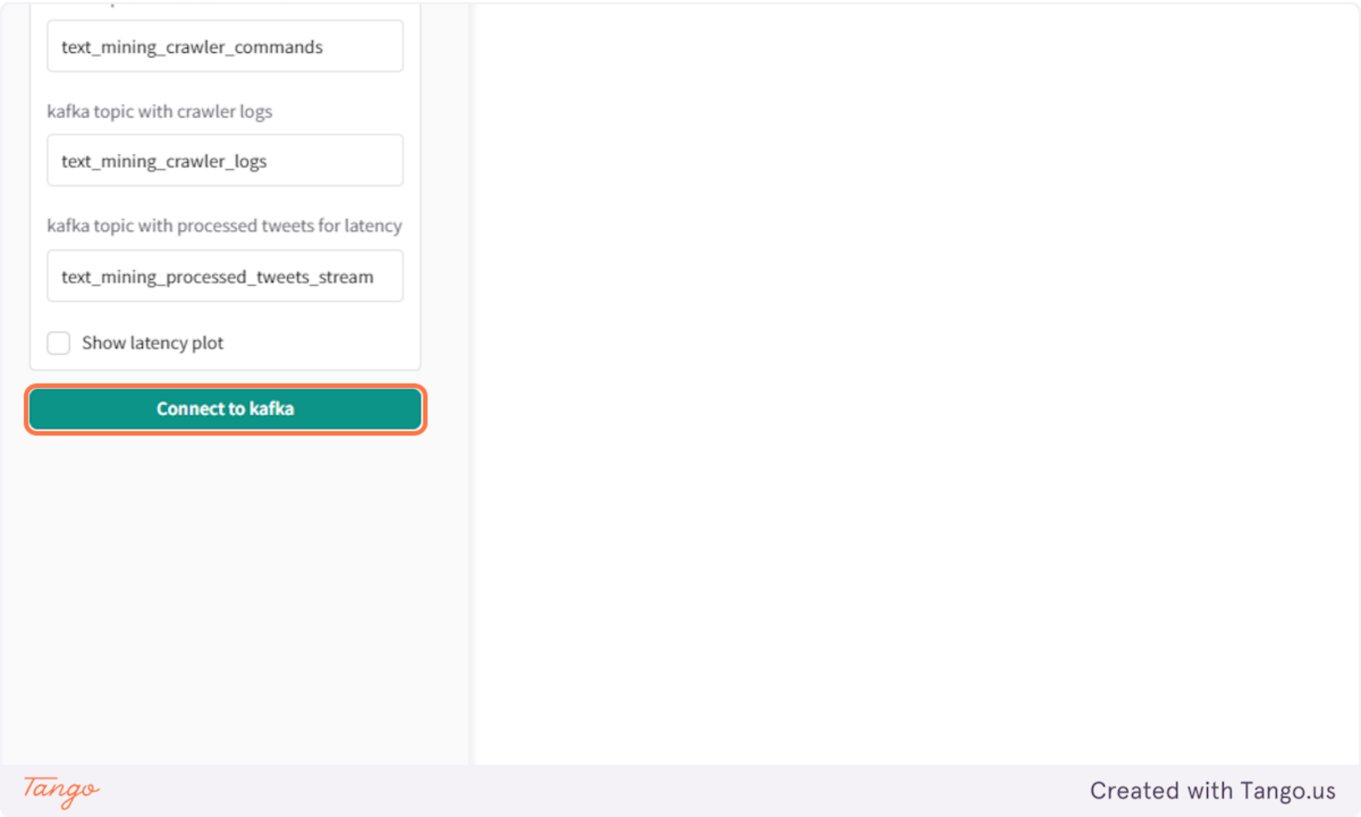
How to Start and Stop a Crawl in the text mining UI

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".



text_mining_crawler_commands

kafka topic with crawler logs

text_mining_crawler_logs

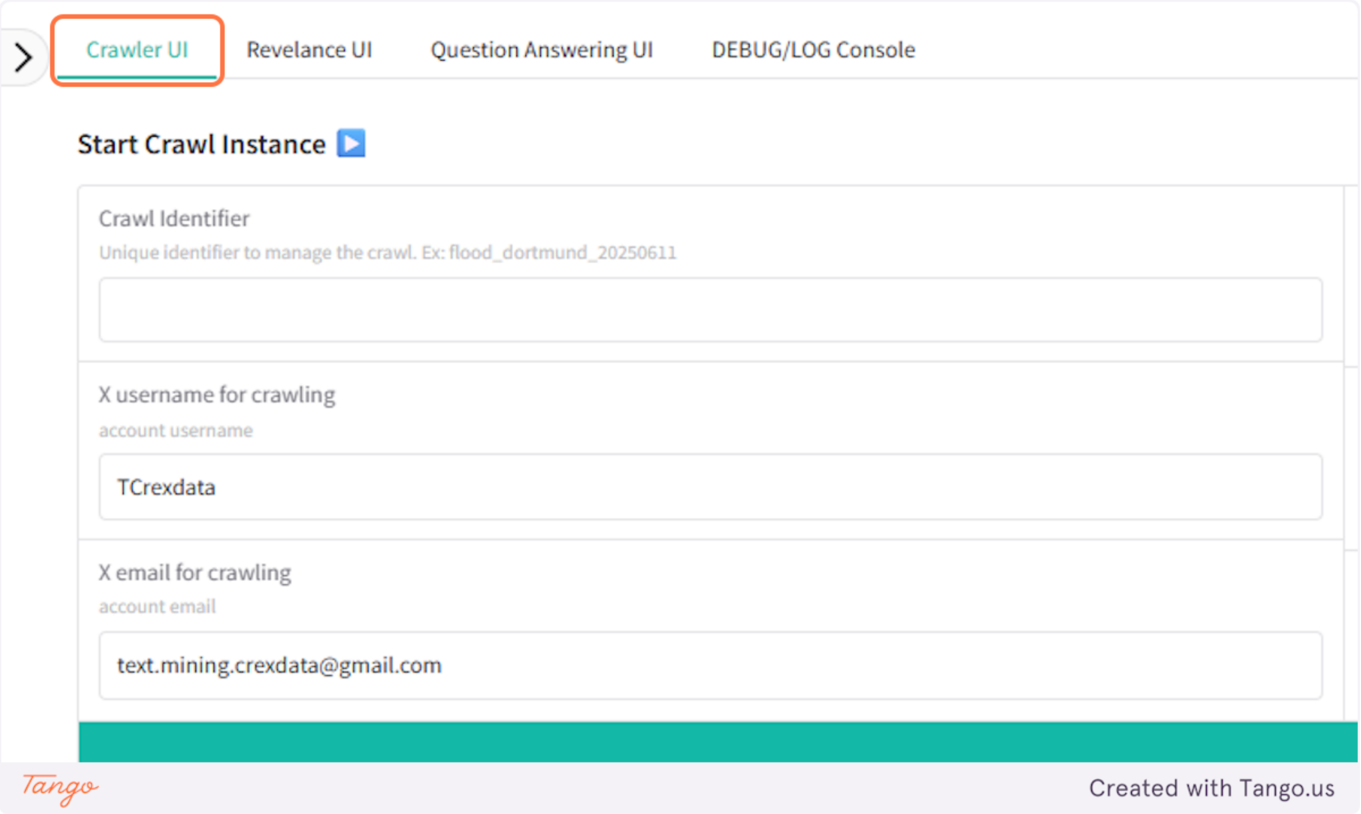
kafka topic with processed tweets for latency

