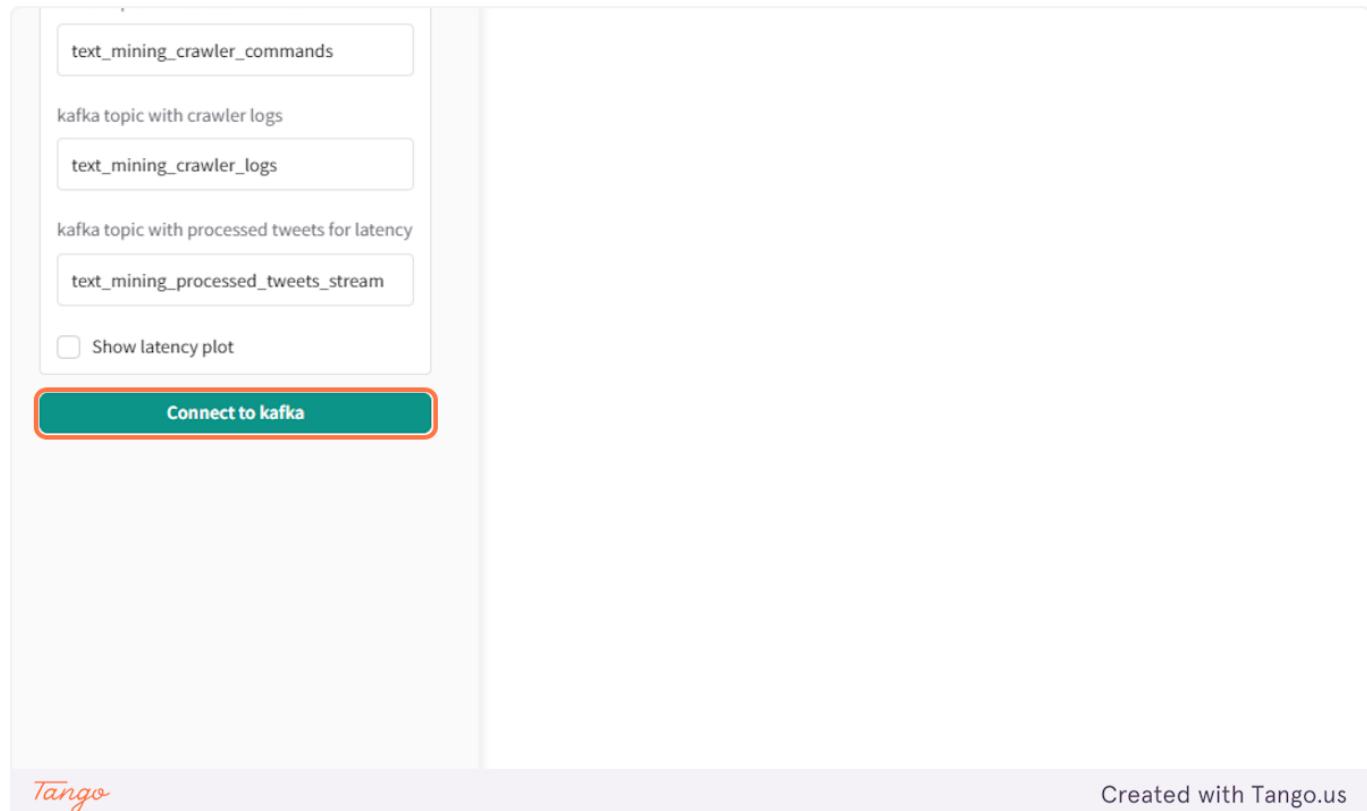


How to Start and Stop a Crawl in the text mining UI

This guide provides a step-by-step process to start and stop a crawl using the text mining UI.

1. Navigate to the link for the text mining app ex:<http://localhost:7860/>
2. Follow steps in setup kafka cluster documentation, then click "Connect to kafka".



3. Navigate to the "Crawler UI" tab.

Crawl Identifier
Unique identifier to manage the crawl. Ex: flood_dortmund_20250611

X username for crawling
account username
TCrexdata

X email for crawling
account email
text.mining.crexdata@gmail.com

Tango Created with Tango.us

4. Click the "Crawl Identifier" field, and type in a unique identify to manage the crawl. ex: flood__.

Crawl Identifier
Unique identifier to manage the crawl. Ex: flood_dortmund_20250611
flood_testplace_testdate

X password for crawling
account password

Gmail app access password
Should be the app pass for the x email, which should be gmail

Crawl language
Only those supported by X
German

Start Crawl

Kill Crawl Instance

Crawl Identifier to kill.
Ex: flood_dortmund_20250611. Ensure the identifier is of a running crawl

Tango Created with Tango.us

5. Fill in the other details, include the X login details, language of posts, lon, lat, and radius of area to crawl from, then click "Start Crawl".

