily read and maintain: 99% of the time they call their classes, methods, parameters with names that are not obscure (e.g. getCustomerName, updateProfile($CreditCardNumber) etc.). As a result, it is possible to approximate (if not determine) that an application processes PII data by scanning its source code and counting occurrences of PII-related keywords. Scanning code to search for patterns? **That’s exactly where Highlight comes into the game**.

CAST Highlight Command Line

The Keyword Scan feature works with the command line and takes the path to your keyword configuration file (–keywordScan “path/to/your/file.xml”).

```
java -jar HighlightAutomation.jar --workingDir "C:\XXX\Work\SCAN_RESULTS" --sourceDir "C:\XXX\Source\SOURCE_CODE" --login
“LOGIN" --password “PASSWORD" --applicationId APP_ID --companyId COMPANY_ID --serverUrl "https://rpa.casthighlight.com" --
keywordScan "C:\XXX\KEYWORD_CONFIG.xml "
```

--workingDir "C:\XXX\Work\SCAN_RESULTS“

>> the location that will host the scan analysis result files including: the health and cloudready index assessment, framework discovery, the keyword results as well as the logs.

--sourceDir "C:\XXX\Source\SOURCE_CODE"

>> the root folder that contains the source code of the application to be analyzed

Connection Credentials

>> type the login (email), password and server URL to establish the connection to the solution
>> Company ID & Application ID can be found in the Manage Portfolio \ Manage Applications console from CAST Highlight UI

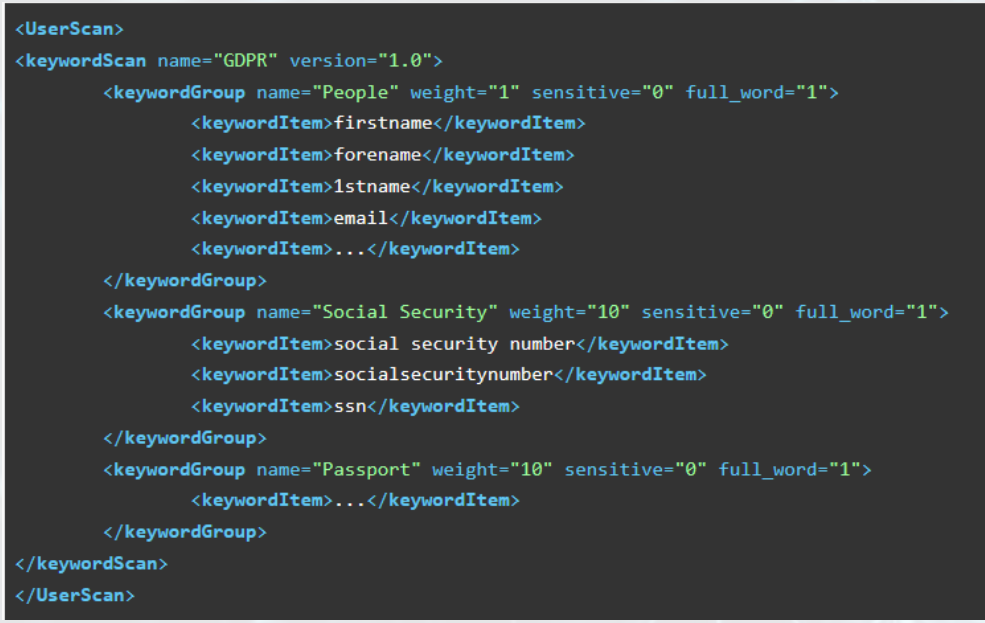
--keywordScan "C:\XXX\KEYWORD_CONFIG.xml

>> define the path where the Keyword Analyzer configuration file is located

Keyword Analyzer Configuration File

This configuration file will tell the analyzers in a structured way what to search during a code scan. Its structure is detailed below:

- **UserScan**: the root node that contains the configuration.
- **keywordScan**: the main node for a keyword topic. You can indicate a name and a version (e.g. name=”GDPR” version=”1.2”). You can have multiple topics in a single configuration file as you may want to search for GDPR-related keywords but also keywords for licenses, specific unauthorized functions, other regulation tags…
- **keywordGroup**: the node that will search in code for a keyword or a set of similar keywords (e.g. “social security number”, “ssn”, “social security nbr”, etc.). For each keyword group, you can define a specific weight (for instance, in a GDPR context, a passport number will weigh more than a firstname) and search options such as case sensitivity or full vs. partial word-matching.
- **keywordItem**: one of the search element. You can have multiple items for a given keyword group.



