



A Mashup Interface for Text Analysis Operations

A web application tool for mashing up automatic text analysis tools and create a customizable visual workflow. This application is linked to the [CATARSI Project](#). It was developed using [DIPAM](#): A Dashboard Interface for Python-based Applications Mashup. MITAO is a customization of DIPAM, which has been configured for the application of text analysis methods based on popular python natural language libraries.

1) Setup and launch	2
1.1) Online version	2
1.2) Setup MITAO on your machine	2
1.2.1) Requirements	2
1.2.2) Installation	2
2) Usage	4
2.1) The interface	4
2.2) Data	5
2.3) Tool	5
2.4) Building a workflow	7
2.4.1) Add a "Data" node	8
2.4.2) Add a "Tool" node	9
2.4.3) Connecting the nodes	10
2.4.4) Running the workflow	10
3) Examples	11
3.1) LDA topic modelling on textual documents	11

1) Setup and launch

Two are the options available for getting started with MITAO:

- 1) The online demo version at <http://163.172.159.152:5000/>
- 2) Setup MITAO on your own local system and use it with no restrictions.

1.1) Online version

MITAO is available at the following address: <http://163.172.159.152:5000/>. You can access it from any modern web browser, although we highly recommend to use Chrome (since it was fully tested on it). The current server hosting MITAO has limited hardware to let users do massive and complex operations, therefore the maximum data limit users can upload is set on 2MB.

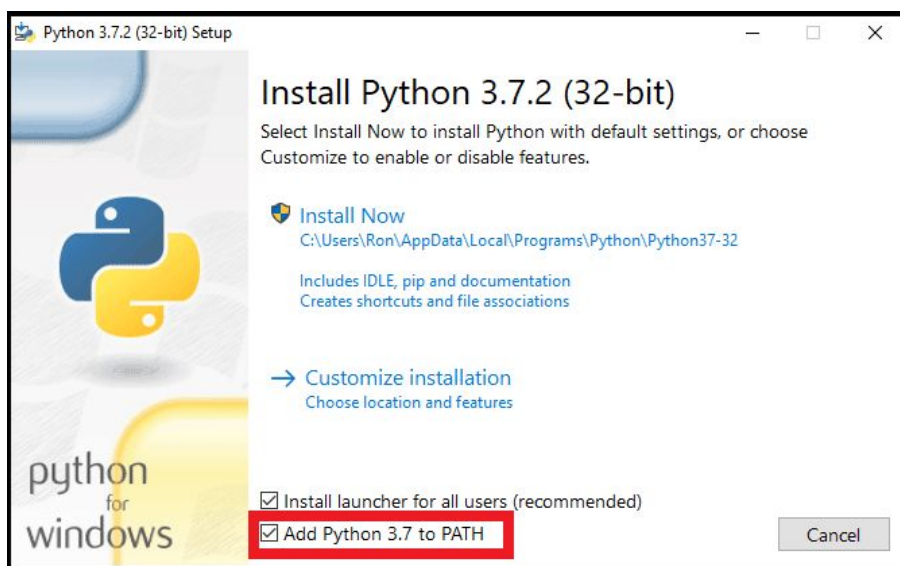
1.2) Setup MITAO on your machine

This section describes how to setup MITAO on your own local machine. It could be installed on Windows, Mac, and Linux operating systems. Once installed it could run on any modern web browser, although we highly recommend to use Chrome (since it was fully tested on it).

1.2.1) Requirements

Before installing MITAO make sure you have the following Requirements available in your system:

1. Python 3.7 programming language, [Download and install it from the official website](#). Note: Check "Add Python 3.7 to PATH".



2. The Chrome web browser, [download it from the official site](#).

1.2.2) Installation

To install MITAO on your local machine, you can choose one of the following options:

a. From the Git repository

Move to the official Git repository of MITAO at:

<https://github.com/catarsi/mitao>, and follow the Installation instructions written in the “Readme” file. (You can find a preview of the “Readme” right on the description of the project).

b. Built-in package

Download and unzip MITAO from

<https://github.com/catarsi/mitao/archive/master.zip>, or from the official beta release page at <https://github.com/catarsi/mitao/releases/tag/v1.1-beta>. Next do the following operations according to your operating system:

Mac or Linux:

Double click on setup_macos to install the tool and all its dependencies, right after that, double click on run_macos, this will run MITAO and open you the browser automatically.

Windows:

Double click on setup_windows to install the tool and all its dependencies, right after that, double click on run_windows, this will run MITAO and open you the browser automatically.