Start Crawl Instance

Crawl Identifier
Unique identifier to manage the crawl. Ex: Rood_dortmund_20250611
`Rood_testplace_testdata`

X password for crawling
account password
XXXXXXXXXX

Gmail app access password
Should be the app pass for the email, which should be gmail
XXXXXXXXXX

Lat of location to crawl from
in Decimal
51.513889

Lon of location to crawl from
in Decimal
7.465278

Radius to crawl from
in Km
100

Crawl language
Only those supported by X
German

Start Crawl

Kill Crawl Instance

Crawl Identifier to kill.
Ex: Rood_dortmund_20250611. Ensure the identifier is of a running crawl.

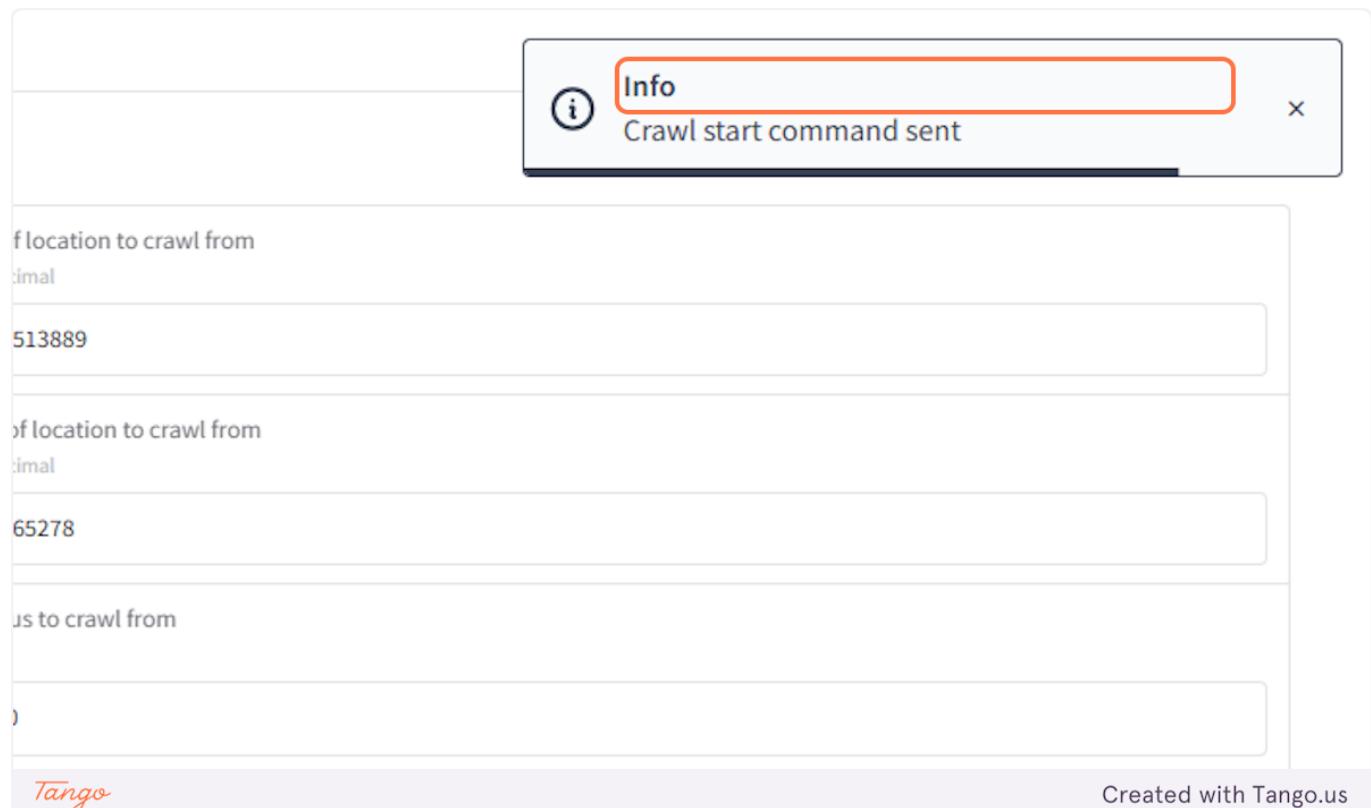
Kill Crawl

Tango

Created with Tango.us

i Note: We advise using a Gmail account to create an X account for crawling, and also setting up an app password. With these steps:
 1. create Gmail account
 2. visit google security settings to turn on 2fa - <https://myaccount.google.com/security>
 3. add mobile number for 2fa
 4. visit <https://myaccount.google.com/apppasswords>
 5. create new app to access account
 6. copy app password

6. Click on Info...A notification will show that the crawl command has been sent.

*Tango*

Created with Tango.us

7. Now you can navigate to "DEBUG/LOG Console" tab to see the crawler logs and

commands.

Crawler UI Revelance UI Question Answering UI **DEBUG/LOG Console**

Start Crawl Instance

Crawl Identifier
Unique identifier to manage the crawl. Ex: flood_dortmund_20250611
flood_testplace_testdate

X password for cra
account password
.....

X username for crawling
account username
TCrexdata

Gmail app access p
Should be the app pa
.....

X email for crawling
account email
text.mining.crexdata@gmail.com

Crawl language
Only those supported
German

Kill Crawl Instance

Tango Created with Tango.us

8. In the Crawler log console, there are two consoles, the left is for crawler commands.

You can set length of latest commands returned using the log tail field.

Kafka Topic Status

Enter topic:

Number of Messages in Topic
NaN

Timestamp of Last Message
N/A

(-) Last Message {..}

Crawler Log console

Settings

Log tail
20

Refresh

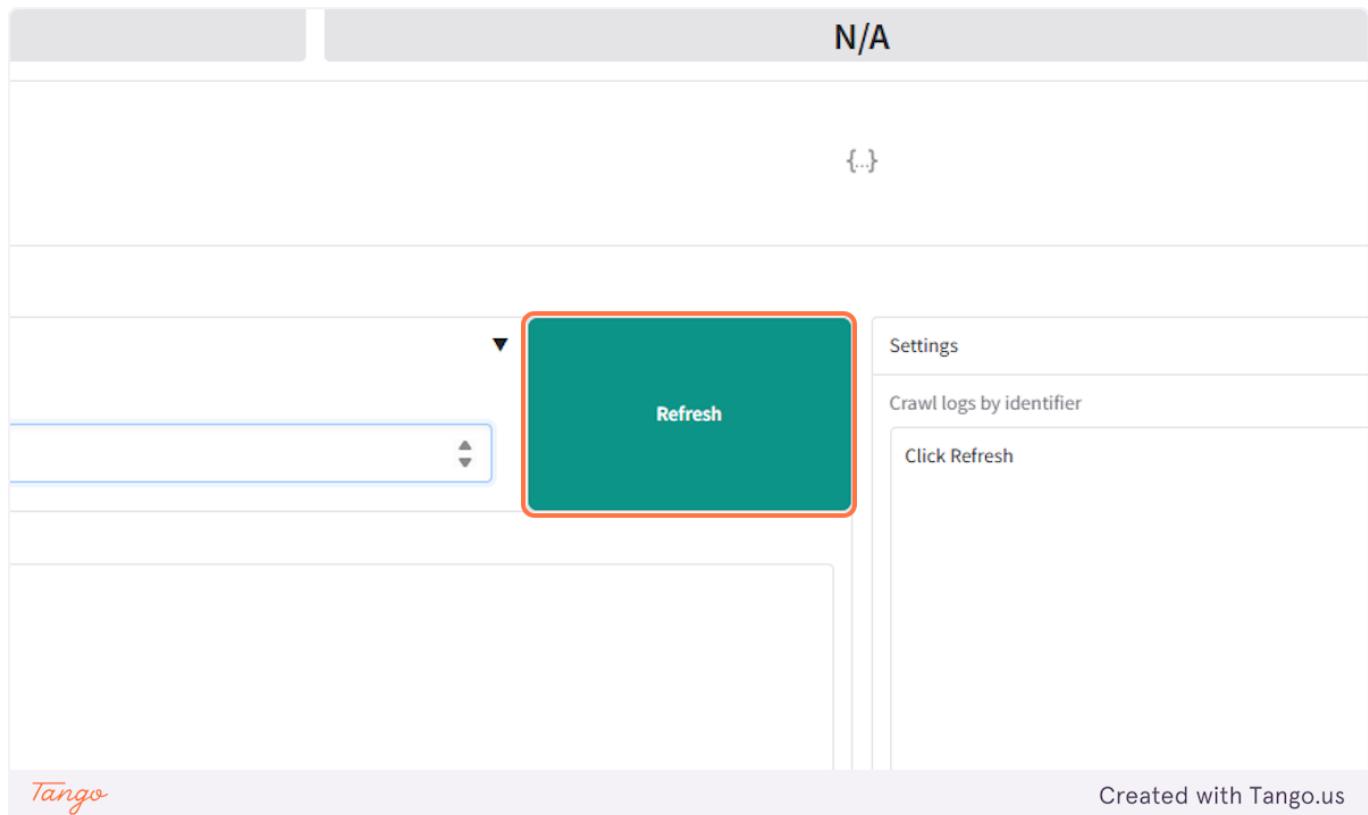
Running crawls logs
Click Refresh

Settings

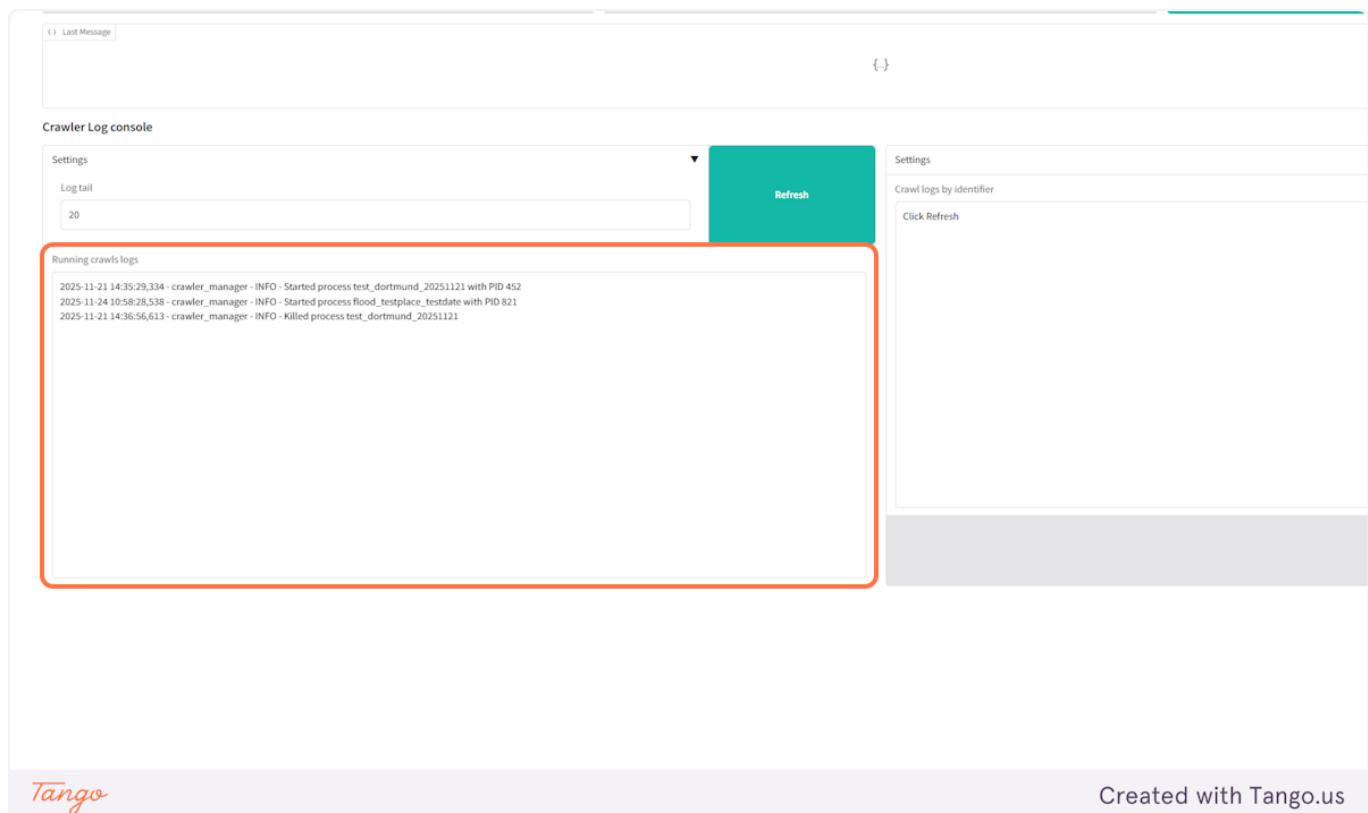
Crawl logs by identifier
Click Refresh

Tango Created with Tango.us

9. Then click "Refresh" to view logs.



