User Manual

HealthDataAnonym: A Tool for the Anonymization of Health Data

English user manual version: 0.9

Version created by: Sergio Ilarri
Creation: April 3, 2024 (Sergio Ilarri)
Last modification: April 3, 2024 (Sergio Ilarri)

OUTLINE

In Section 1, the installation manual of the application, which can be run on a command-line console, is shown. In Section 2, the different user manuals of two versions of the application (one with no GUI, to execute unattended, and another one with a GUI) are shown. Section 3 explains how to generate a .jar file of the application. Finally, in Section 4, we explain how an automated test of the impact of anonymization on data mining (classification) tasks can be carried out.

1. Installation

In this section, we describe the basic installation steps. We will also explain the folders and files that will be needed for the complete setup of the execution environment. In the following, we consider the Ubuntu operating system with version 20.04.3 LTS, but the installation for other operating systems would be similar.

Firstly, we will begin with the installation of the necessary tools, programming languages, and libraries needed to correctly run the application. Since the core of the application is implemented in the Java programming language, it requires Java to be available on the system. Similarly, the application uses a series of script files implemented in the Python programming language. Below are the required tested versions of each programming language:

- **Java 11**. Version: 11.0.11 or higher.
- **Python 3**. Version: 3.9.14 or higher.

Below are the commands necessary to perform the installation of the Java programming language:

```
sudo apt update
sudo apt install default-jre
```

It is possible that if you install this programming language for the first time, certain errors may occur. One of them may be due to the use of two *PPAs (Personal Package Archives)*. To successfully resolve this, both PPAs must be removed, and then the commands shown earlier should be executed again. In order to remove the PPAs, you can execute the following:

```
sudo apt-add-repository -r ppa:gnome3-team/gnome3
sudo apt-add-repository -r ppa:philip.scott/spice-up-daily
sudo apt update
```

Once the Java programming language has been installed correctly, you can proceed with the installation of the Python language. In this case, Python 3 will be installed. Below are the commands necessary for its installation:

```
sudo apt update
sudo apt install software-properties-common
sudo add-apt-repository ppa:deadsnakes/ppa
sudo apt install python3.9
python3.9 -- version
```

Some of the files programmed in Python utilize a library called *Pattern*. This library may or may not be included with the previous installation of Python 3. For its installation, the following commands can be executed:

```
pip install pattern3
```

Next, you should install the *Jython* library, which will be use to enable the communication between the Java code and the Python code. For that purpose, you should do the following:

```
sudo apt install jython
```

Finally, the application uses a natural language processing tool called *spaCy*. Below are the commands needed for installing *spaCy* for the Linux operating system, in Spanish:

```
pip install -U pip setuptools wheel
pip install -U spacy
python -m spacy download es_core_news_sm
```

The installation of *spaCy* can also be done for other operating systems using pipes, Conda, or "from source". The steps for this procedure are explained on the official page of *spaCy*, here. It is of vital importance that the selected language be Spanish, as the developed tool targets documents written in Spanish.

Once the installation process of the necessary tools for execution is completed, you can proceed with the configuration of the environment. Inside the ".zip" file where this manual was located, you can also find a folder named "Pruebas" (or "Tests", in English), which contains all the necessary folders and files needed for execution:

- **Folder *"Colecciones"*** (*"Collections",* in English). This folder contains the data collections that the application queries during its execution.
- **Folder *"Resultados"*** *("Results",* in English). Anonymized files will appear in this folder when the anonymization process is completed.
- **File *"Configuration.txt"***. This is the application's configuration file. It is a key file when executing the tool without GUI (Graphical User Interface). It allows configuring the level of anonymization to apply and/or select the set of attributes to anonymize from the input file. Additionally, you will need to indicate the paths of the files to anonymize, as well as the paths of the folders described above.
- **Python files (.py)**. These are the files implemented in the Python programming language that the application requires for its proper functioning. It has files with names such as *"analyzer.py"*, *"analyze_names.py"*, *"analyze_names_surnames.py"*, *"personas.py"* ("personas" is the Spanish term for "people" in English), and "syntactic_analysis.py".
- **Folder *"Informes"*** *("Reports",* in English). In this folder, you can include the files (i.e., medical reports) to anonymize with the application. The files can have extensions ".txt" or ".zip". The program will automatically detect the type of file.
- **Executable *"App.jar"***. This is the application to be executed, in jar format.

