Beeg Meme Project 3rd Deliverable Tests

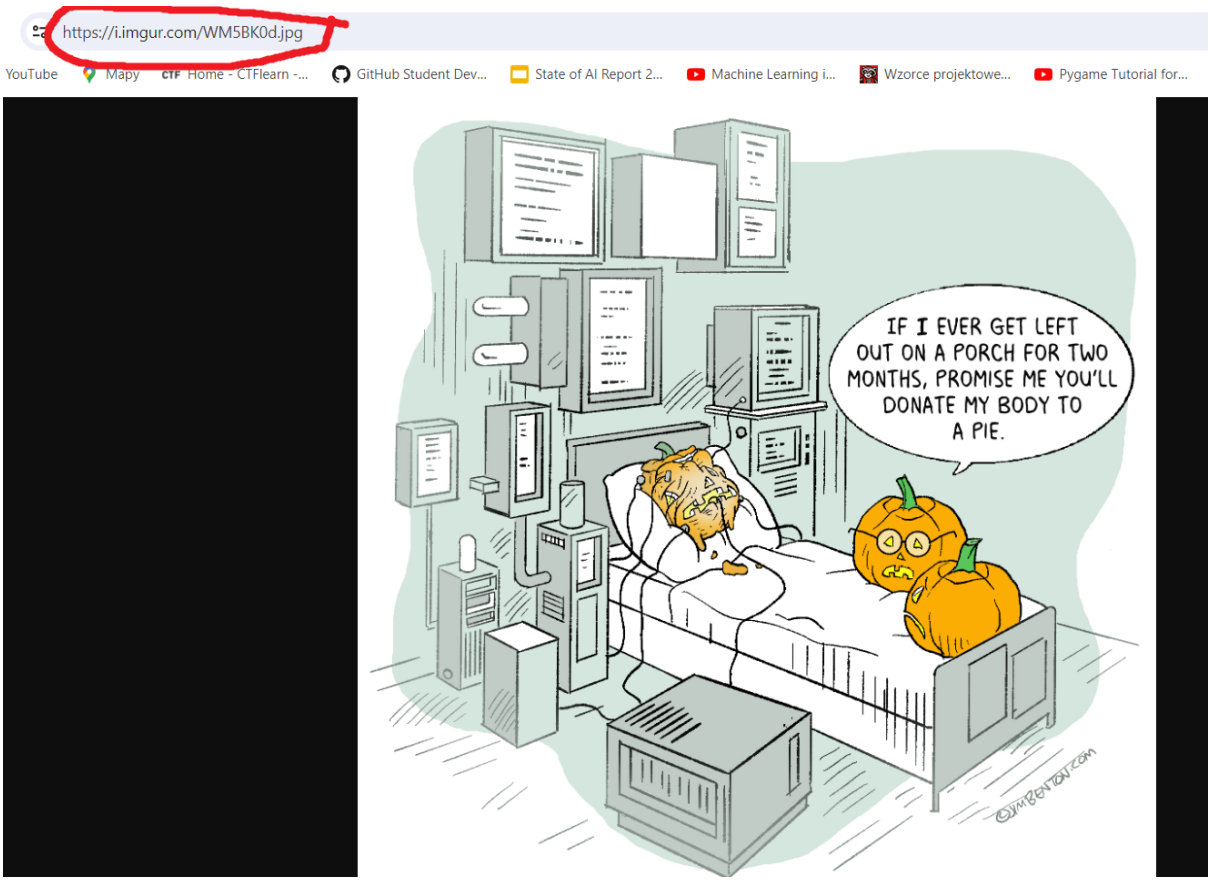
Test table

Test objective	Test steps	Expected Result	Actual Result
Test the image backup capabilities	1. Start the Nifi flow for the Imgur path. Download a single post 2. View the image associated with the post 3. Check that the image has been downloaded into the backup bucket.	The image connected to the downloaded post is saved into a dedicated GCS bucket.	As expecte  d (proof below)
Test the re-emittance Nifi path	1. Pick a date and an hour and find an appropriate .tar file on the Master Log bucket 2. Retrieve this item on the re-amittance path 3. Post deduplicated posts from it to kafka	The posts with changed url (so the posts from the re-amittance path) have been published to kafka.	As expected (proof below)
Test the ViT module	1. Ensure Ingest module and Kafka is running 1a. If Ingest module is unavailable, the test can be performed by using a Kafka Console Producer and a sample of data 2. Start the ViT Spark Job and verify [UDF] tagged log entries are appearing 3. Open a Kafka	Messages containing images that the model is capable of running inference on are appearing in the Kafka console consumer, with their embeddings. Messages that cannot be processed are arriving with an empty string.	Mostly as expected (proof below) Some “normal” images weren’t processed due to issues with inferring their size.