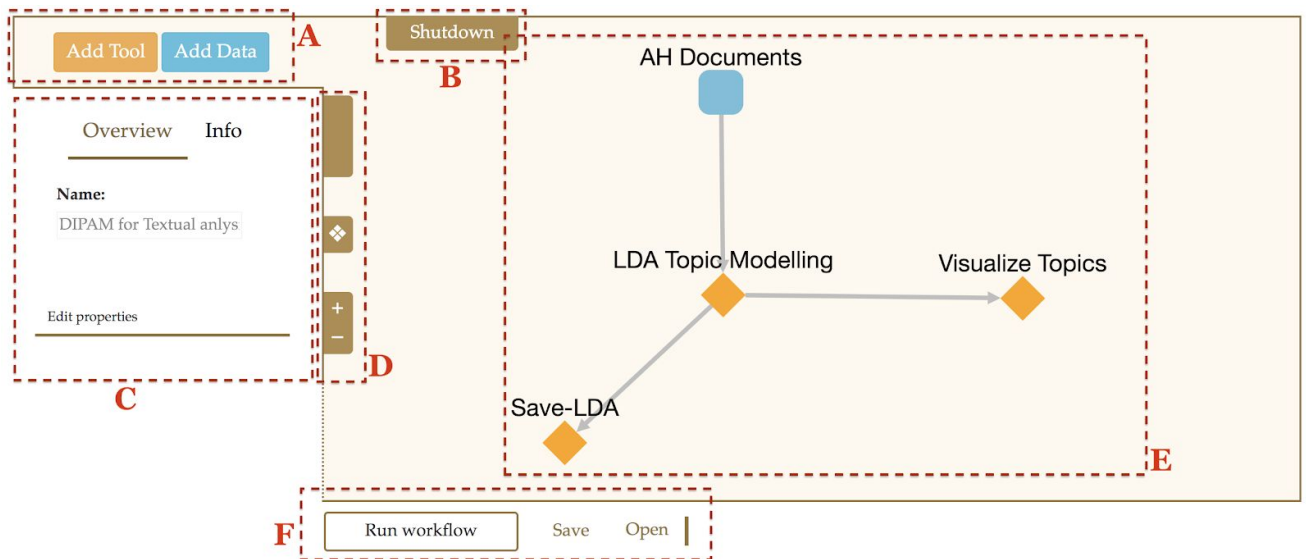
Note: some Operating System require a check to allow scripts developed by unknown users to be launched.

2) Usage

In this section we will first talk about the general interface of MITAO, then we move on defining the “Data” and “Tool” nodes. The final subsection will guide you on how to correctly build and run a workflow using MITAO.

2.1) The interface

The following image shows the interface of MITAO once it's launched.



Here we give a brief description for each annotated component from the above MITAO interface image:

- A. Buttons for adding “Tool” and “Data” nodes into the workflow:
The “Data” nodes will appear as blue rectangles, while the “Tool” nodes are represented as orange rhombuses.
- B. Shutdown button:
(this button will not appear when using MITAO from its online version)
Note: to launch again MITAO you need to double click again on the run file (see section 1.2.2).
- C. The Info panel:
When selecting a node or an edge from the diagram the info panel will display its related information under the “info” menu. Clicking on “Overview” will display general information about the Diagram. The “Edit properties” and “Remove element” buttons will be displayed according to the selected item typology.
- D. General operations:
Available supporting options to apply while editing the workflow diagram: (a) undo/redo the modifications made on the diagram; (b) fitting the entire workflow diagram into the displayed window; (c) zoom in/out (this option could be made also through your PC mouse)

E. The workflow panel:

It's the main panel of MITAO and it will contain the body of the workflow diagram. Users have the ability to move/drag the inner nodes, along with the creation of new edges to connect them.

F. Run, save and open

The "Run workflow" button launches the built workflow; "Save" let users store the workflow locally on their machines, and save the current MITAO status (it will become the default workflow to load when opening MITAO); "Open" let users choose and load a workflow.

2.2) Data

When adding a "Data" node into the workflow the Info panel ([see section 2.1](#)) displays the default attributes of the just added component, and users will have the ability to modify such attributes right on its first drop into the diagram by clicking on "Edit properties". Here we list the data nodes attributes with a brief definition, and the corresponding available values:

Attribute	Description	Value
Name	The name of the node	Any textual value.
Type	The type of data the node represents	<div>Textual document/s</div> <div>PDF document/s</div> <div>Chart image</div> <div>Chart legend</div> <div>Topics X Words table</div> <div>Documents X Topics table</div> <div>LDA coherence value</div> <div>LDA perplexity value</div> <div>Stopwords list</div>
File/s or Directory	A single or multiple files which stores the corresponding data	One or multiple files selected locally from the hosting PC.