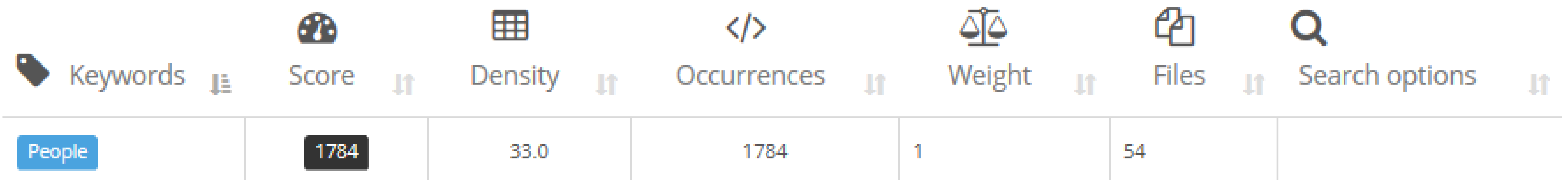
```
<UserScan>
<keywordScan name="GDPR" version="1.0">
        <keywordGroup name="People" weight="1" sensitive="0" full_word="1">
                <keywordItem>firstname</keywordItem>
                <keywordItem>forename</keywordItem>
                <keywordItem>1stname</keywordItem>
                <keywordItem>email</keywordItem>
                <keywordItem>...</keywordItem>
        </keywordGroup>
        <keywordGroup name="Social Security" weight="10" sensitive="0" full_word="1">
                <keywordItem>social security number</keywordItem>
                <keywordItem>socialsecuritynumber</keywordItem>
                <keywordItem>ssn</keywordItem>
        </keywordGroup>
        <keywordGroup name="Passport" weight="10" sensitive="0" full_word="1">
                <keywordItem>...</keywordItem>
        </keywordGroup>
</keywordScan>
</UserScan>
```

Keyword Analyzer behavior

Example #1 – A Keyword Group with multiple items

```
<keywordGroup name="People" weight="1" sensitive="0" full_word="0">
                <keywordItem>firstname</keywordItem>
                <keywordItem>1stname</keywordItem>
                <keywordItem>name</keywordItem>
                <keywordItem>lastname</keywordItem>
                <keywordItem>birthdate</keywordItem>
                <keywordItem>nationality</keywordItem>
                <keywordItem>citizenship</keywordItem>
                <keywordItem>email</keywordItem>
                <keywordItem>e-mail</keywordItem>
</keywordGroup>
```

The Keyword Analyzer looks for all the keyword items and applies a weight of 1 each time an occurrence is found. The configuration is not Case Sensitive nor restrictive regarding the string i.e. the items can be prefixed or suffixed e.g. “name” counts by its own but it also if the engine finds “myname”, “name1st” or “TestnameSuffixe”.



Keywords	Score	Density	Occurrences	Weight	Files	Search options
People	1784	33.0	1784	1	54	

The Keyword Analyzer detected 1784 occurrences amongst 54 different files. Because the weight is set at 1, the total score equals 1784. The Density corresponds to the total score divided by the number of files.

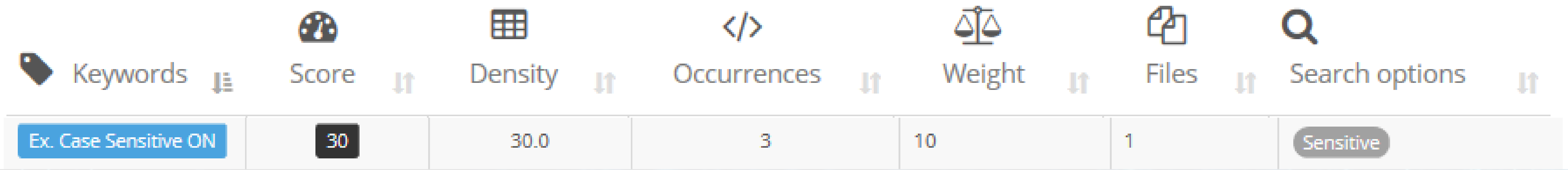
Bear in mind that this score embeds all the keyword items that belong to the keyword group called PEOPLE.

Keyword Analyzer behavior

Example #2 – A Keyword Group using the Case Sensitive Option

```
<keywordGroup name="Ex. Case Sensitive ON" weight="10"
sensitive="1“ full_word="0">
            <keywordItem>myTestCaseSensitive</keywordItem>
            <keywordItem>MYtestCASESensitivE</keywordItem>
            <keywordItem>MYTESTCASESENSITIVE</keywordItem>
</keywordGroup>
```

The Keyword Analyzer looks for all the keyword items and applies a weight of **10** each time an occurrence is found. The configuration is Case Sensitive but not restrictive regarding the string i.e. the items can be prefixed or suffixed - as long as they use the same CASE, they will be counted.