The structure of folders should resemble what is shown in Figure 1.

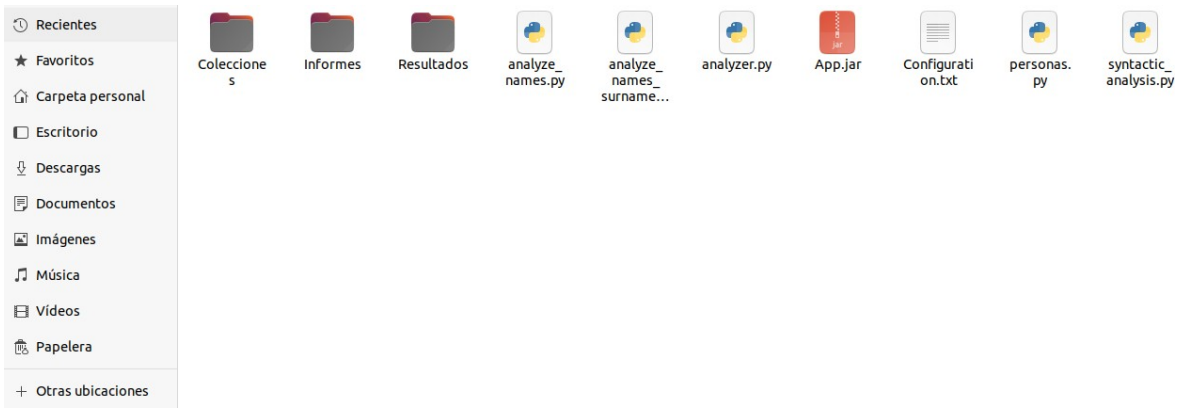

Figure 1 Final structure or folders

Once you have reached this point, the environment is correctly configured to execute the application. Additionally, we have also built a virtual machine with the Ubuntu operating system version 20.04.3 LTS, which reproduces a complete execution environment without the need to follow the previous steps.

2. User manual

In this section, we describe the basics of the two versions of the application developed. Firstly, in Section 1.2.1, we focus on the version without graphical user interface (useful to perform large-scale unattended evaluations). Secondly, in Section 1.2.2, we describe the GUI-based version of the application (useful to use the tool interactively).

2.1 Version with no GUI (exploiting a configuration file)

If you are going to execute the application from a provided virtual machine (mentioned in Section 1), you should download and import it into VirtualBox. The download may take a few minutes. Once it is downloaded and imported into VirtualBox, you should access it using the username *"usuario"* and the password *"usuario"* ("usuario" means "user" in Spanish; of course, these default credentials can be easily changed as required). Once you run the Virtual Machine (VM), you will see the desktop. In the home folder, there is a folder named "Pruebas" ("Tests", in English), where all the necessary files and folders for the correct operation of the application are located, as well as the application itself; the contents of this folder were explained in detail in Section 1.

The configuration file should specify the options corresponding to the type of anonymization you wish to apply. In Figure 2, the structure of the content of this configuration file is shown.

```
Write an X on those attributes you want to anonymize:

-Personal information:
-Context information:
-Style life information:
-Events information:

Write an X on the level of anonymization you want to apply: (You only can choose one of these four options. You must choose one.)
-Do not apply:
-Low level:
-Medium level:
-High level:

Write the path to the file which you want to get anonymized in the next line:

Write the path to the results folder in the next line:

Write the path to the collections folder in the next line:
```

Figure 2 Structure of the configuration file

As you can see in Figure 2, the first group of options to select refers to the set of attributes you want to anonymize: *"Personal information"*, *"Context information"*, *"Lifestyle"*, and/or *"Events information"*. In this case, you should mark with an X those attributes you wish to anonymize. It can be one or more sets of attributes. To clarify which attributes each group collects, Table 1 is presented.