	Console consumer with the “vit2analytics” topic		
Test the Clustering module	1. Ensure Ingest and ViT module as well as Kafka is running 1a. If Ingest/ViT module is unavailable, the test can be performed by using a Kafka Console Producer and a sample of data 2. Start the Cluster Spark Job and verify [UDF] tagged log entries are appearing 3. Open a Kafka Console consumer with the “cluster2analytics” topic	Messages containing embeddings that the model is capable of running inference on are appearing with a non-negative cluster in the Kafka console consumer. Messages where the inference failed are appearing with cluster “-1”, and messages where embeddings were missing have their cluster set to “-2”.	Mostly as expected (proof below) Occasional error with “-1” cluster appeared unexpectedly.

Image backup test

First, we download a single post from the Imgur path. We copy the link from it and check the picture it leads to.



Next, we check if the image has been saved in the backup bucket. As we can see, the image is indeed saved under the appropriate URL.

Szczegóły obiektu

Zasobniki > images_beeg_meme > https: > / > i.imgur.com > WM5BK0d.jpg

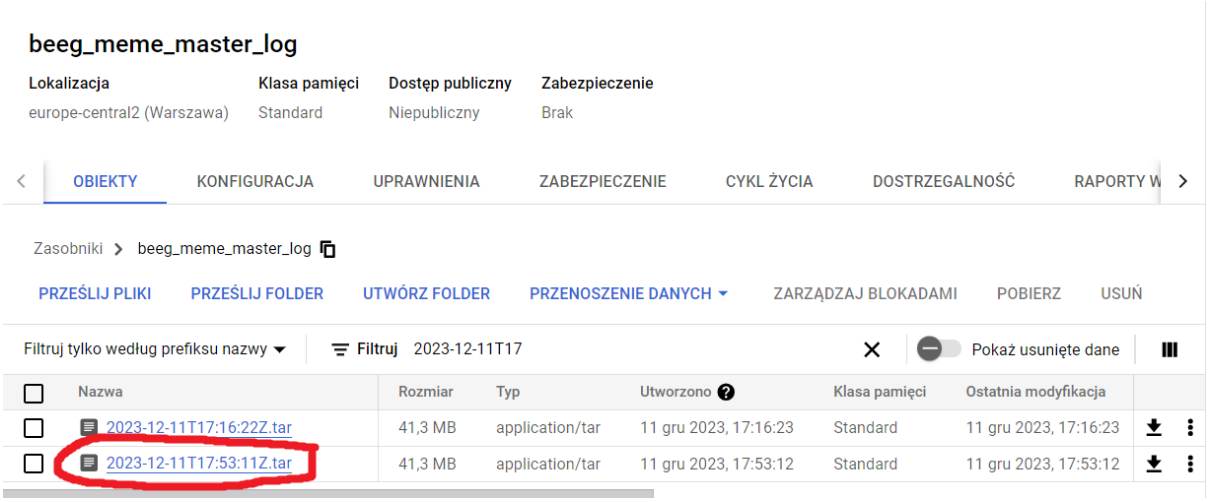
Opis	
Typ	image/jpeg
Rozmiar	377,5 KB
Utworzono	17 gru 2023, 23:38:29
Ostatnia modyfikacja	17 gru 2023, 23:38:29
Klasa pamięci	Standard
Czas niestandardowy	—
Publiczny adres URL	Nie dotyczy
Uwierzytelniony adres URL	https://storage.cloud.google.com/images_beeg_meme/https%3A//i.imgur.com/WM5BK0d.jpg
Identyfikator URI polecenia gsutil	gs://images_beeg_meme/https://i.imgur.com/WM5BK0d.jpg
Uprawnienia	
Dostęp publiczny	Niepubliczny
Zabezpieczenie	
Historia zmian	—
Zasada przechowywania	Brak
Stan blokady	Brak
Typ szyfrowania	Zarządzany przez Google