text_mining_processed_tweets_stream

□ Show latency plot

Created with Tango.us

3. Navigate to the "Crawler UI" tab.



Start Crawl Instance ▶

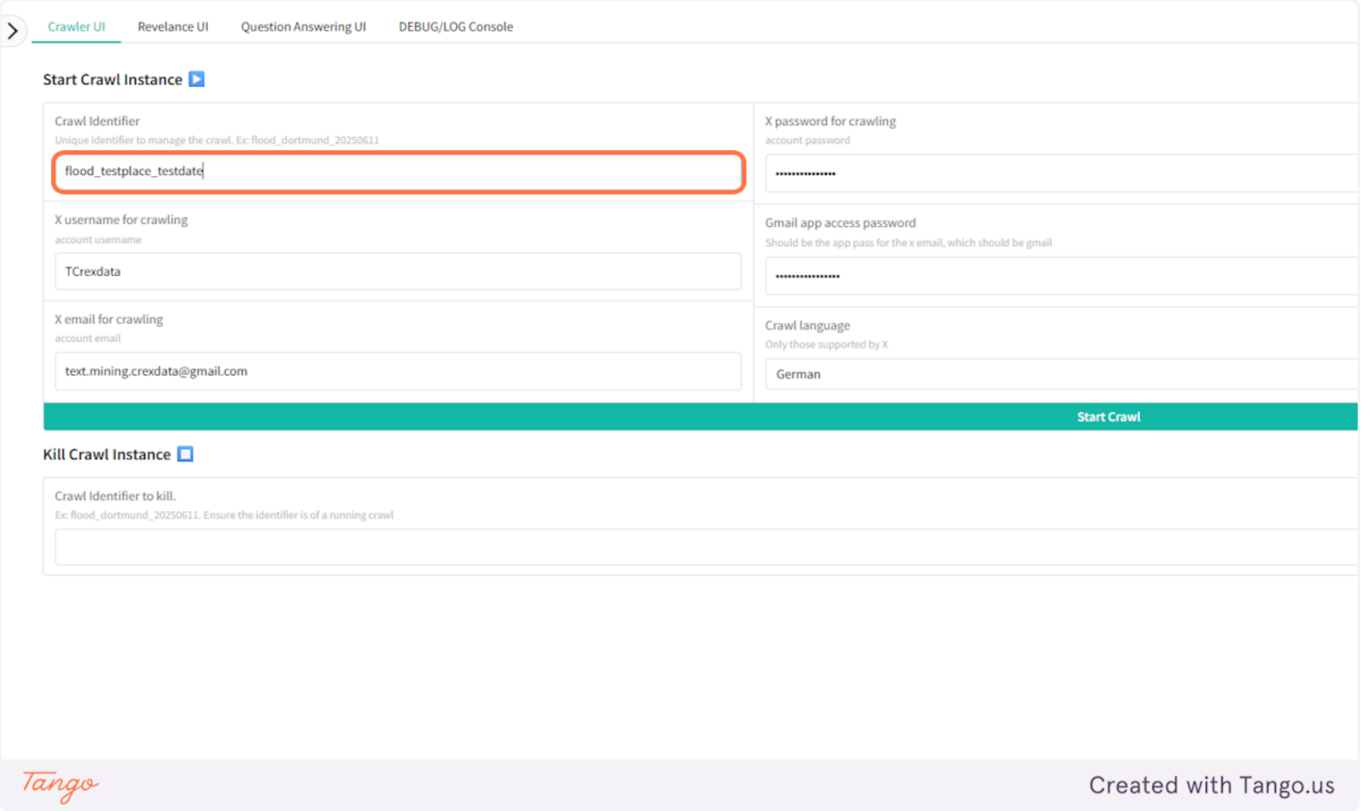
Crawl Identifier
Unique identifier to manage the crawl. Ex: flood_dortmund_20250611

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__.



Start Crawl Instance ▶

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

X username for crawling
account username
TCrexdata

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

X password for crawling
account password
•••••••••••••••

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

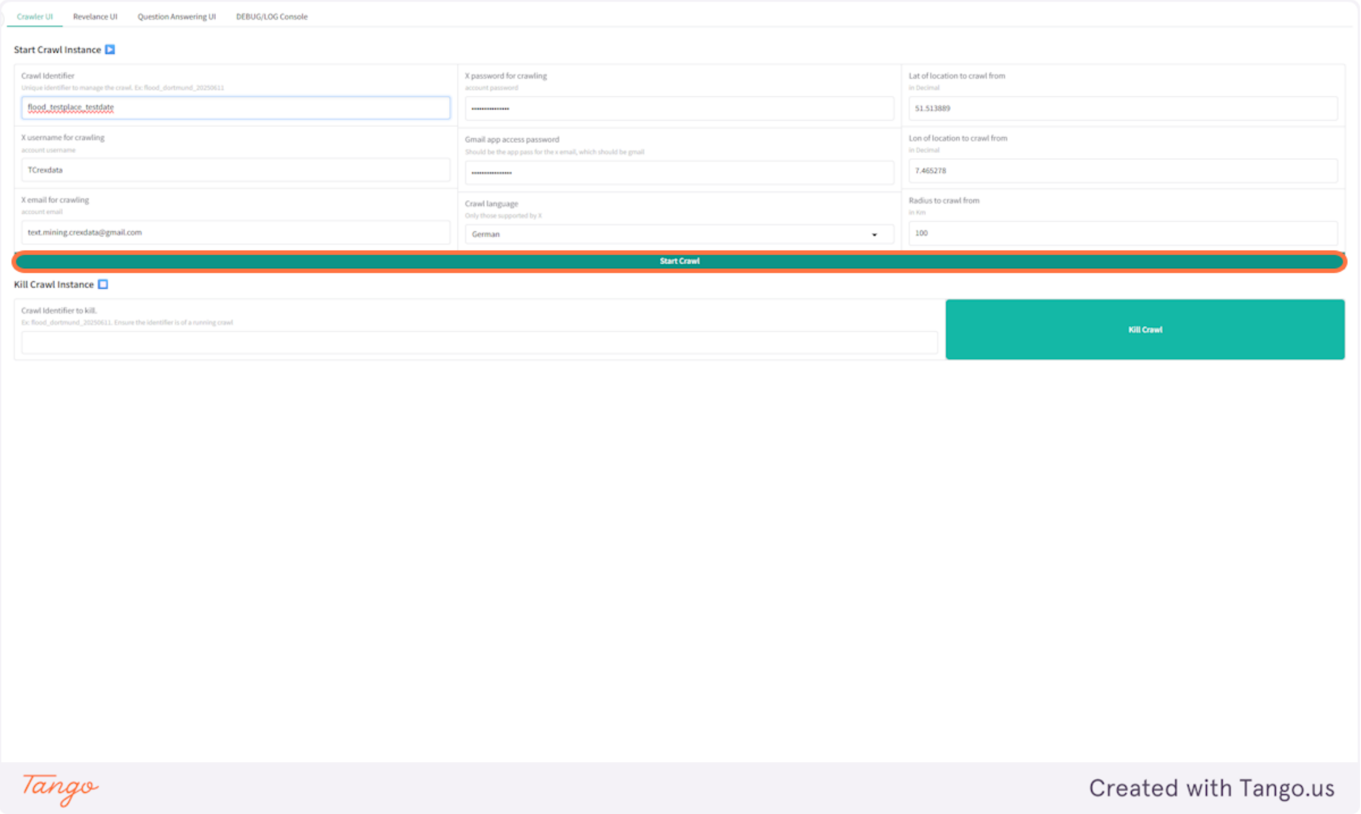
Crawl language
Only those supported by X
German

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: 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
TCrexdata

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

Lon of location to crawl from
in Decimal
7.465278

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

Crawl language
Only those supported by X
German

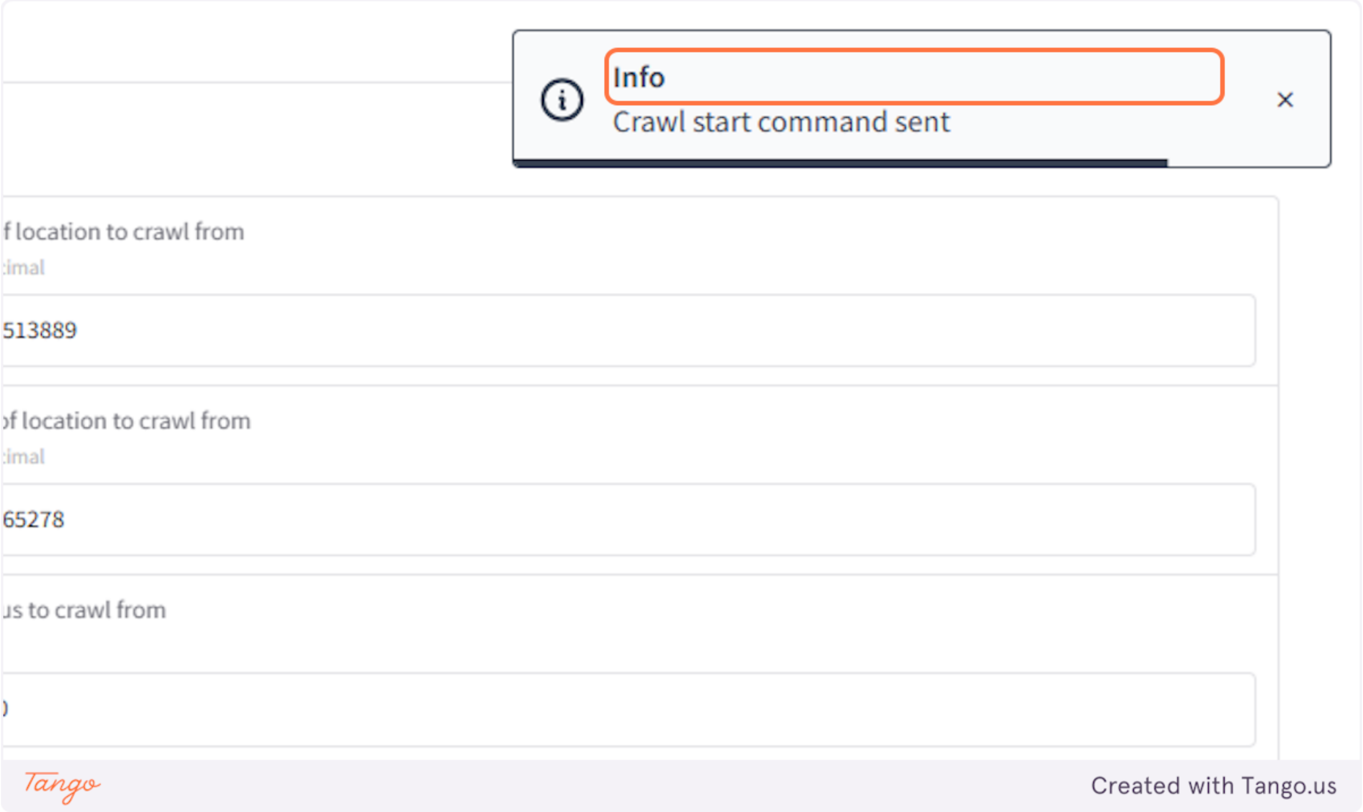
Radius to crawl from
in Km
100

Kill Crawl Instance

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

ℹ Note: We advise using a Gmail account to create an X account for crawling, andalso setting up an app passwords. With these steps:1. create Gmail account2. visit google security settings to turn on 2fa - https://myaccount.google.com/security3. add mobile number for 2fa4. visit https://myaccount.google.com/apppasswords5. create new app to access account6. copy app password