Category	Attributes
Personal information (PI)	Name, surname, age, ethnicity, ID number, phone number, address, email, date of birth, date of death, Social Security Number (SSN), professional license number (for doctors), and gender.
Contextual information (CI)	Relatives, work, places, countries, cities, and municipalities.
Lifestyle information (LS)	Sports and habits.
Clinical events information (CE)	Hospitalization, admission date, date associated to emergencies or other events, and discharge date.

Table 1 Categories of data

The second group of options presented in the configuration file refer to the level of anonymization you wish to apply: *"Low level"*, *"Medium level"*, *"High level"*, or *"Don't apply"*. Unlike the previous case, only one of these levels can be selected (as they are, obviously, mutually exclusive); if you do not wish to apply any, then you should select the *"Don't apply"* option. The attributes considered by each group in this section are as follows:

- **Don't apply**: no anonymization level is applied.
- **Low level**: it implies removing attributes with high sensitivity (and also those with other lower sensitivity levels but which appear alongside other attributes that increase their sensitivity to high).
- **Medium level**: it implies removing attributes with high or medium sensitivity (and also those with other lower sensitivity levels but which appear alongside other attributes that increase their sensitivity to medium or high).
- **High level**: it implies removing all potentially-sensitive attributes, including those with low sensitivity.

Next, in the configuration file, you can find fields to fill in regarding configuration parameters that must be provided to the application:

- **File to anonymize**. Here, you should indicate the path of the file to anonymize. The files to anonymize can be included in the *"Informes"* (*"Reports"*, in English) folder.
- **Results folder**. Here, you should indicate the path of the *"Resultados"* (*"Results"*, in English) folder, where the anonymized files will be obtained after execution. It is advisable that this folder be empty before launching the application. If you want to save them in another folder, you can specify the path of the desired folder.
- **Collections folder**. Here, you should indicate the path of the *"Colecciones"* (*"Collections"*, in English) folder.

It should be emphasized that paths should be adjusted according to the path formatting rules of the operating system where the application will be executed. That is, if the operating system is Windows, the path separator should be \\ and / should be used instead for Linux and macOS, as illustrated in Figure 3 and Figure 4.

```
Write the path to the file which you want to get anonymized in the next line:
C:\Users\Public\Desktop\Informes\report1.txt
Write the path to the results folder in the next line:
C:\Users\Public\Desktop\Resultados
Write the path to the collections folder (dictionaries) in the next line:
C:\Users\Public\Desktop\Colecciones
```

Figure 3 Example of path in Windows (path parameters defined in the configuration file)

```
La ruta del fichero introducido es la siguiente: /home/usuario/Pruebas/Informes/prueba2.txt
La ruta de la carpeta de salida es la siguiente: /home/usuario/Pruebas/Resultados
La ruta de la carpeta de colecciones es la siguiente: /home/usuario/Pruebas/Colecciones
```

Figure 4 Example of path in Linux (feedback provided when the tool is executed)

Once all of this is written, the configuration file *"Configuration.txt"* should have an appearance similar to the one shown in Figure 5.

```
Write an X on those attributes you want to anonymize:

-Personal information:
-Context information:
-Style life information:
-Events information:

Write an X on the level of anonymization you want to apply: (You only can choose one of these four options. You must choose one.)
-Do not apply:
-Low level: X
-Medium level:
-High level:

Write the path to the file which you want to get anonymized in the next line:
/Volumes/OWCEnvoyProFX/Workspace/healthAnonym/TextAnonym/FunctionalAppForWebsite/Informes/prueba1.txt
Write the path to the results folder in the next line:
/Volumes/OWCEnvoyProFX/Workspace/healthAnonym/TextAnonym/FunctionalAppForWebsite/Resultados
Write the path to the collections folder (dictionaries) in the next line:
/Volumes/OWCEnvoyProFX/Workspace/healthAnonym/TextAnonym/FunctionalAppForWebsite/Colecciones
```

Figure 5 Example of configuration file

Now everything is ready to execute the application. You should indicate the path of the configuration file *"Configuration.txt"* as an argument. The following one is an example of command to launch the application:

```
java -jar App.jar /home/usuario/Pruebas /Configuration.txt
```

When the application has finished its execution, it will display on the screen an output similar to what is shown in Figure 6. As shown in the figure, the anonymized files are located in the folder specified in the configuration file as *"carpeta de resultados"* (*"results folder"*, in English); the program also informs about the path specified in the arguments.