Keywords	Score	Density	Occurrences	Weight	Files	Search options
Ex. Case Sensitive ON	30	30.0	3	10	1	Sensitive

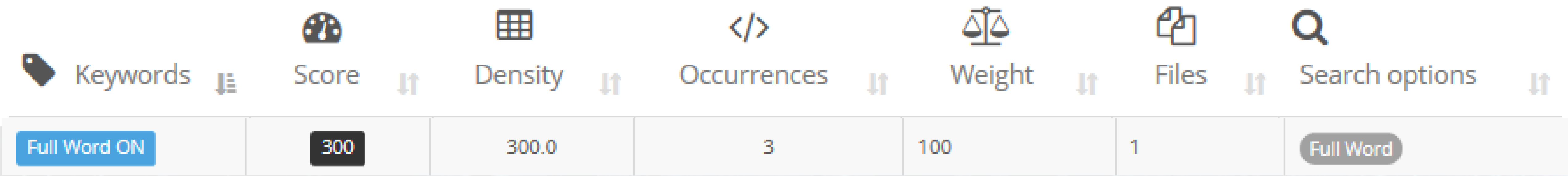
The Keyword Analyzer detected 3 occurrences within a one unique file. Because the weight is set at 10, the total score equals 30. The Density corresponds to the total score divided by the number of file(s).
Bear in mind that this score embeds all the keyword items that belong to the keyword group called “Ex. Case Sensitive On”.

Keyword Analyzer behavior

Example #3 – A Keyword Group using the Full Word Option

```
<keywordGroup name="Full Word ON" weight="100" sensitive="0"
full_word="1">
                <keywordItem>casthighlight</keywordItem>
</keywordGroup>
```

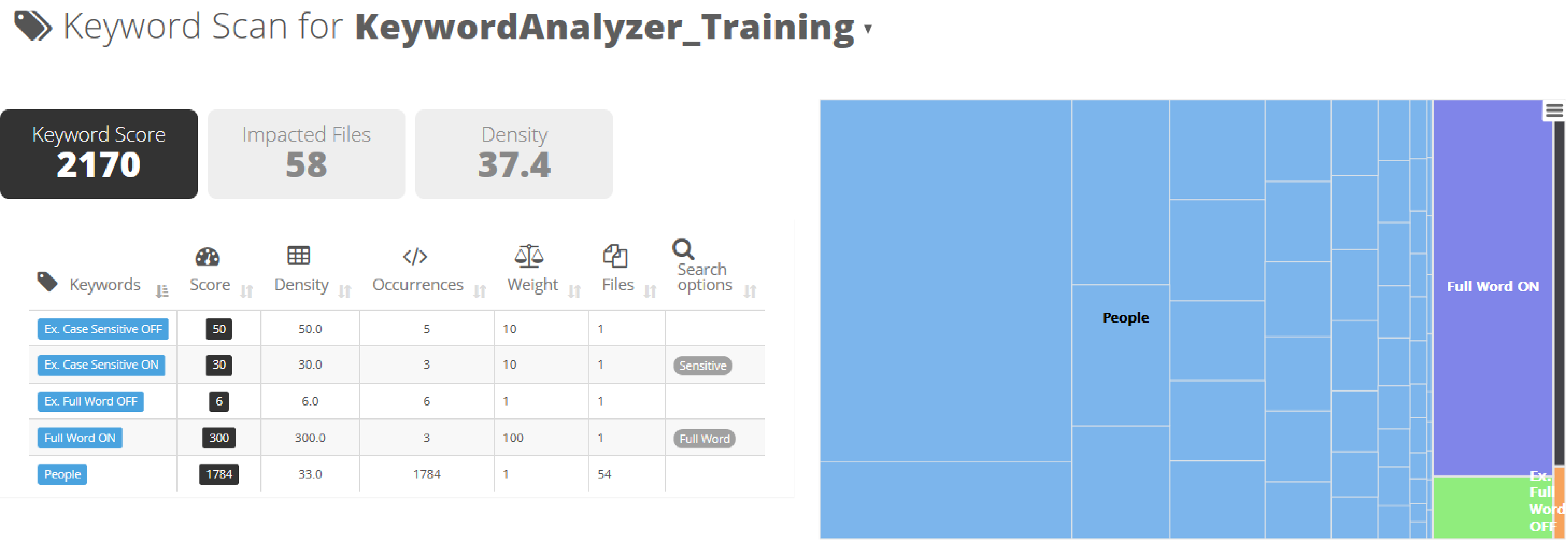
The Keyword Analyzer looks for all the keyword items and applies a weight of **100** each time an occurrence is found. The configuration is not Case Sensitive but restrictive regarding the string i.e. if some characters precede or/and follow the searched string, it won't count. It must be exactly the same characters.



Keywords	Score	Density	Occurrences	Weight	Files	Search options
Full Word ON	300	300.0	3	100	1	Full Word

The Keyword Analyzer detected 3 occurrences within one unique file. Because the weight is set at 100, the total score equals 300. The Density corresponds to the total score divided by the number of file(s).