2.3) Tool

As for the "Data" nodes, adding a "Tool" node into the workflow, will trigger the Info panel ([see section 2.1](#)), and display the default attributes of the just added node. Users can modify such attributes by clicking on "Edit properties". Here we list the "Tool" nodes attributes with a brief definition, and their corresponding available values:

Attribute	Description	Value
Name	The name of the node	Any textual value.
Type	The type of tool the node represents	Users can select one of the following values: Topic modelling with LDA, Filter the text, save the files, visualize document topics, visualize topics words, an convert PDF document/s to text.

Additional attributes will be displayed according to the “Tool” type. The underneath table lists the attributes according to the type of the selected tool, we give a brief description of the tool typology, the inputs it requires, and its outputs. Notice that the inputs and the outputs are all “Data” node types ([see section 2.2](#))

Tool	Description	Attributes	Input	Output
Topic modelling with LDA	The application of the LDA topic modelling on a given corpus.	<div>Stopwords language</div> <hr/> <div>The number of topics to generate</div> <hr/> <div>The number of words to retrieve for each topic</div>	<div>Textual document/s</div> <hr/> <div>Stopwords list</div>	<div>Documents X Topics table</div> <hr/> <div>Topics X Words table</div> <hr/> <div>LDA perplexity value</div> <hr/> <div>LDA coherence value</div>
Filter the text	Filter a given textual input, and outputs the same files given as input filtered.	<div>Filter and remove a set of predefined values such as: dates, the header, or the reference list.</div> <hr/> <div>Filter a specific textual value or regular expression that matches some values in the input corpus.</div>	<div>Textual document/s</div>	<div>Textual document/s</div>

Save the files	Gives the possibility to save and download the input files	NONE	ANY DATA TYPE	ANY DATA TYPE
Visualize document topics	Shows a graphical visualization of the topics per document table	NONE	Documents X Topics table	Chart image
Visualize topics words	Shows a graphical visualization of the words per topic table	NONE	Topics X Words table	Chart image
Convert PDF document/s to text	Converts a given PDF file/s into a text file type.	NONE	PDF document/s	Textual document/s

2.4) Building a workflow

In this section we will guide you into how to correctly define a workflow and merge the nodes we described in sections (2.3) and (2.2) using the MITAO interface features of section (2.1). Here we list an ordered and recommended procedure to follow when getting started with MITAO for building a complete workflow.

Each subsection describes a basic operation and gives a step-by-step description on how to apply such operation on MITAO. Building and running a workflow in MITAO requires a complete understanding of these operations.

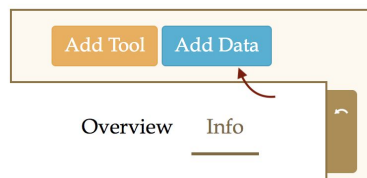
2.4.1) Add a “Data” node

In order to build a correct functional workflow we need to have at least one file containing some data to process. So first we should add a “Data” node then we should select the type of data such node represents and upload the file/s from our local machine.

Do it in MITAO

(1)

Click on the “Add Data” button ([see section 2.1](#))



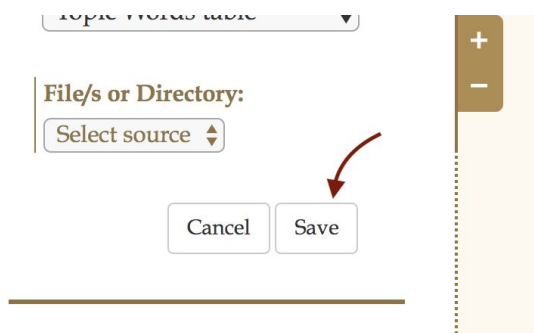
(2)

From the “The Info panel” ([see section 2.1](#)) select and define the: name, type, and files to upload.



(3)

Click on the “save” button to save the changes



(Notes)

To modify again the attributes of the node. Click on the node and redo the same

2.4.2) Add a “Tool” node

This step answers the question: what I want to do with the data nodes I have inserted in the workflow? We should select and add the right tools to the workflow according to the analysis we want to do, and the type of data we have (to take the right decision look again at [section 2.3](#)).