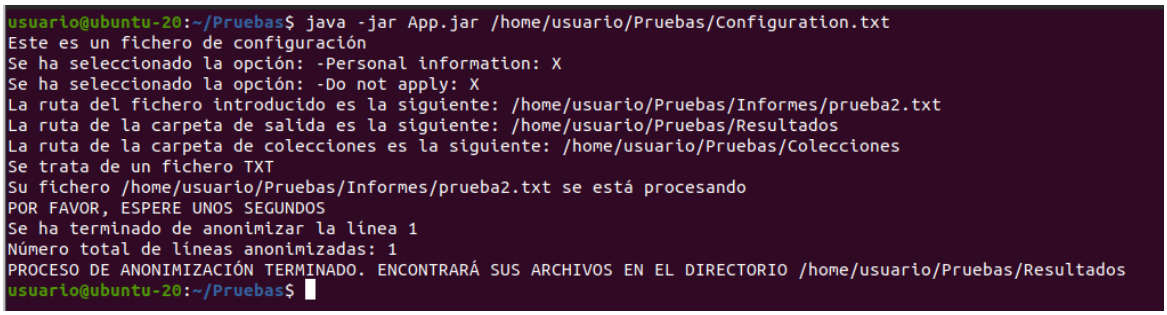

Figure 6 Execution without GUI finished

2.2 Version with GUI

The GUI tool consists of two distinct parts: anonymization process and data mining.

Anonymization process

The anonymization process is found in the *"Data"* window. In this window, documents to process can be loaded. It can be a single file in ".txt" format or a set of text files compressed into a ".zip" file.

Next, the user must select the attribute or attributes that he/she wishes to eliminate. As explained in Section 2.1, these attributes are grouped into four categories: *"Personal information"*, *"Context information"*, *"Lifestyle"*, and *"Clinical events"*. The user can select one, several, or none of the groups. The attributes collected by each group were specified in Table 1.

Then, as explained in the previous section, the user can select the anonymization level, which is complementary to the choice of categories of attributes to remove. As indicated before, the system offers three levels of anonymization: *"Low level"*, *"Medium level"*, and *"High level"*. The *"Low level"* option implies removing attributes with high sensitivity, the *"Medium level"* implies removing attributes with medium or high sensitivity, and finally, the *"High level"* implies removing all potentially-sensitive attributes, as it will remove those classified with high, medium, or low sensitivity. In addition, the user can select the option *"Don't apply"*, which means that no

anonymization level will be applied to the document being processed, and therefore it will only remove attributes in the categories explicitly selected by the user.

After completing the configuration of the anonymization process to be performed, the user clicks the *"Download folder"* button. When the process is finished, they can find the anonymized documents in the *"Resultados"* (*"Results"*, in English) folder.

Data mining

On the other hand, the data mining part is located in the *Mining* window, where you can configure the *Cross Validation* evaluation technique to be carried out. Before explaining how the data mining GUI is organized, it is important to highlight how to structure the folders to ensure a successful evaluation.

Once the datasets of documents to be evaluated are ready, a folder should be created for each of the labels to be considered in the evaluation. Each of these folders should contain the documents belonging to that label. For example, if you want to classify between personal and non-personal documents, you should create two folders: one folder (named, for example, *True*) that contains the personal documents, and another folder (named, for example, *False*) that contains the non-personal documents. This is how labeling is done. In a separate folder, the same procedure should be followed for the anonymized dataset.

The tool allows loading these folders. Therefore, you should select in the application the parent folder where the subfolders representing the evaluation labels reside. Once both folders are set, you should configure the evaluation to be performed. Firstly, you need to select the classifier, the number of "folds" for cross-validation, and the target class. For the classifier, you can choose among ZeroR, OneR, Naive Bayes, and SMO. Additionally, if you select the *"All classifiers"* option, then the evaluation will be performed with all classifiers available.

Once the configuration is completed, you can click on the *"Compare"* button. When the evaluation process finishes, a new *"Results"* window will appear on the screen with the results obtained.