Insights at the Application Level



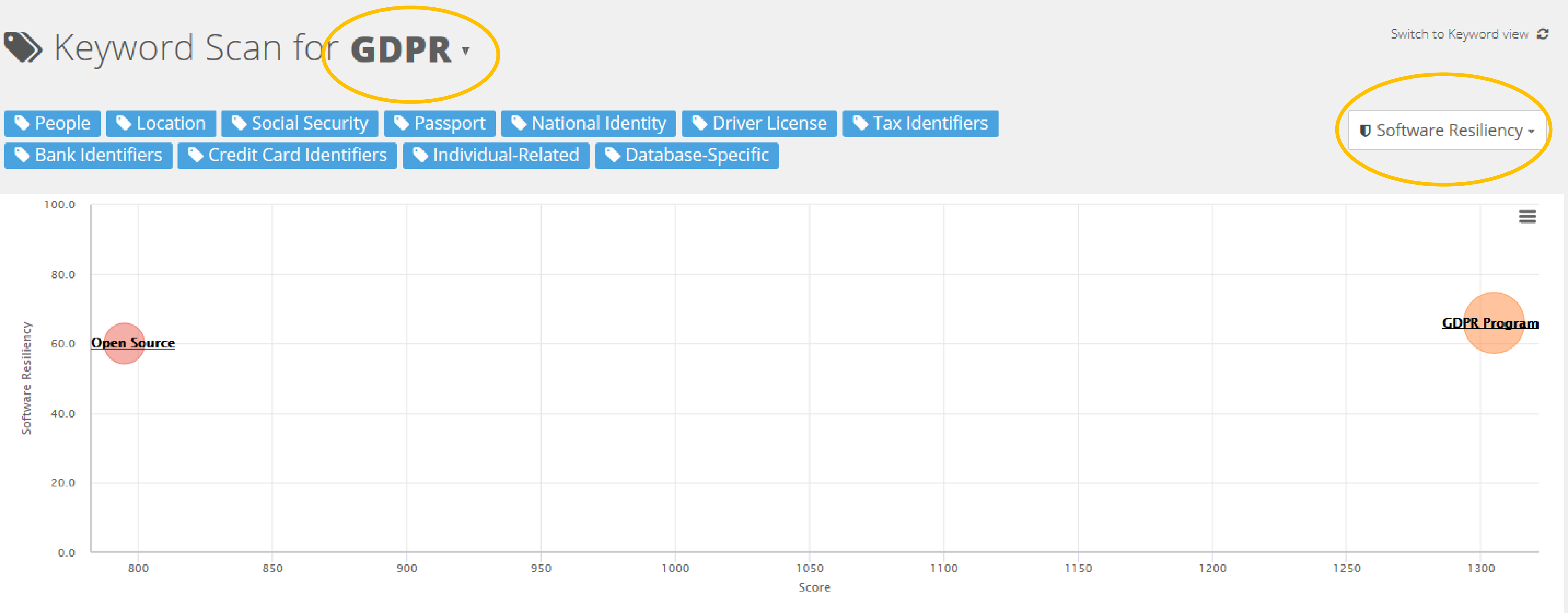
Keyword Scan for **KeywordAnalyzer_Training**

Keyword Score	Impacted Files	Density
2170	58	37.4

Keywords	Score	Density	Occurrences	Weight	Files	Search options
Ex. Case Sensitive OFF	50	50.0	5	10	1	
Ex. Case Sensitive ON	30	30.0	3	10	1	Sensitive
Ex. Full Word OFF	6	6.0	6	1	1	
Full Word ON	300	300.0	3	100	1	Full Word
People	1784	33.0	1784	1	54	

The Keyword Score sums any independent Keyword Groups score as well as the number of impacted files in order to compute an overall density of Keywords amongst the application.
The diagram on the right hand side displays a visual representation of the files per Keyword group. The size of the shape depends on the density. By clicking on a specific Keyword Group, the diagram shows the file names.