Do it in MITAO

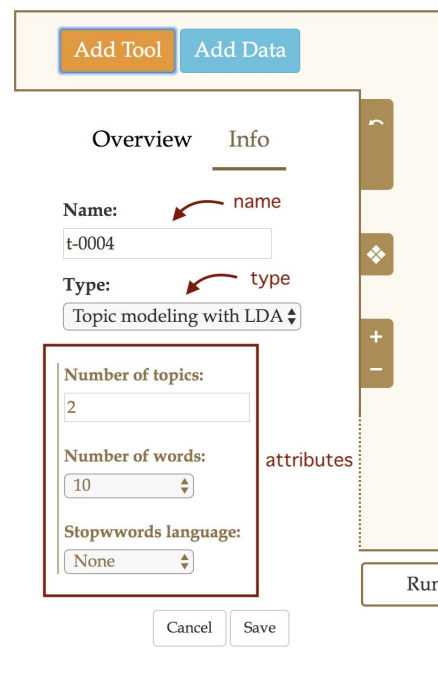
(1)

Click on the *“Add Tool”* button ([see section 2.1](#))



(2)

From the *“The Info panel”* ([see section 2.1](#)) select and define the: name, type, and related attributes.



(3)

Click on the *“save”* button to save the changes

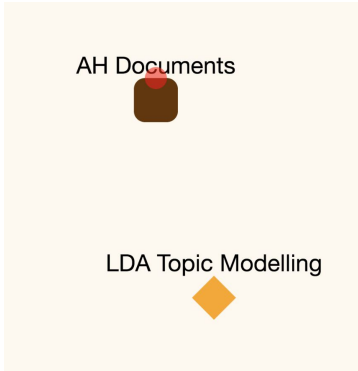
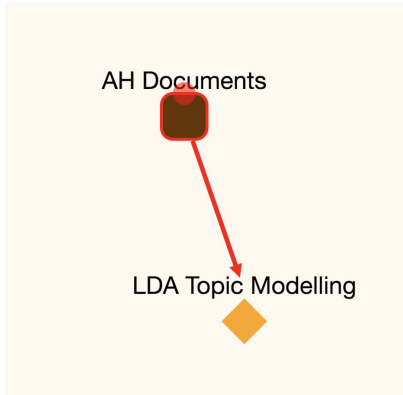


(Notes)

To modify again the attributes of the node. Click on the node and redo the same

2.4.3) Connecting the nodes

Each node inside the diagram (“Tool” or “Data”) could be connected only to its compatible nodes, which could take its value as input. E.g: a “PDF document/s” can be connected only to a “Convert PDF document/s to text” tool, since none of the other tools takes a such type of data as input ([see section 2.3](#)).

Do it in MITAO		
(1)	(2)	(Notes)
<p>Move the cursor on the node you want to connect</p> 	<p>Move the mouse pointer to the displayed red circle above of it, then hold down the left click of the mouse and drag the mouse pointer to the destination node(MITAO disables the non compatible nodes).</p> 	<p>In case you want to remove an edge, click on the edge you want to removed and click on the button “Remove element” from the “Info panel”</p>

2.4.4) Running the workflow

Once you have built your workflow and you want to run it you should click on “Run Workflow” ([see section 2.1](#)). The processing time depends on the complexity of the workflow defined, while processing a timeline block will appear on the right of the “Run Workflow” button. In case the workflow includes “Save the Files”, “Visualize document topics”, or “Visualize topics words” tools typology, the timeline will also generate some links to get access at the produced results.

3) Examples

In this section we introduce some real use cases and how to build a workflow to handle them using MITAO.

3.1) LDA topic modelling on textual documents

In this example we want to build a workflow which takes in input 10 .TXT files and applies an LDA topic modelling algorithm to them. We want to visualize the final topics and save the results locally on our machine. To do so we will go through the following steps:

- 1) Add a “Data” node and set the following attributes ([see section 2.4.1](#)):
 - a) Name: “General Documents”
 - b) Type: “Textual document/s”
 - c) File/s or Directory: a collection of 10 .TXT files from our machine.
- 2) Add a “Tool” node and set the following attributes ([see section 2.4.2](#)):
 - a) Name: “LDA Topic Modelling”
 - b) Type: “Topic modelling with LDA”
 - c) Stopwords language: “English”
 - d) Number of words: “25”
 - e) Number of topics: “10”
- 3) Add a “Tool” node and set the following attributes ([see section 2.4.2](#)):
 - a) Name: “Visualize Topics”
 - b) Type: “Visualize topics words”
- 4) Add a “Tool” node and set the following attributes ([see section 2.4.2](#)):
 - a) Name: “Save LDA”
 - b) Type: “Save files”
- 5) Run the built workflow ([see section 2.4.4](#))