3. Generate.jar file

We have set up a Maven project that allows generating a .jar file by using a command. To do this, you need to navigate to the configurationProject folder. Inside it, you should execute the following command:

```
mvn clean install package
```

Once it has been executed, you should navigate to the folder named *"target"*, where you will find a .jar file initially named "configurationProject-0.0.1-SNAPSHOT.jar". This is the final file with which you can run the developed application. Of course, you can easily adjust the Maven configuration according to your needs and goals.

4. Automated tests for data mining

In this section, we explain how automated tests can be conducted for the data mining part. Firstly, you need to have the dataset to be used for the *Cross Validation* technique. You should have both the original dataset and the anonymized dataset to perform the evaluation.

To label the datasets, you need to create a folder for each of the classes you want to consider (one for each label). As an example, let us imagine a situation where we want to evaluate the impact of data anonymization on a classification task that needs to distinguish between two different types of documents: documents containing personal data and documents corresponding with clinical guidelines. For this scenario, two folders could be created: the "true" folder, where documents related to personal data are stored, and the "false" folder, where documents related to clinical guidelines are stored. The same procedure should be followed for the anonymized datasets. As a result, you should have a folder where the original data are stored in the corresponding subfolders for labeling, and the same for another folder which stores the anonymized data.

Next, you can proceed to execute the application with the GUI version. As mentioned in Section 2.2, this tool offers two tabs: *Data* and *Mining*. In this case, you should access the *Mining* tab. This tab is shown in Figure 7.

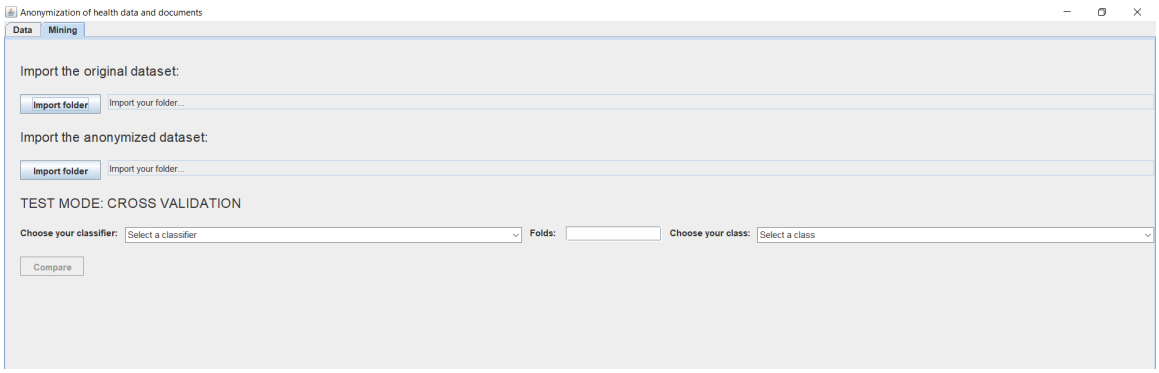

Figure 7 Data mining window

The *Mining* tab allows the user to load the data for the evaluation. To load the data correctly, the parent folder containing the subfolders with labeling should be selected. This should be done for both the original and anonymized data. Next, the user should configure the evaluation options. He/she could either select *"All classifiers"* in the *"Choose your classifier"* option or select an individual classifier to test. He/she should also specify the number of "folds" for evaluation and the target class to consider.

For example, by selecting the *"All classifiers"* option, the evaluation will be conducted with the four classifiers currently offered in the tool: ZeroR, OneR, Naive Bayes, and SMO. When this process finishes, all the results obtained in the evaluation with each classifier and each dataset will be displayed on the screen.

Contributors

Researchers

- Sergio Ilarri (University of Zaragoza)
- Carlos Tellería (IACS research health institute)

Students (final degree projects)

- Marta Morales

Acknowledgments

- Project PID2020-113037RB-I00 / AEI / 10.13039/501100011033 — Next-gEnerATion dAta Management to foster suitable Behaviors and the resilience of cItizens against modErN ChallEnges (NEAT-AMBIENCE)
- Government of Aragon (COSMOS research group; last group reference: T64_23R; previous group reference: T64_20R)