10. The crawl commands are show in the text console below.



11. For the right console, you can see the crawler logs.

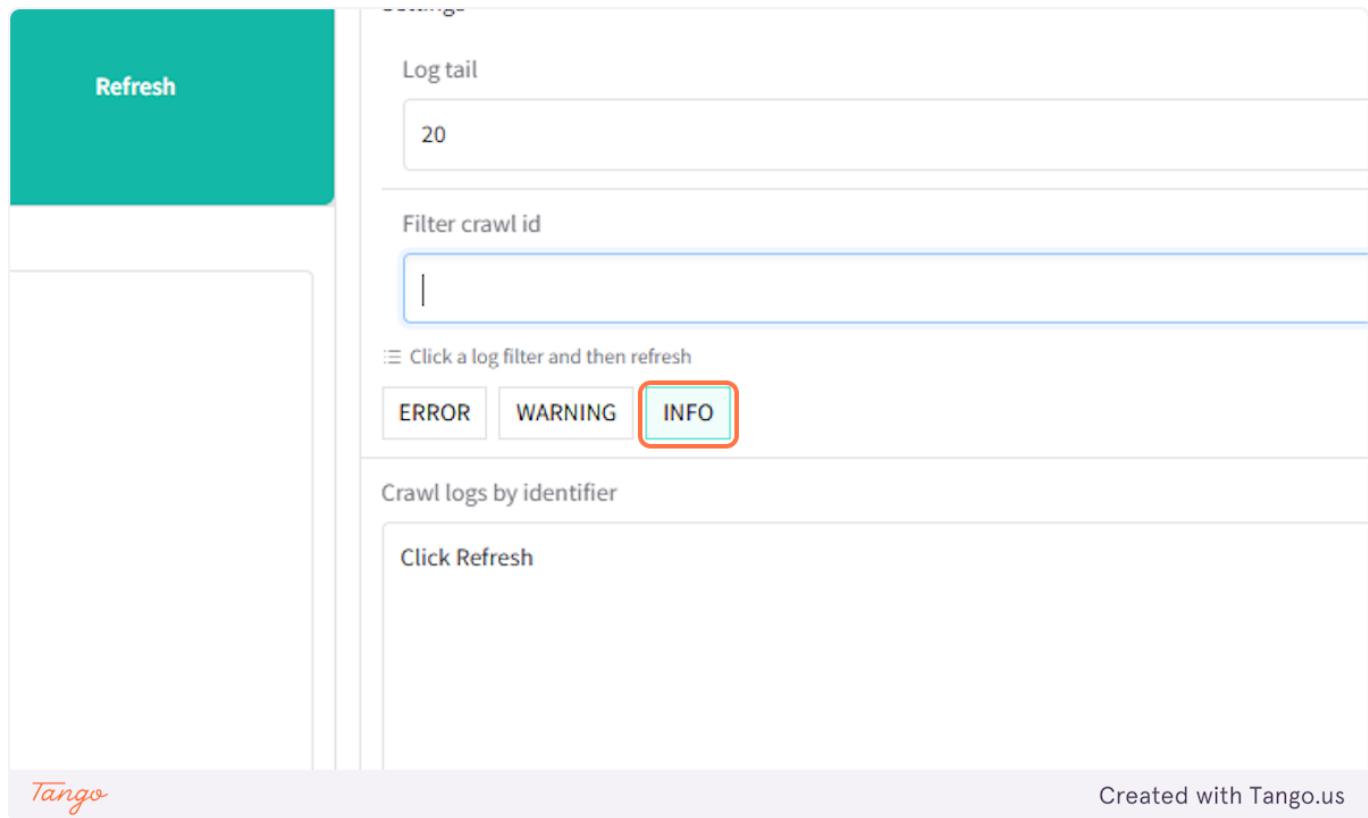
The screenshot shows a user interface for managing crawl logs. At the top, there's a grey header bar with the text "Timestamp of Last Message" and "N/A". To its right is a teal button labeled "Refresh". Below this is a large white area containing the text "{...}" and a redacted log entry. On the left side, there's a teal sidebar with the word "Refresh" and a "Tango" logo at the bottom. In the center, there's a "Settings" button with a red border. Below it is a section titled "Crawl logs by identifier" with a "Click Refresh" button. The footer of the page includes the "Tango" logo and the text "Created with Tango.us".