Zasobniki > images_beeg_meme > https: > / > i.imgur.com > WM5BK0d.jpg

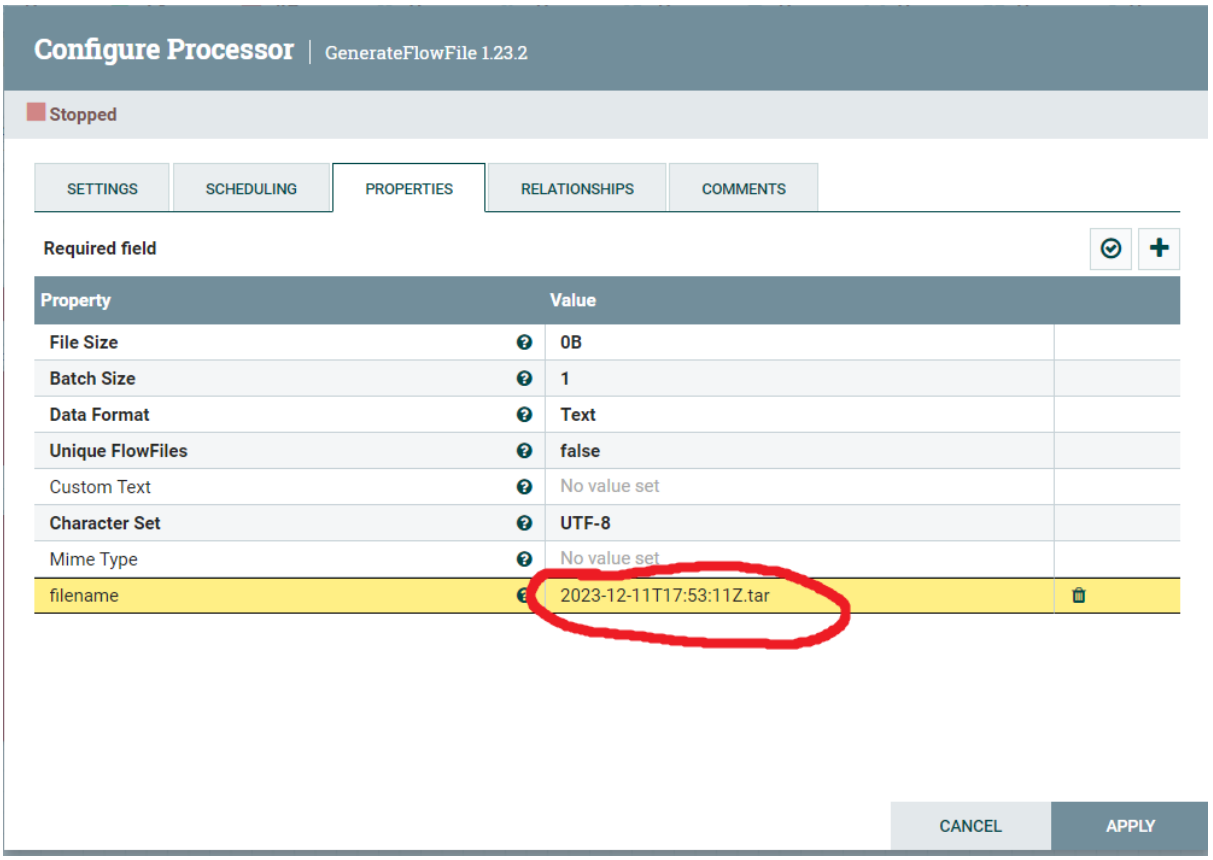
We can see the image inside the bucket (as indicated by the path above it).

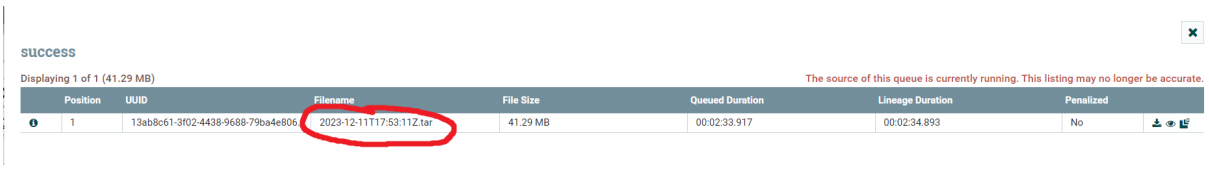
Re-emittance path test

We first pick a date and time. We have picked 5pm from the 11th of december. We see that there is a .tar log file for this date.