Insights at the Portfolio Level

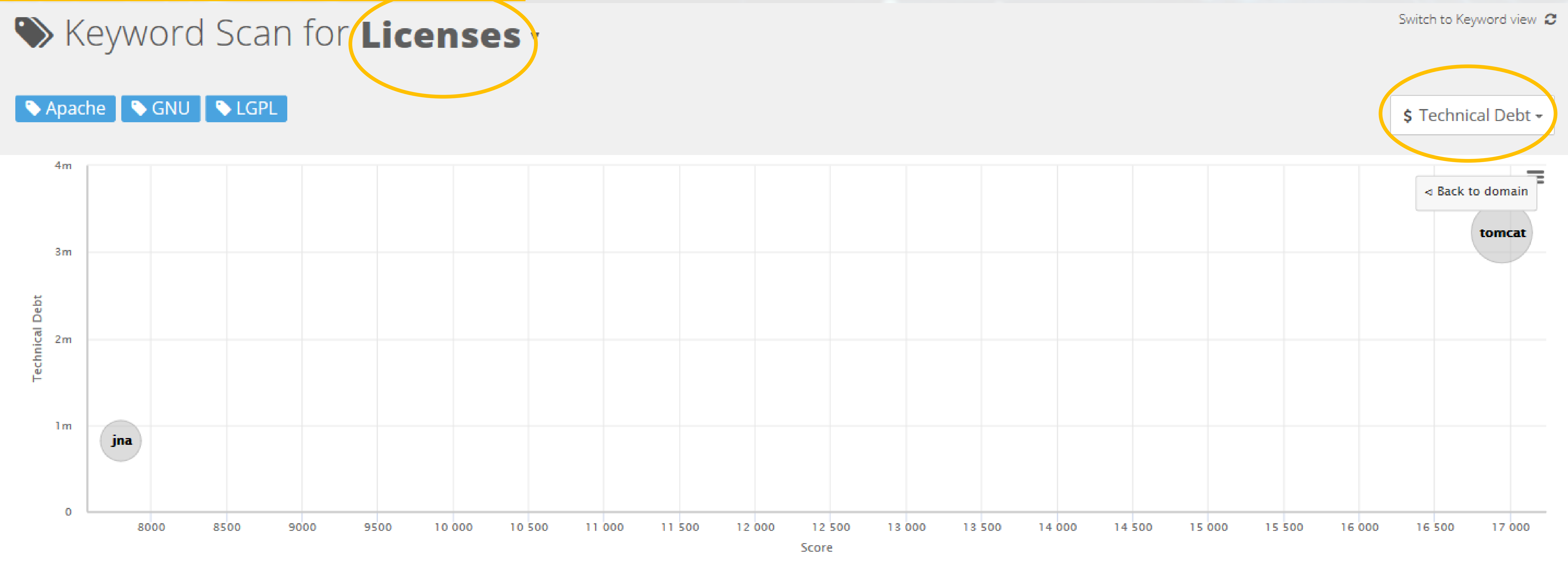


Keyword Scan for GDPR

At the top, you can use the drop down list to pick the Keyword Scan Name, here “GDPR”. The Y-axis represents the Software Resiliency by default but it can be changed by other factors such as CloudReady Index, FTEs, Size … The X-axis is the score. All the Keyword Groups appear as blue tags at the top.

The bubbles are CAST Highlight Domains which represent logical groupings of applications. When selecting a domain, we drill down to all the applications that are attached AND contain associated keywords.

Insights at the Portfolio Level



Keyword Scan for **Licenses**

Here is another example where the Keyword scan has been switched from “GDPR” to “**Licenses**”. We count 3 Keyword Groups: “Apache”, “GNU” and “LGPL”. The Y-axis corresponds to the Technical Debt and the bubbles represent the applications.

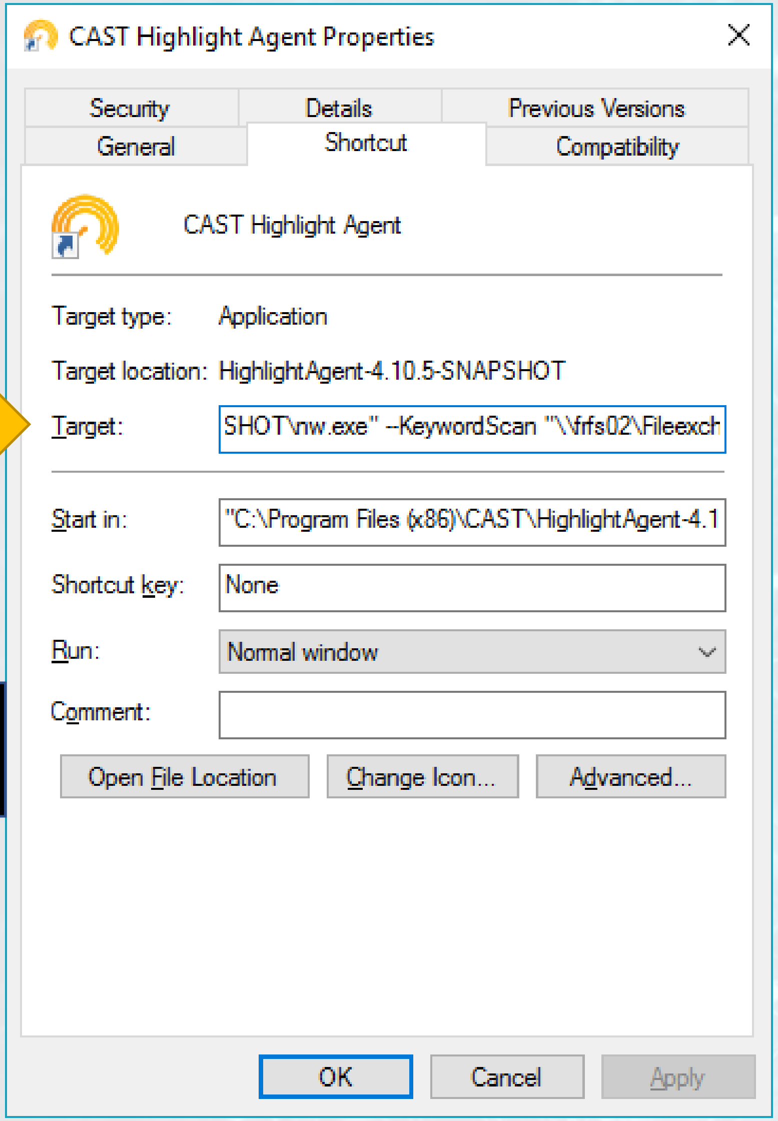
Additional Info #Enable the Local Agent UI

Enable the Keyword Analysis from the Local Agent UI
In case you don’t use the command line to orchestrate & execute the scans, there is an option to activate the Keyword Analysis directly from the Local Agent User Interface.

