semantia LAB	**Project:** GallitoAPI **Contact:** info@semantialab.es	**Version:** 2	**Date:** 04-01-2014
	Name: mainFiles_v1.docx	**Document Type:** User Manual	

Main Files

Document information

This document describes the main files in the **GallitoAPI** environment, some of which are necessary for its operation, while other files are optional and can be stored unchanged when not required by the functions used.

Introduction

Generally speaking, **GallitoAPI** can be said to use the files that define the underlying thematic space (loaded, active space), as well as a "cfg" configuration file, a "log" file, and a retrieval file when **GallitoAPI** creates a cache. The other files support **GallitoAPI's** functionalities and are described below.

Files

1. cfg.ini file

This is the file in which **GallitoAPI** is parametrised. Its location is:

C:\inetpub\wwwroot\GallitoAPI\bin\data

Its parameters and values are as follows:

- **[CI]** Identifies whether the Construction-Integration algorithm is available to project a text into the thematic space (for more information, see the research papers: Jorge-Botana, León, Olmos, Hassan-Montero, 2010; Jorge-Botana, G., Olmos, R., León J.A, 2009; Jorge-Botana, Olmos and Barroso, 2015). Parameters are defined by the number of neighbours to be extracted and by the number of list terms that serve to calculate the final vector.

  - **available=[true,false]**
    Use of this model is not recommended in texts longer than one sentence. If you are not sufficiently familiar with this model, we recommend assigning the value "false".
  - **neighbours=0**
  - **list=20**

- **[DestinySelection]** This parameter is used by functions that manage the partial texts to select the category of the text(s) that are most similar to the analysed text (for more information, see the research papers: Jorge-Botana, Olmos, and Barroso, 2012; Jorge-Botana, Olmos, and Barroso, 2015).

  Modes:

  "One by one": a text is compared with each of the partial texts. The criterion is one to one, and scores for each category are extracted with the partial text from each category that displays the highest degree of similarity.

  "byAverage4": as in the previous case, a text is compared with each of the partial texts. Scores for each category are extracted with the average similarity to the four partial texts of each category that are most similar to the analysed text.

  These two modes to calculate the categories that best fit the analysed text can be applied to the functionalities that use partial summaries using **distances based on LSA** or **Levenshtein distances**. Both possibilities can be separately parametrised.

  - **ModeLSA=[1,2]**
  - **ModeLev=[1,2]**

  However, the values ModeLSA=2 and ModeLev=1 are recommended.

- **[logisticThreshold]** For the "DoRoute" function, a type of correction of the final category score can be applied (for more information about use of this parameter, see the research paper: Jorge-Botana, Olmos & Barroso, 2012).

- **logisticFunctionForCompareContentToGold]** A transformation that has an impact on the value of the final holistic note can be added to the CompareContentToGold

function. This value is calculated on the basis of empirical data or through application of a theoretical (analytical) logistics function.

- **[Lemmatisation] Type of lemmatisation used in the training**
  - **language=[1, 2, 3, 4, 5]** (0=nothing; 1=Spanish; 2=English; 3=French; 4=German; 5=Italian)
    Note: If you are unsure of which type of lemmatisation is being used, "open" the space in **Gallito Studio** and check this data under the "properties" command.

- **[otherParameters]**

  - **availableClasses=[true,false]**
    If the training has been performed using classes, and these classes are available in the same file under the [classes] label, the class file used by Gallito must be "pasted" into this file.
  - **availableCode=[true,false]**
    If a specific code is used to identify a request, the source of the request can be known.

- **[classes]: a class list "pasted" into the file (see a little sample below)**

  .
  .
  office=classApp
  excel=classApp
  windows=classApp
  castilla=classRegion
  aragon=classRegion
  galicia=classRegion
  .
  .

- **[contentReference]** Functionalities that use a labelled thematic space require information about the meaning of the set of the "n" first dimensions. To this end, a label will be specified for each dimension in the "cfg.ini" file, as well as the descriptors selected to define the labelled concepts in the labelled space (see change of base manual). The format of this parameter is as follows:

  LABEL=DESCRIPTOR1 DESCRIPTOR2 DESCRIPTOR3…

  Example: In the following example a thematic space has been used where the 3 first dimensions have specific meanings. The labels for their meanings are INCIDENTSC, INVOICESC, and TERMINALS for dimensions 1, 2, and 3 respectively. The descriptors that were used in Gallito Studio for the change are given after the equality sign in each of the labels (for more information, see the research papers: Olmos, Jorge-Botana,

León, and Escudero, 2014; Olmos, Jorge-Botana, Luzón, Martín-Cordero, and León, 2015).

INCIDENTSC=incident problems solve connection anomaly cut
INVOICESC=expenses data invoice month account bank direct debit pay
TERMINALS= puk pin telephone malfunction blackberry mobile vibration Nokia

Note: Neither the labels nor the descriptors have any impact on the meaning or the vectorial representation of the thematic space, as meaning is not imposed in **GallitoAPI** but in **Gallito Studio**. This simply means that in this file labels and descriptors are simply specified to identify each dimension using its own label and descriptors.

2. Files with ".bnl" extension

These files represent the n-dimensional thematic space. They are generated by **Gallito Studio** once the training has ended and are automatically placed in this folder:

C:\inetpub\wwwroot\GallitoAPI\bin\data

- **US.bnl** file is the **Terms Matrix**
- **termList.bnl** is the **list of terms represented in the space**
- **features.bnl** are **characteristics of the space**
- **S.bnl** file are the **dimension weighting**
- **weights.bnl file:** a standardised figure for the entropy, specificity, or neighbourhood density of each word. This binary file can be generated mentioning any of these indices. Use of specific indices will depend on the purpose of the task. By default, **Gallito Studio** generates the training entropy file, but other files can be generated as desired.

3. “log” files

In the registry system, "log" files record the data for each call or request, as well as error data if applicable. They also record the success of the operations performed (see the "log" and trace registry system manual). The logsemanticService.log file is located in this folder:

C:\inetpub\wwwroot\GallitoAPI\log

4. Grammar file (grxml)

A grammar file in .grxml (SRGS) (http://www.w3.org/TR/speech-grammar/) format to identify the existence of linguistic constructions or structures previously defined in each of the sentences of a text. It is what it known as a "deterministic automaton". The grammar is included in the explicitContent.grxml file, which is located in this folder:

C:\inetpub\wwwroot\GallitoAPI\bin\grammars

5. Golden text file

The functions **compareGoldenToGold** and **compareGoldenToGoldByPartial** require the existence of a reference text. This text is kept in the "goldEssay.txt" file, located in this folder:

C:\inetpub\wwwroot\GallitoAPI\bin\content

For more information, see the research papers: Olmos, Jorge-Botana, León y Escudero (2014); Olmos, Jorge-Botana, Luzón, Martín-Cordero and León (2015).

6. Partial texts file

Functions requiring the existence of **partial texts** (e.g. **compareGoldenToGoldByPartial, DoRoute, DoContentAnalysisByPartial, DoContentAnalysisByPartialByLev)** require the creation of a file that includes the texts as well as information about their categories. This information has the following format:

Category ID | Category Name | Text

and it is located in this folder:

C:\inetpub\wwwroot\GallitoAPI\bin\data

7. fastRecover file

This file recovers the System in the event of an unexpected stoppage or recycling. Its name is "estructura.bnl" and it is located in this folder:

C:\inetpub\wwwroot\GallitoAPI\fastRecover.

For security reasons, the user who is authorised to write in this folder is: "**IIS_IUSRS"**.