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



Info
Crawl start command sent

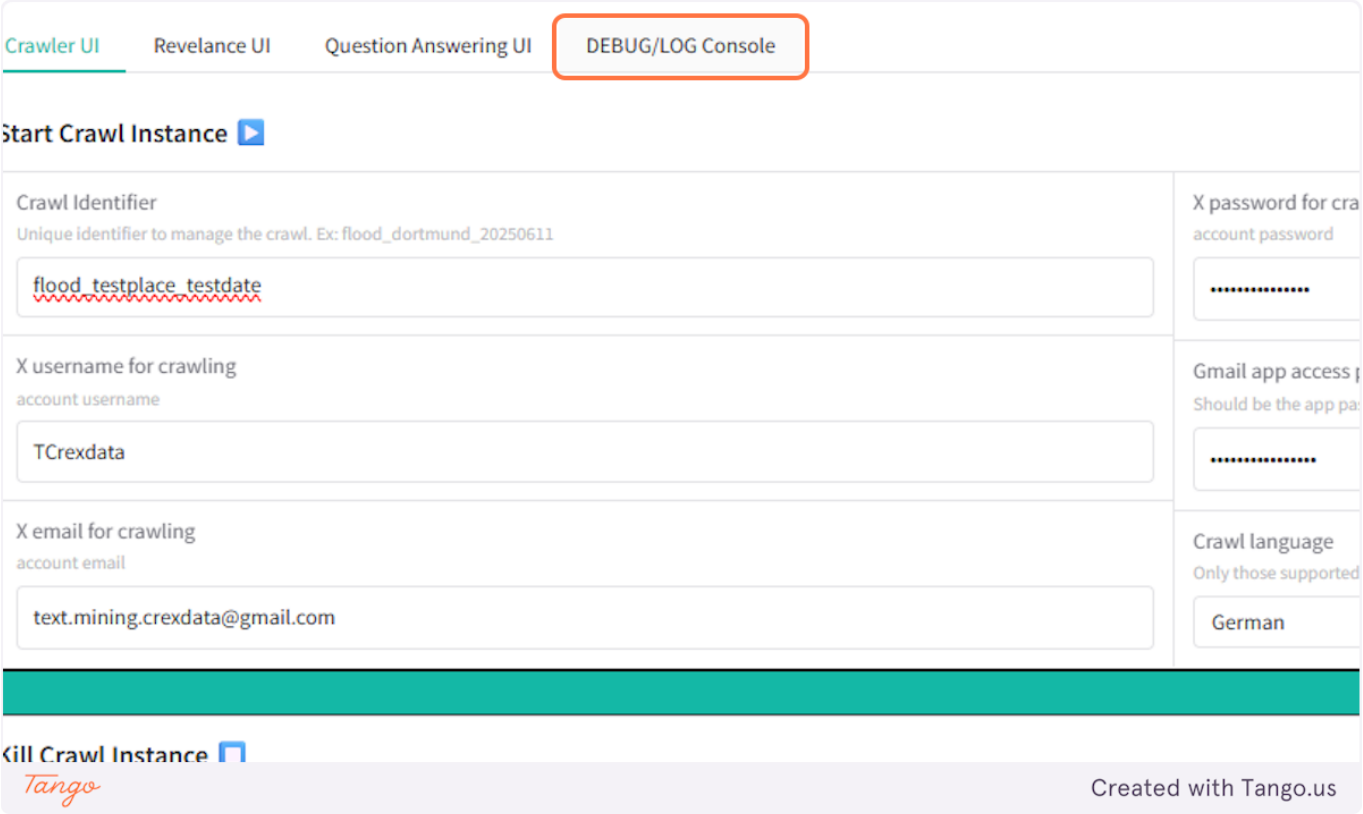
f location to crawl from
:imal
513889

of location to crawl from
:imal
65278

us to crawl from
)

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

commands.



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

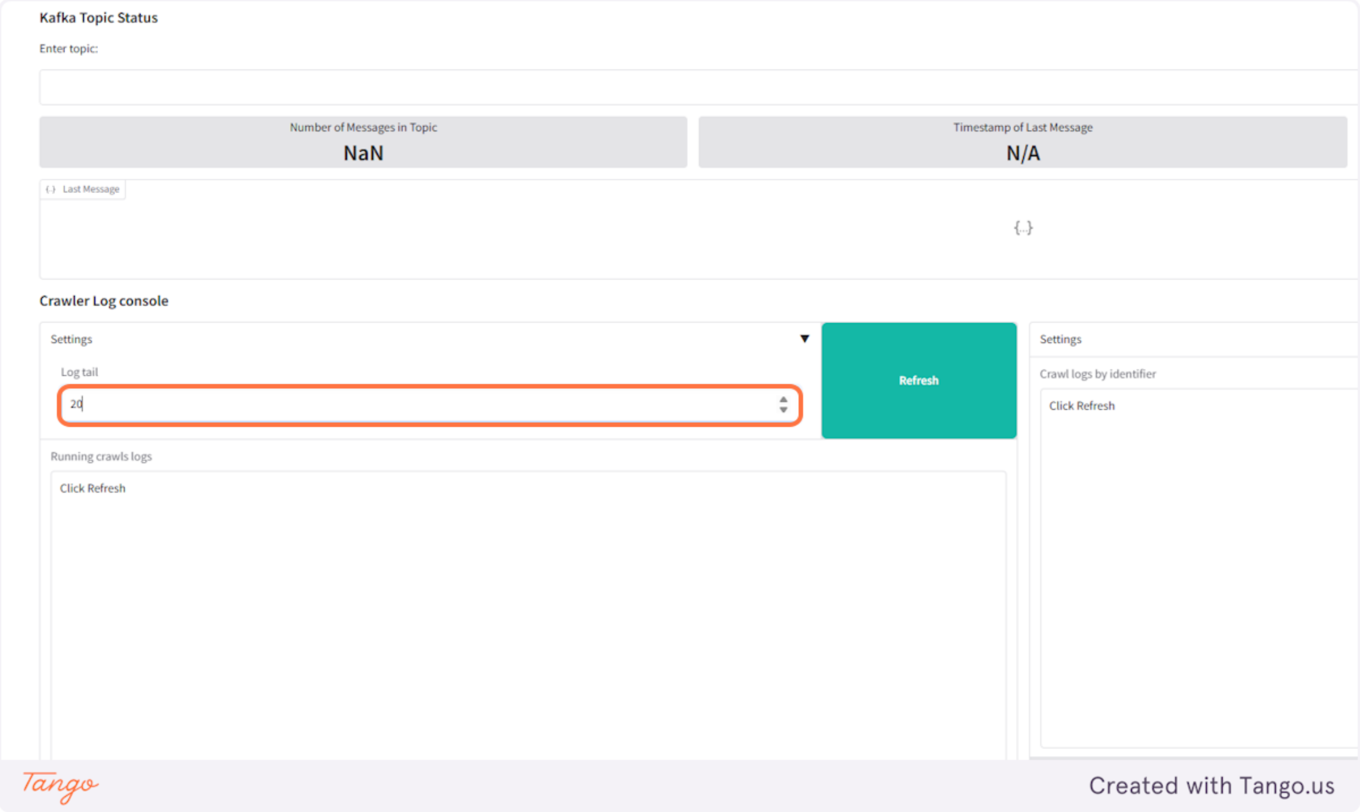
Crawl language
Only those supported
German