- Reach the Shortcut Properties of “CAST Highlight Agent”
- Identify the Target field
- Don’t change the path but simply add the following expression
- **--KeywordScan “PATH\KEYWORD_CONFIG_FILE.xml"**
- Here is an example:

```
"C:\Program Files (x86)\CAST\HighlightAgent\nw.exe" --KeywordScan "C:\CLI\keywordanalyzer\hl_keywords_GDPR_simple.xml"
```

Once the standard scanning process is over, the generated zip file will also embed the keyword analysis result CSVs. From here, simply upload the zip file to CAST Highlight Application Scans Page.



CAST Highlight Agent Properties

CAST Highlight Agent

Target type: Application

Target location: HighlightAgent-4.10.5-SNAPSHOT

Target: SHOT\nw.exe" --KeywordScan "\\frfs02\Fileexch

Start in: "C:\Program Files (x86)\CAST\HighlightAgent-4.1

Shortcut key: None

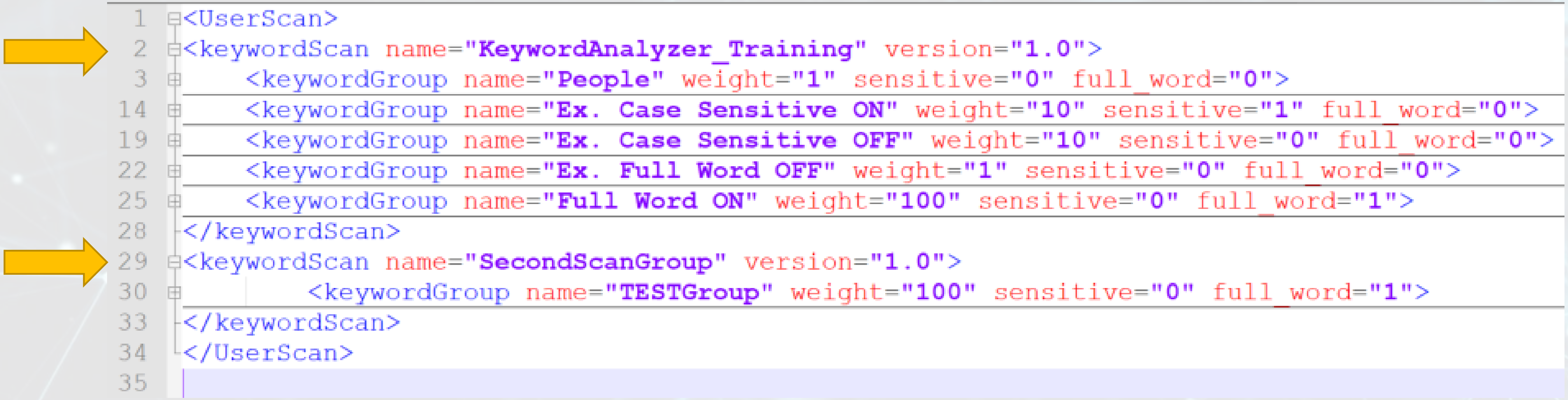
Run: Normal window

Comment:

Additional Info #multiple Keyword topics

One Configuration file for multiple Keyword Scan Analysis

Note that it’s totally feasible to include several Keyword Scans through the same configuration file