12. You can also set the log tail and further filter the logs by, ERROR logs, WARNING logs, or INFO logs by clicking on the corresponding button to populate the filter field.

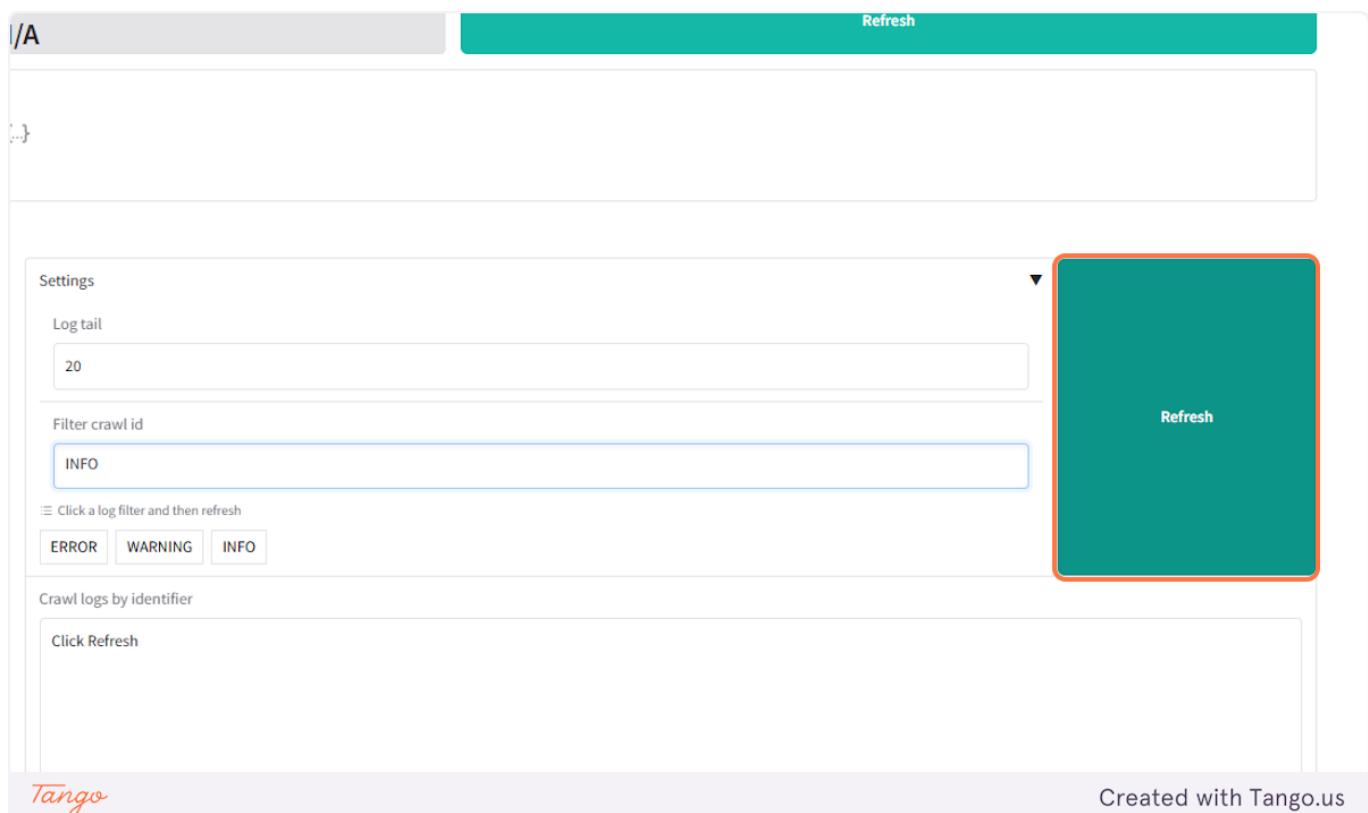
Note: you can also enter the unique crawl id used to start the crawl to view logs specific to that crawl.