Kill Crawl Instance

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
{..}

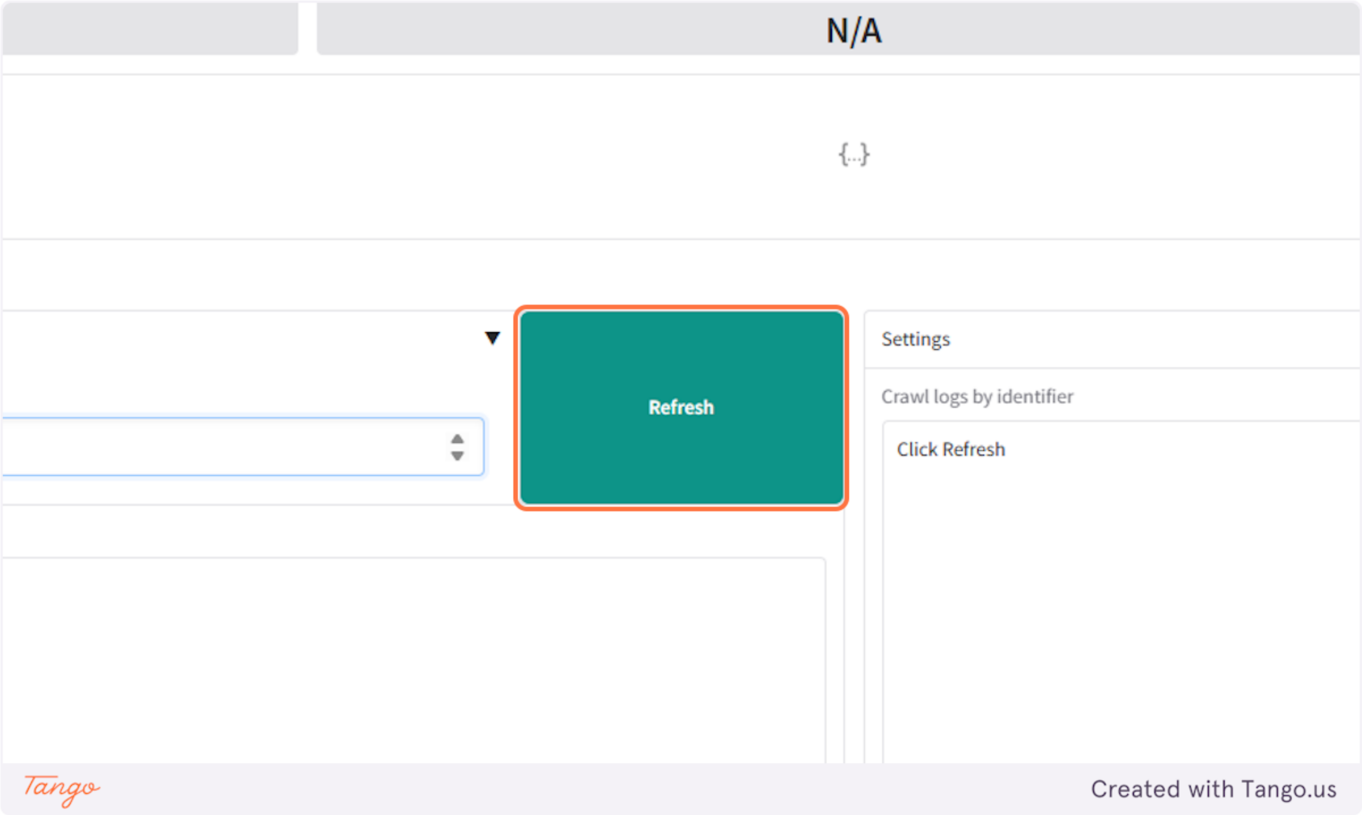
Crawler Log console

Settings
Log tail
20

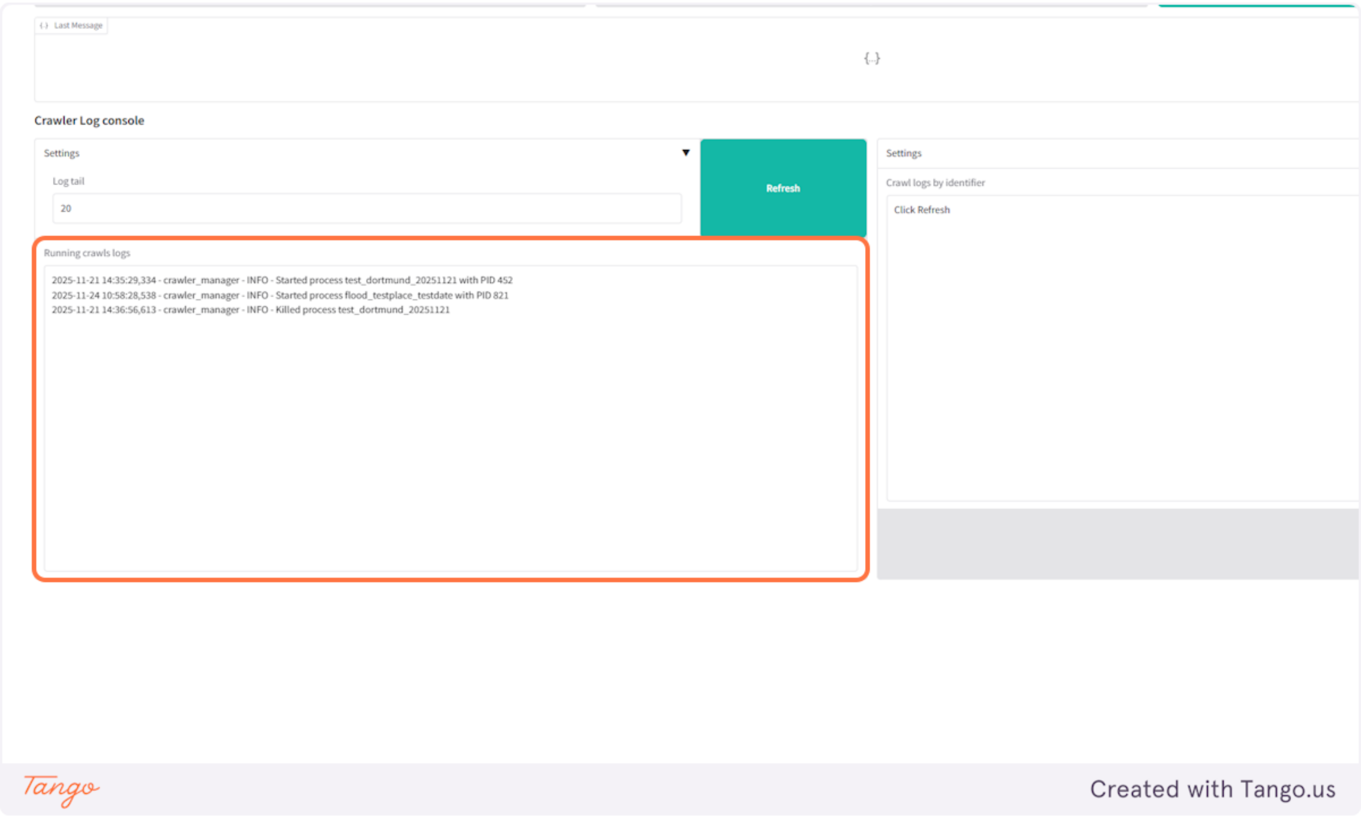
Running crawls logs
Click Refresh

Settings
Crawl logs by identifier
Click Refresh

9. Then click "Refresh" to view logs.



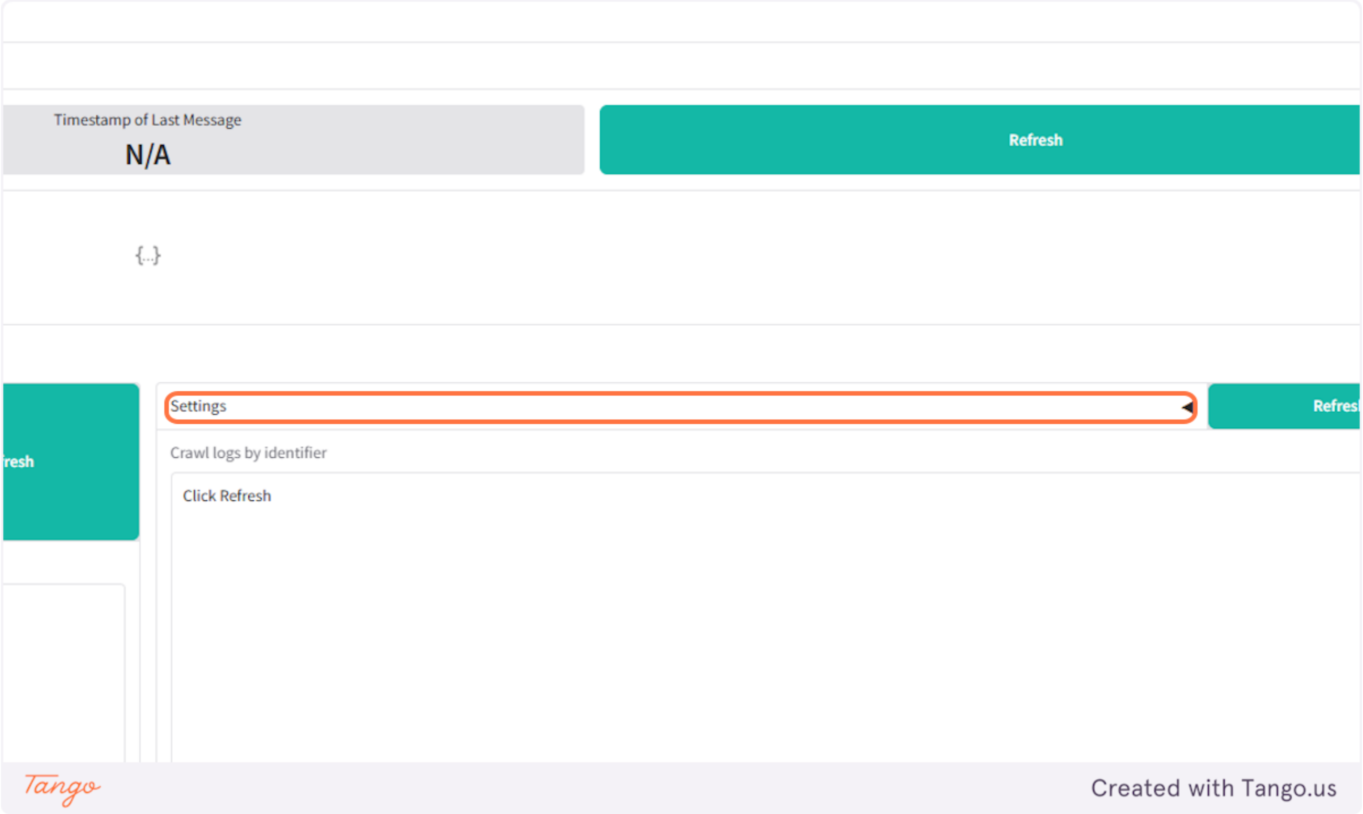
10. The crawl commands are show in the text console below.