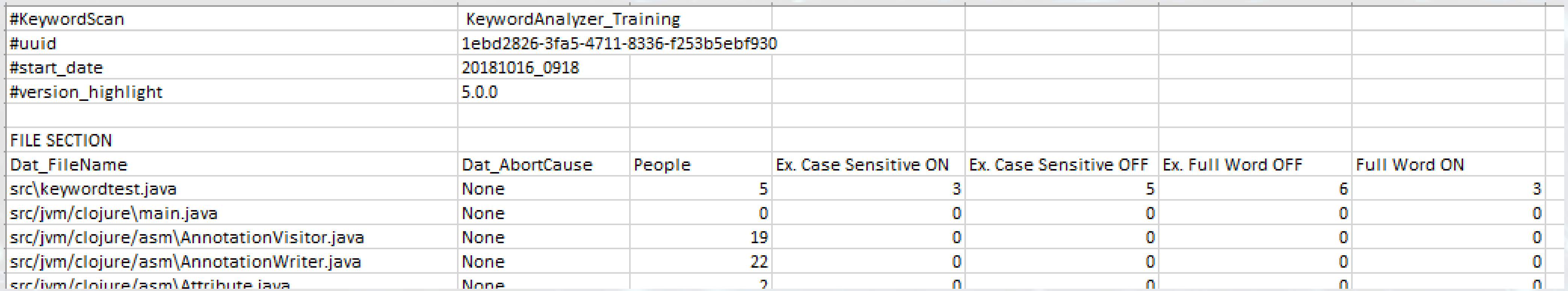


```
<UserScan>
<keywordScan name="KeywordAnalyzer_Training" version="1.0">
    <keywordGroup name="People" weight="1" sensitive="0" full_word="0">
    <keywordGroup name="Ex. Case Sensitive ON" weight="10" sensitive="1" full_word="0">
    <keywordGroup name="Ex. Case Sensitive OFF" weight="10" sensitive="0" full_word="0">
    <keywordGroup name="Ex. Full Word OFF" weight="1" sensitive="0" full_word="0">
    <keywordGroup name="Full Word ON" weight="100" sensitive="0" full_word="1">
</keywordScan>
<keywordScan name="SecondScanGroup" version="1.0">
        <keywordGroup name="TESTGroup" weight="100" sensitive="0" full_word="1">
</keywordScan>
</UserScan>
```

Important: you can only associate **1** Keyword Analyzer Configuration file by **Scan**. Which forces you to use the technique above in case you want to separate the **Topics** (i.e. GDPR, License…).

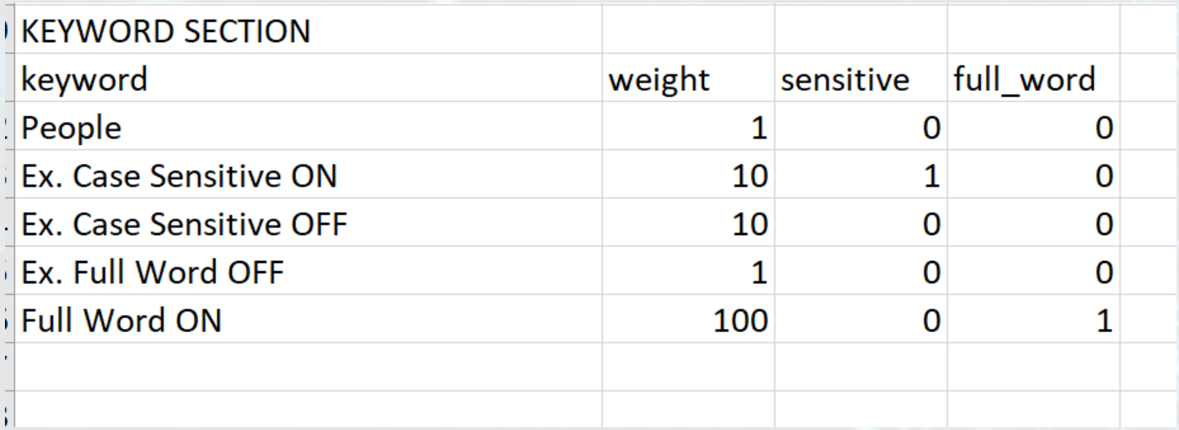
Additional Info #Keyword Detailed Report

Excel Reports count the occurrences by file and by Keyword Group.
The Keyword Analyzer engine generates CSV reports. There is one file by technology. Here is the naming convention: Java-YYMMDD_HHMM.KeywordScan.*KeywordAnalyzer_Training*.csv