This screenshot shows the same interface as the previous one, but with more detailed settings. The "Settings" button is still highlighted with a red box. Below it, the "Log tail" input field contains the value "20", which is also highlighted with a red box. There are three buttons below the log tail input: "ERROR", "WARNING", and "INFO", with "INFO" being the selected option. The "Filter crawl id" input field is empty. The "Crawl logs by identifier" section has a "Click Refresh" button. The footer includes the "Tango" logo and "Created with Tango.us".

13. You can select "INFO" or any other predefined filter.



14. The click "Refresh" to view the logs.



15. The logs are displayed in the text console below.

The screenshot shows a log viewer interface with a teal header bar. On the left, there's a sidebar with the word 'esh'. The main area has a 'Settings' section with a 'Log tail' input set to '20'. Below it is a 'Filter crawl id' input field containing 'INFO', which is highlighted with a red border. A note below says 'Click a log filter and then refresh'. There are three buttons: 'ERROR', 'WARNING', and 'INFO'. Underneath is a section titled 'Crawl logs by identifier' containing several log entries. At the bottom right, it says 'Created with Tango.us'.

16. You can also type in the unique crawl ID used to start the crawl in this filter field, so you see only the logs related to that crawl.

The screenshot shows a log viewer interface with a teal header bar. On the left, there's a sidebar with the word 'esh'. The main area has a 'Settings' section with a 'Log tail' input set to '20'. Below it is a 'Filter crawl id' input field containing a single vertical bar '|', which is highlighted with a red border. A note below says 'Click a log filter and then refresh'. There are three buttons: 'ERROR', 'WARNING', and 'INFO'. Underneath is a section titled 'Crawl logs by identifier' containing several log entries. At the bottom right, it says 'Created with Tango.us'.

17. Finally, you can keep changing the filter and clicking "Refresh" as many times as you need.

The screenshot shows the 'Crawl logs by identifier' section. A red box highlights the 'INFO' filter button. To the right, a large green box highlights the 'Refresh' button.

Settings

Log tail
20

Filter crawl id
INFO

Click a log filter and then refresh
ERROR WARNING INFO

Crawl logs by identifier

```
2025-11-24 10:58:46,895 - crawler_kafka - INFO - flood_testplace_testdate - Logging in
2025-11-21 14:35:45,691 - crawler_kafka - INFO - test_dortmund_20251121 - Logging in
2025-11-21 14:36:31,150 - crawler_kafka - INFO - test_dortmund_20251121 - Fetching and storing tweets
2025-11-21 14:36:56,459 - crawler_kafka - INFO - test_dortmund_20251121 - Received SIGTERM in crawler_kafka.py. Cleaning up...
2025-11-24 10:58:28,538 - crawler_manager - INFO - Started process flood_testplace_testdate with PID 821
2025-11-24 10:58:30,704 - crawler_kafka - INFO - flood_testplace_testdate - Driver initialized
```

Tango Created with Tango.us

18. To kill crawl, navigate back to the "Crawl UI" and enter the unique id for a running crawl in the "Kill Crawler Instance" section.

The screenshot shows the 'Kill Crawl Instance' section. A red box highlights the 'Kill Crawl' button.

Start Crawl Instance

Crawl Identifier Unique identifier to manage the crawl, Ex: flood_dortmund_20250611 flood_testplace_testdate	X password for crawling account password *****	Lat of location to crawl from in Decimal 51.513889
X username for crawling account username TCredata	Gmail app access password Should be the app pass for the email, which should be gmail *****	Lon of location to crawl from in Decimal 7.465278
X email for crawling account email test.mining.credata@gmail.com	Crawl language Only those supported by X German	Radius to crawl from in Km 100

Kill Crawl Instance

Crawl Identifier to kill. This will stop the crawl. Enter the identifier to kill a running crawl. flood_testplace_testdate	Kill Crawl
---	-------------------

Tango Created with Tango.us

19. Then click on "Kill Crawl".

Radius to crawl from
in Km

100

Kill Crawl

Tango

Created with Tango.us

20. A final notification will pop up showing the kill command has been sent.

Info

Crawl kill command sent

f location to crawl from
:imal

513889

o location to crawl from
:imal

65278

us to crawl from

)

Tango

Created with Tango.us