Log tail: 20

```
2025-11-21 14:35:29,334 - crawler_manager - INFO - Started process test_dortmund_20251121 with PID 452
2025-11-24 10:58:28,538 - crawler_manager - INFO - Started process flood_testplace_testdate with PID 821
2025-11-21 14:36:56,613 - crawler_manager - INFO - Killed process test_dortmund_20251121
```

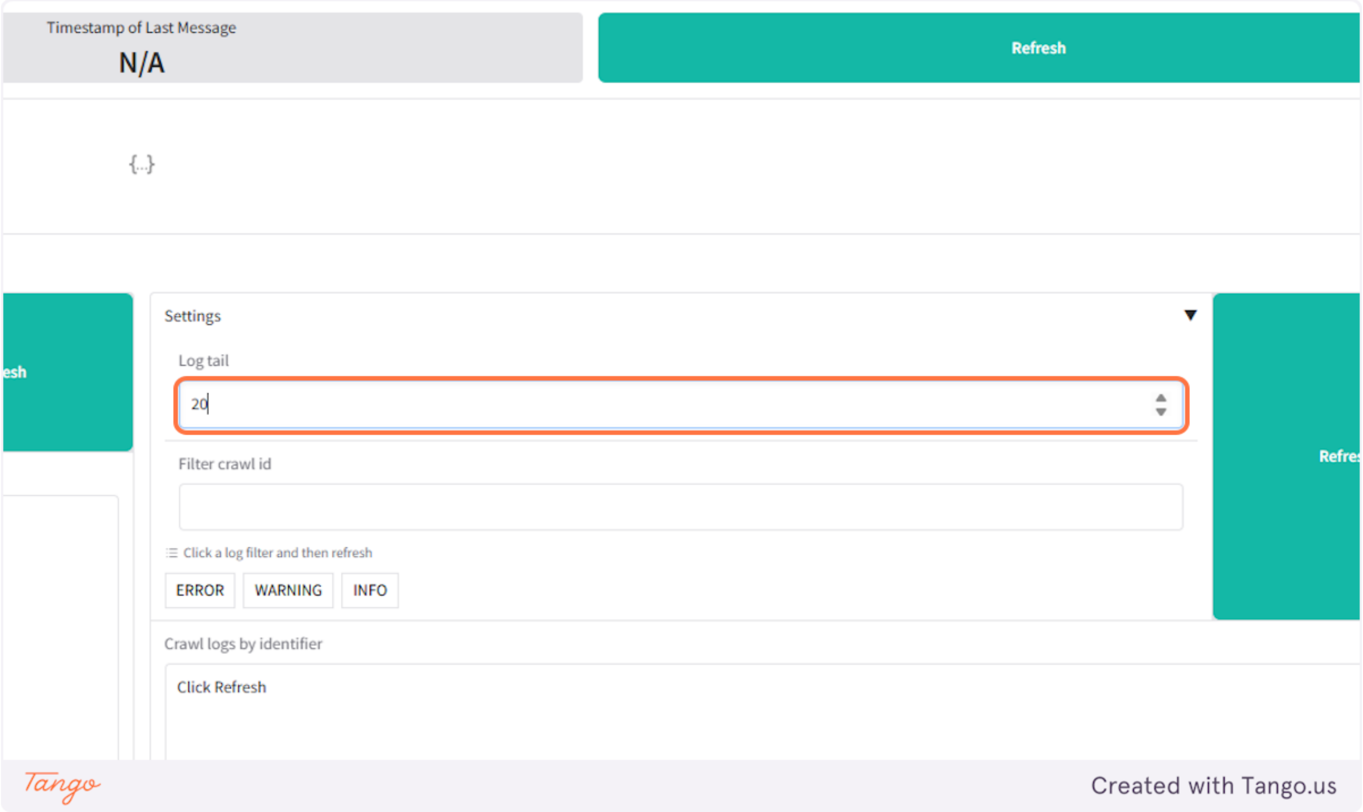
11. For the right console, you can see the crawler logs.



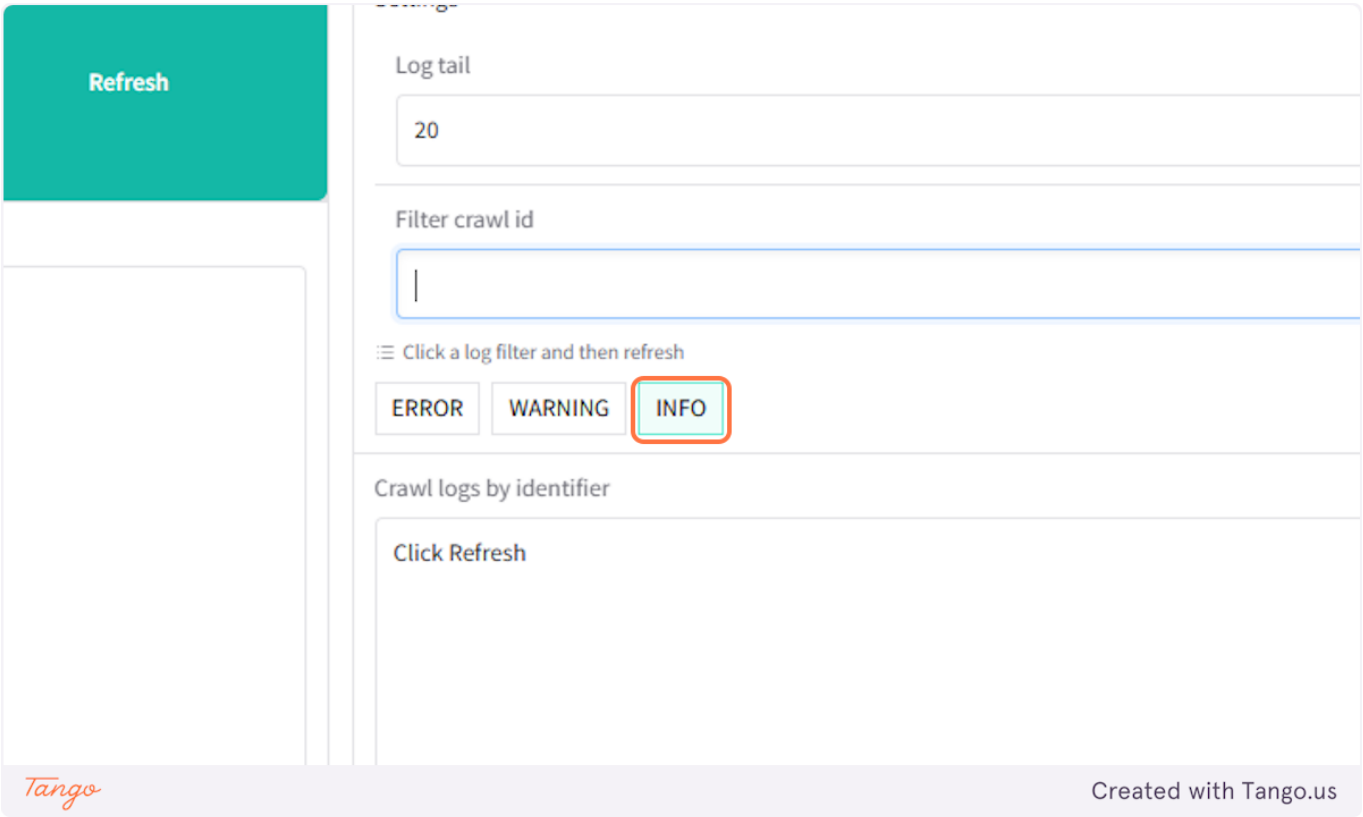
12. You can also set the log tail and further filter the logs by, ERROR logs, WARNING

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.