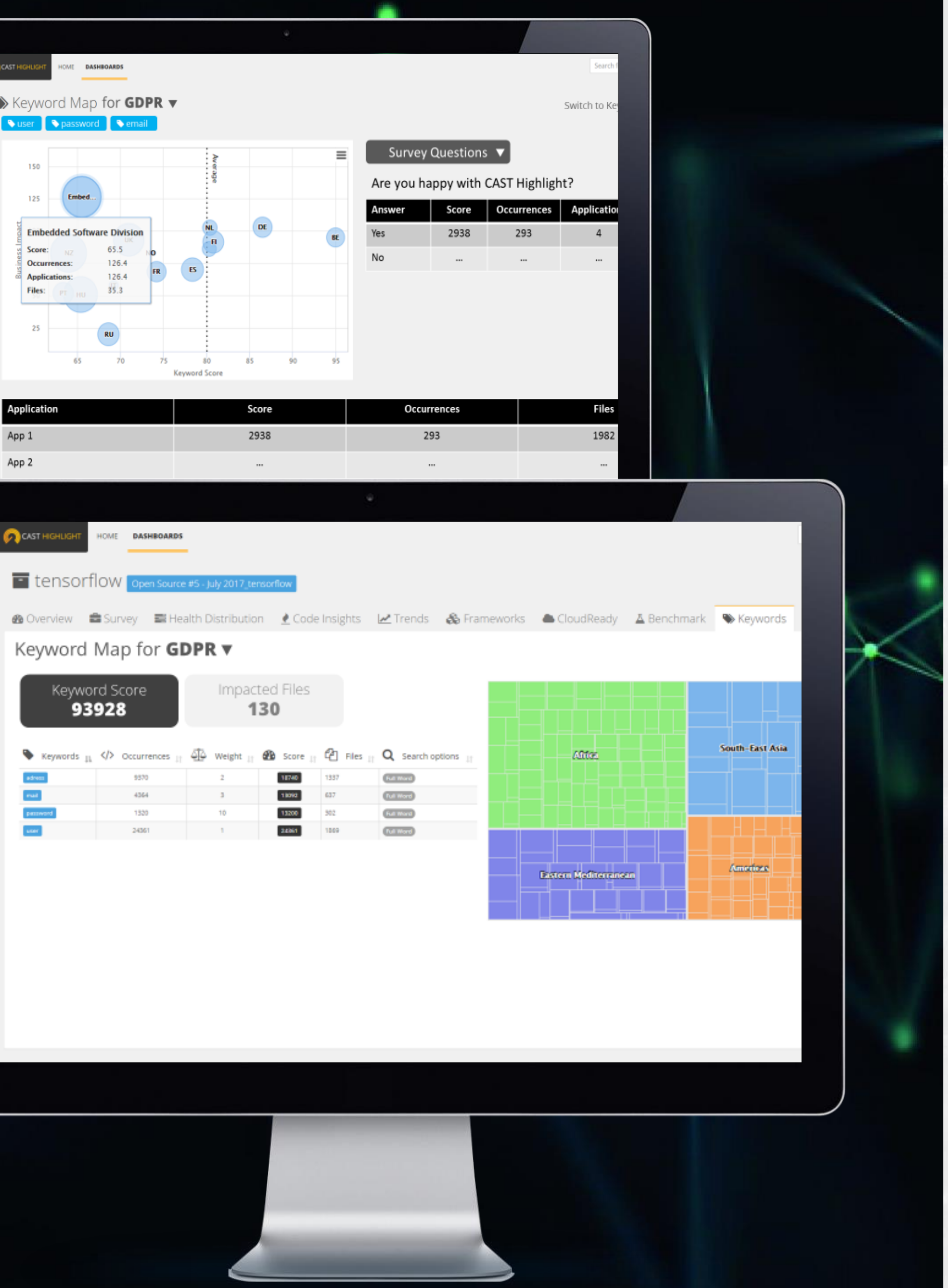


#KeywordScan	KeywordAnalyzer_Training					
#uuid	1ebd2826-3fa5-4711-8336-f253b5ebf930					
#start_date	20181016_0918					
#version_highlight	5.0.0					
FILE SECTION						
Dat_FileName	Dat_AbortCause	People	Ex. Case Sensitive ON	Ex. Case Sensitive OFF	Ex. Full Word OFF	Full Word ON
src\keywordtest.java	None	5	3	5	6	3
src/jvm/clojure\main.java	None	0	0	0	0	0
src/jvm/clojure/asm\AnnotationVisitor.java	None	19	0	0	0	0
src/jvm/clojure/asm\AnnotationWriter.java	None	22	0	0	0	0
src/jvm/clojure/asm\Attribute.java	None	2	0	0	0	0

The file also contains the initial Configuration, including the weight and If sensitive & full word options are activated or not.



KEYWORD SECTION			
keyword	weight	sensitive	full_word
People	1	0	0
Ex. Case Sensitive ON	10	1	0
Ex. Case Sensitive OFF	10	0	0
Ex. Full Word OFF	1	0	0
Full Word ON	100	0	1



Keyword Map for **GDPR** ▼

Embedded Software Division

Score:	65.5
Occurrences:	126.4
Applications:	126.4
Files:	35.3

Are you happy with CAST Highlight?

Answer	Score	Occurrences	Applicatio
Yes	2938	293	4
No	...	...	...

Application	Score	Occurrences	Files
App 1	2938	293	1982
App 2	...	...	...

tensorflow Open Source #5 - July 2017_tensorflow

Keyword Map for **GDPR** ▼

Keyword Score 93928

Impacted Files 130

Keywords	Occurrences	Weight	Score	Files	Search options
adress	9370	2	18740	1337	Full Word
mail	4364	3	13092	637	Full Word
password	1320	10	13200	302	Full Word
user	24361	1	24361	1869	Full Word

Highlight’s Main Features for Keyword Analysis…

- A survey to get a first level of app filtering/assessment
- A code scan with customizable keywords
- Keywords can be grouped and weighted
- Search options: full/partial word, case sensitive
- Portfolio-Level dashboard to consolidate keyword results
- Application-Level dashboard to investigate in depth

CAST HIGHLIGHT

GDPR and beyond…

HiPAA, PCI DSS, CIP Standards, MiFID, …

CAST