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



Log tail

20

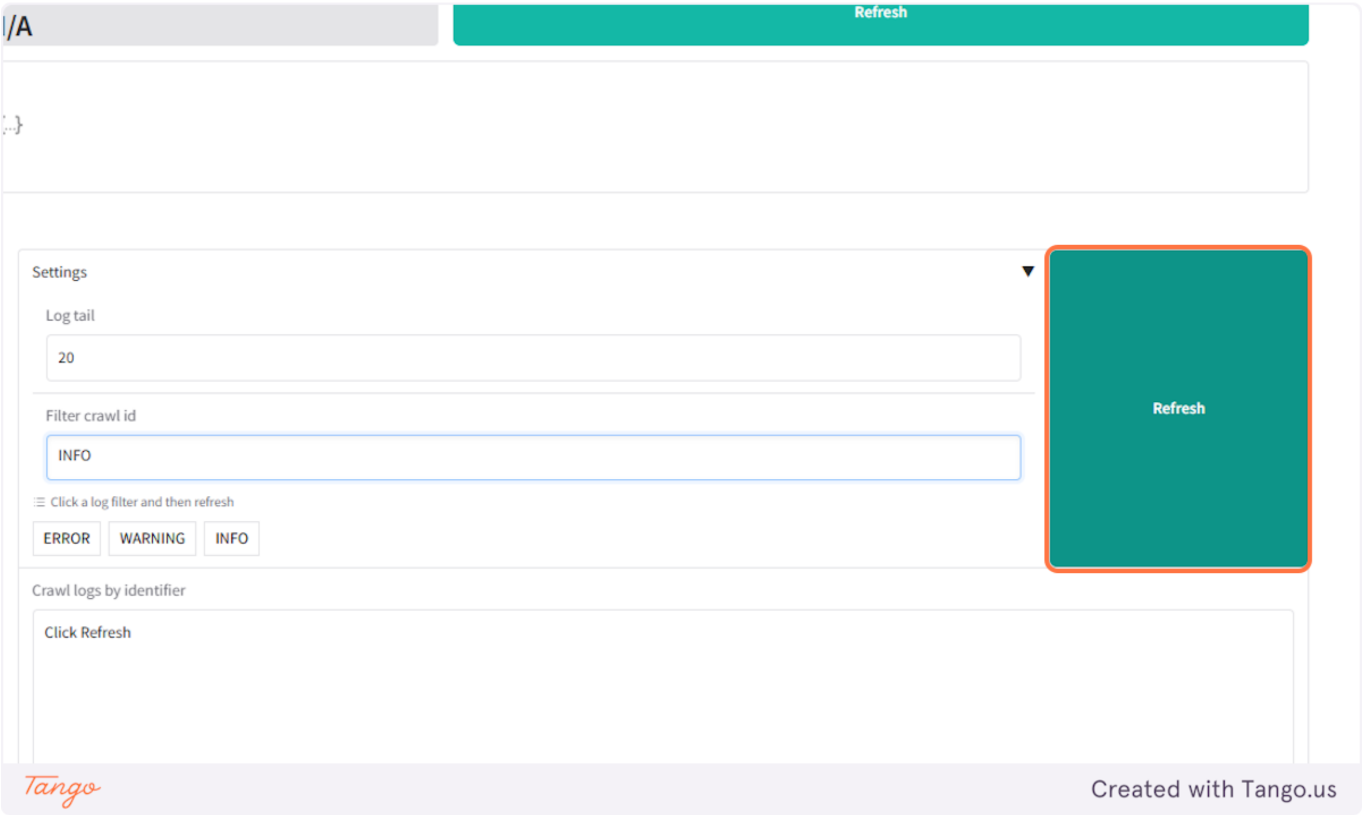
Filter crawl id

Click a log filter and then refresh

Crawl logs by identifier

Click Refresh

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



Settings

Log tail

20

Filter crawl id

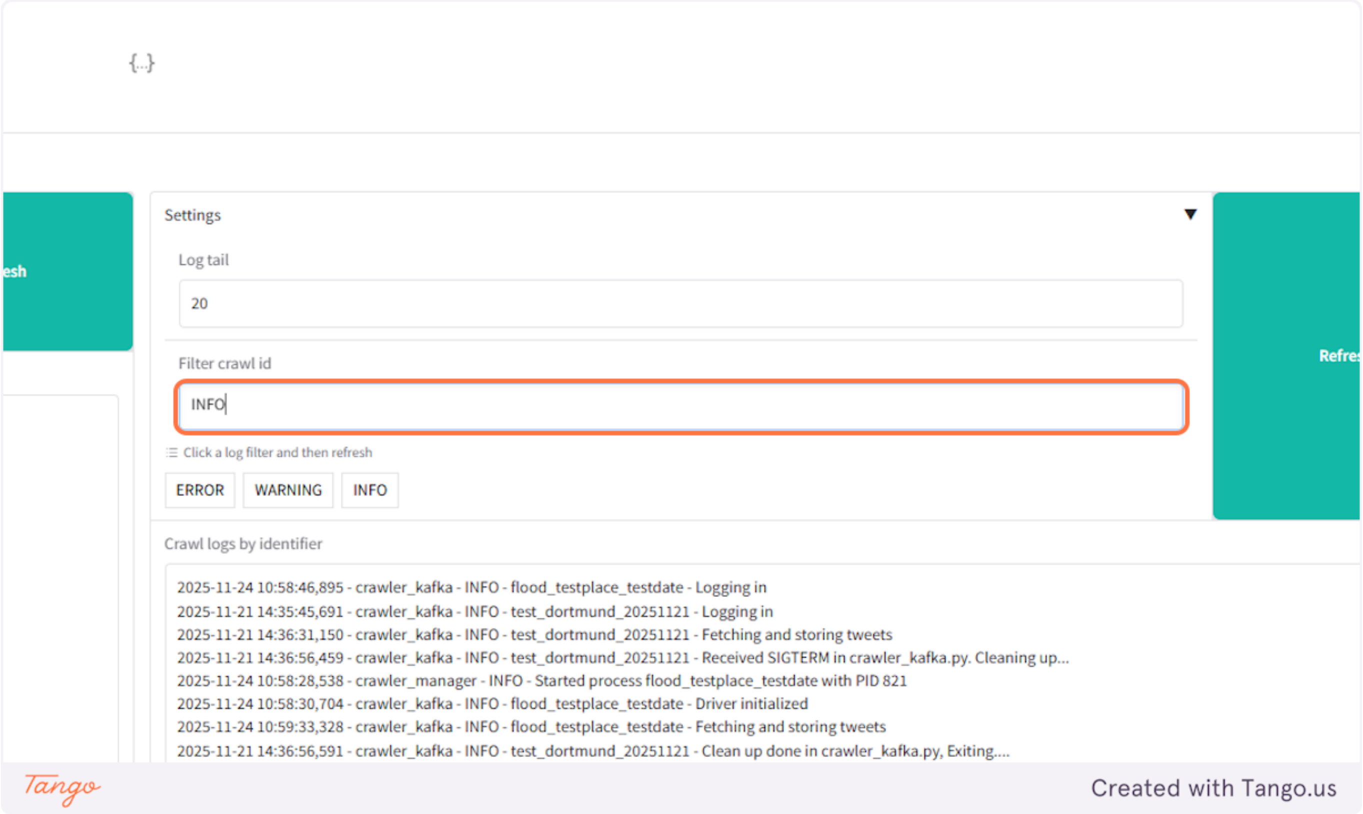
INFO

Click a log filter and then refresh

Crawl logs by identifier

Click Refresh

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



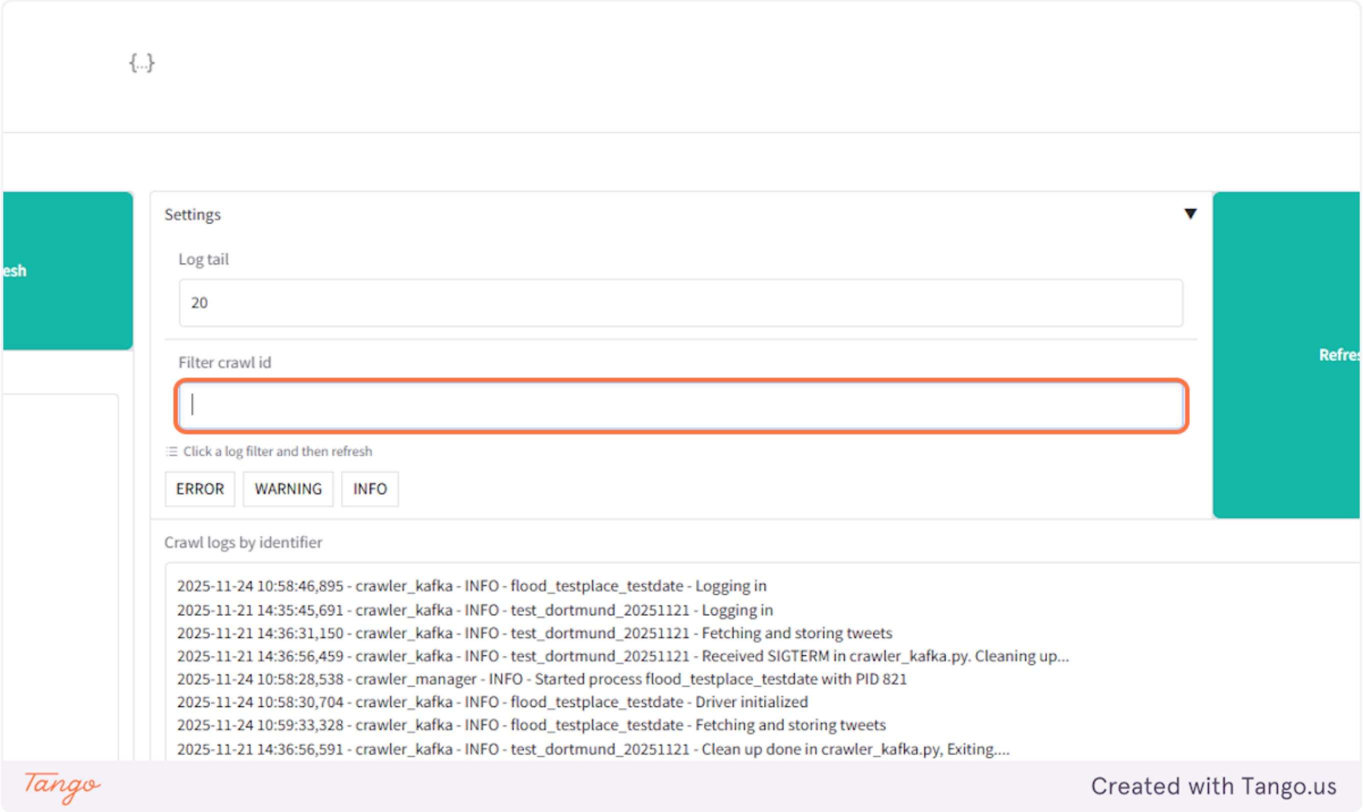
Log tail: 20

Filter crawl id: 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
2025-11-24 10:59:33,328 - crawler_kafka - INFO - flood_testplace_testdate - Fetching and storing tweets
2025-11-21 14:36:56,591 - crawler_kafka - INFO - test_dortmund_20251121 - Clean up done in crawler_kafka.py, Exiting....
```

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.



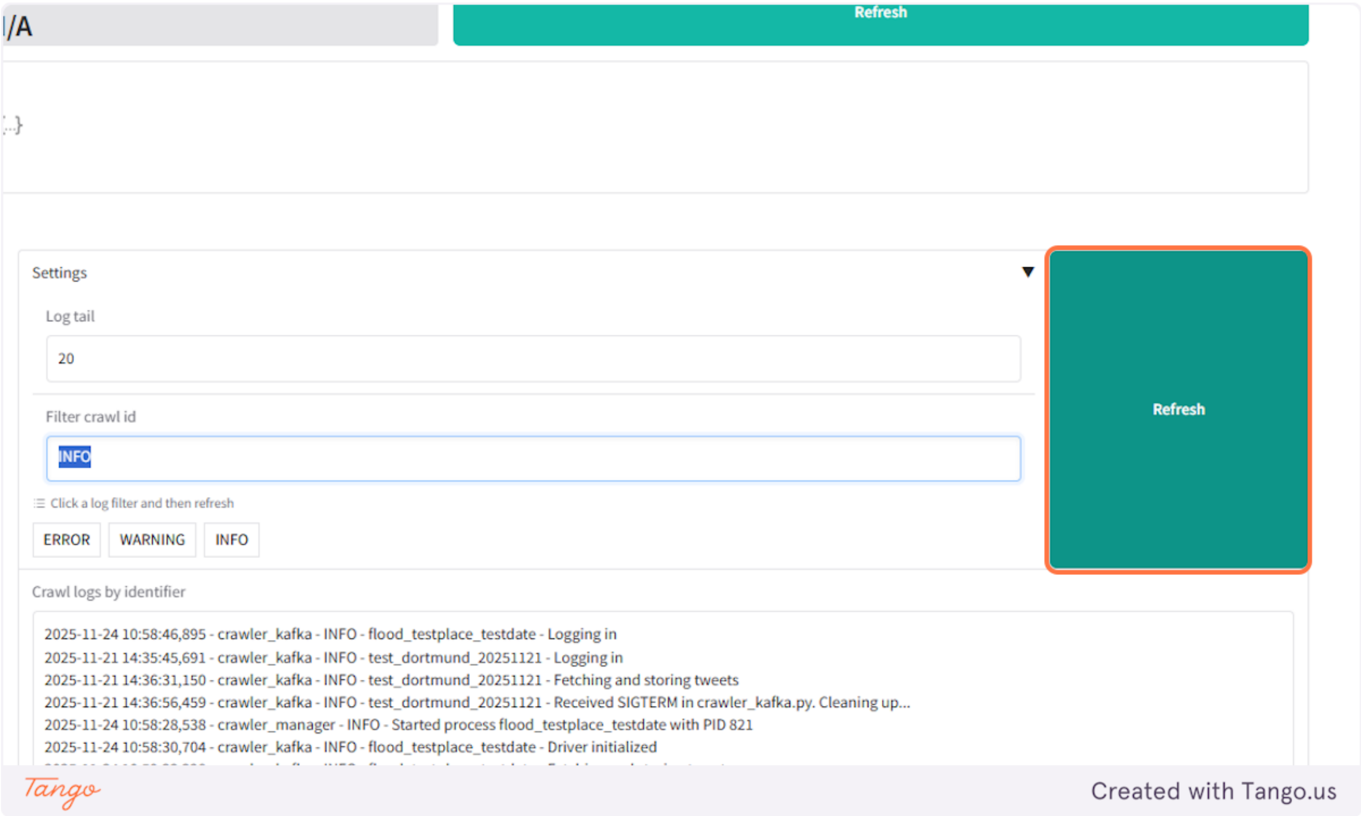
Log tail: 20

Filter crawl id:

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
2025-11-24 10:59:33,328 - crawler_kafka - INFO - flood_testplace_testdate - Fetching and storing tweets
2025-11-21 14:36:56,591 - crawler_kafka - INFO - test_dortmund_20251121 - Clean up done in crawler_kafka.py, Exiting....
```

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



Log tail: 20

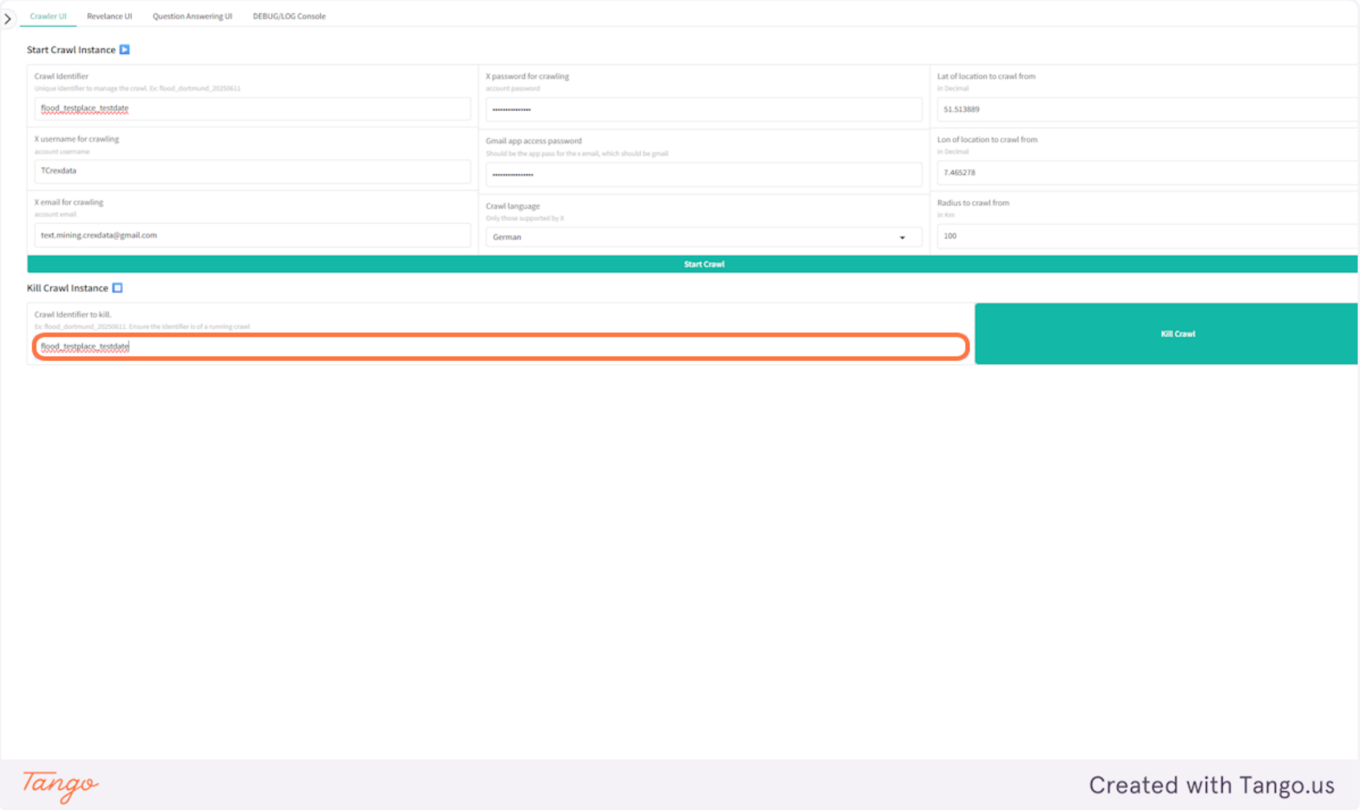
Filter crawl id: INFO

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
```

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.



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
TCrexdata

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

Lon of location to crawl from
in Decimal
7.465278

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

Crawl language
Only those supported by X
German

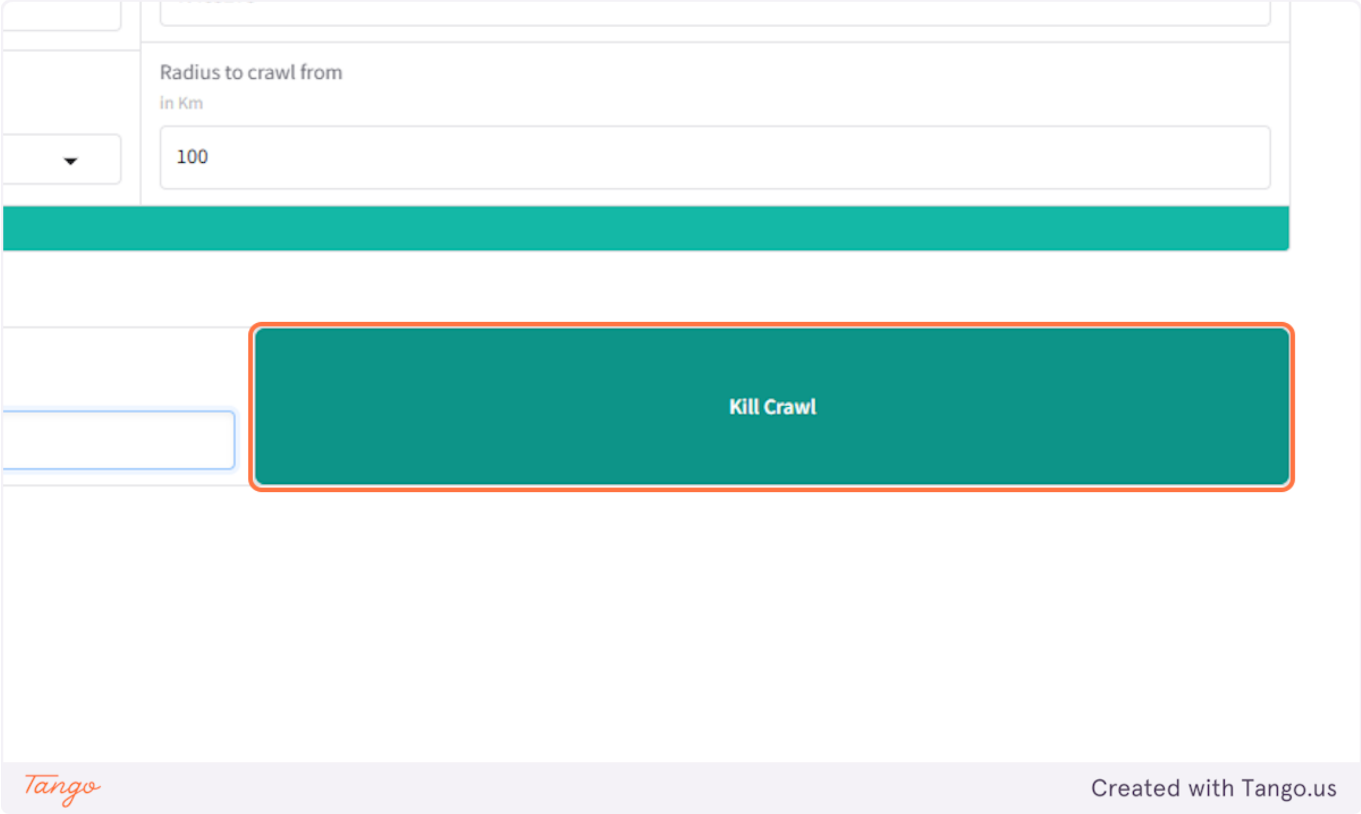
Radius to crawl from
in Km
100

Kill Crawl Instance

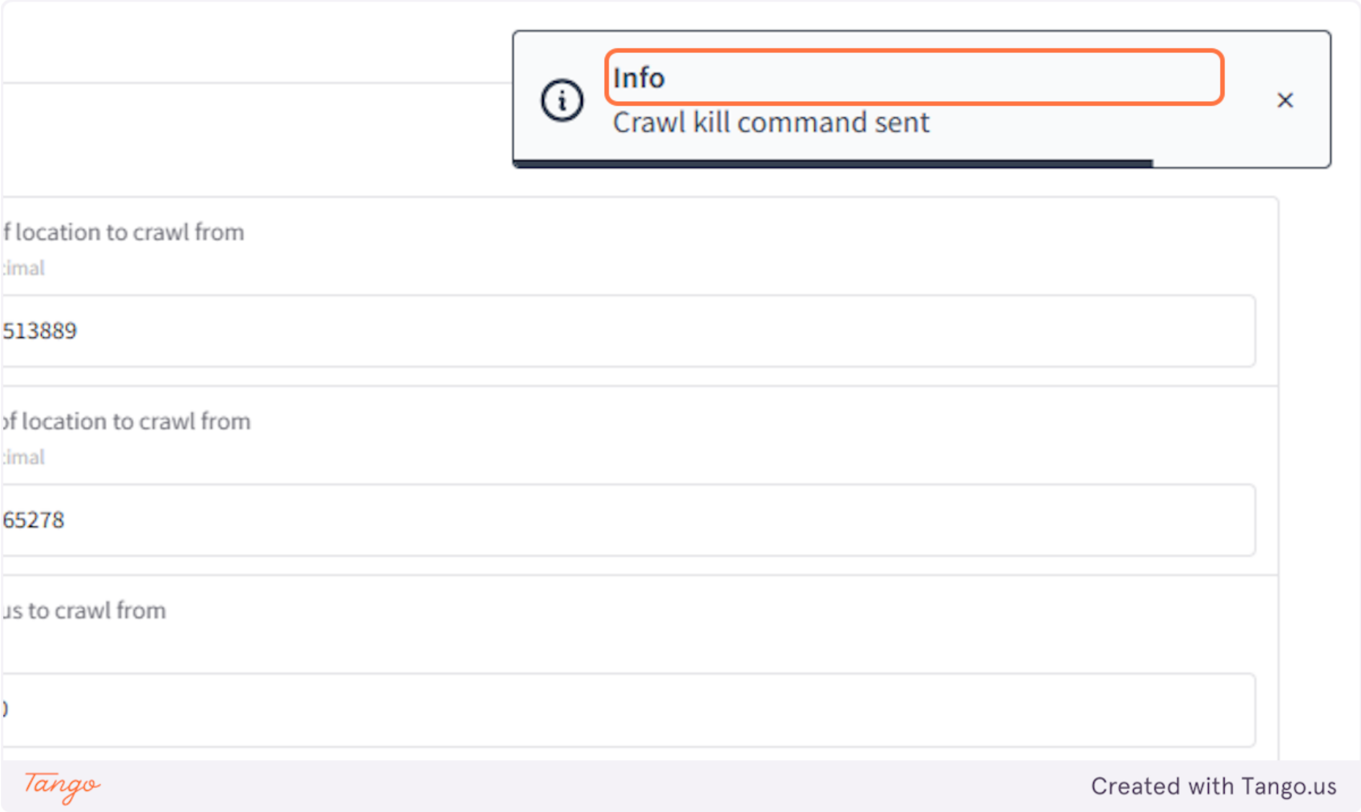
Crawl Identifier to kill.
Ex: flood_dortmund_20250611. Ensure the identifier is of a running crawl
flood_testplace_testdate

Created with Tango.us

19. Then click on "Kill Crawl".


Radius to crawl from
in Km
100

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

of location to crawl from
:imal
65278

us to crawl from
)

Created with Tango.ai