We retrieve it through the re-emittance path.





After the preprocessing, we can see that the files once again have unified fields. What is more, the URL address have been changed to reflect the one in the bucket.

View as: formatted


```
{
  "global_id" : "18fy3ww-reddit",
  "author" : "t2_6oczyrju",
  "created_time" : "2023-12-11T16:13:06",
  "desc" : "I doubt it at this point",
  "score" : 1,
  "url" : "https://storage.cloud.google.com/images_beeg_meme/https://i.redd.it/qxfxzfi5xo5c1.jpeg",
  "source" : "reddit"
}
```


Finally, we publish to Kafka and see that the new messages indeed have a different URL (pointing to the backup bucket).

```
jakub_foltyn1217@nifi:/home/nifi/kafka_2.13-3.6.0$ sudo bin/kafka-console-consumer.sh --bootstrap-server localh
ost:9092 --topic nifi2vit
{"global_id":"18fxvru-reddit","author":"t2_i4fmb3gm","created_time":"2023-12-11T16:03:17","desc":"","score":6,"
url":"https://storage.cloud.google.com/images_beeg_meme/https://i.redd.it/xj0vfih7vo5c1.jpg","source":"reddit"}
{"global_id":"18fxqa0-reddit","author":"t2_s7470p9u","created_time":"2023-12-11T15:56:52","desc":"","score":1,"
url":"https://storage.cloud.google.com/images_beeg_meme/https://i.redd.it/mgfz6h09uo5c1.png","source":"reddit"}
{"global_id":"18fxak6-reddit","author":"t2_8qo73k0tu","created_time":"2023-12-11T15:37:35","desc":"","score":8,
"url":"https://storage.cloud.google.com/images_beeg_meme/https://i.redd.it/gmvg2x4sqo5c1.png","source":"reddit"
}
```

ViT module test

The ViT module test is dependent on the ingestion module and Kafka working correctly. This is obviously less than ideal, but the test could be adapted to use a Kafka Console Producer with the appropriate messages.

First, in one SSH session, a Spark structured streaming job is started without detaching the console output. Eventually, given enough messages arriving in Kafka, a minibatch will start. The model works by using a UDF which logs which images are currently being processed.

```
  File "<stdin>", line 10, in process
  File "/usr/local/lib/python3.9/dist-packages/transformers/image_processing_utils.py", line 549, in __call__
    return self.preprocess(images, **kwargs)
  File "/usr/local/lib/python3.9/dist-packages/transformers/models/mobilevit/image_processing_mobilevit.py", li
ne 281, in preprocess
    input_data_format = infer_channel_dimension_format(images[0])
  File "/usr/local/lib/python3.9/dist-packages/transformers/image_utils.py", line 189, in infer_channel_dimensi
on_format
    raise ValueError("Unable to infer channel dimension format")
ValueError: Unable to infer channel dimension format

[UDF] Processing https://i.redd.it/yykzxarj795c1.jpg
[UDF] Processing https://i.redd.it/ipm5tue77a5c1.jpeg
[UDF] Processing https://i.redd.it/exade8dl3a5c1.jpeg
[UDF] Processing https://i.redd.it/c4tar6lx0a5c1.jpg
[UDF] Processing https://i.redd.it/7cliqa5bw95c1.png
Traceback (most recent call last):
  File "<stdin>", line 10, in process
  File "/usr/local/lib/python3.9/dist-packages/transformers/image_processing_utils.py", line 549, in __call__
    return self.preprocess(images, **kwargs)
  File "/usr/local/lib/python3.9/dist-packages/transformers/models/mobilevit/image_processing_mobilevit.py", li
ne 281, in preprocess
    input_data_format = infer_channel_dimension_format(images[0])
  File "/usr/local/lib/python3.9/dist-packages/transformers/image_utils.py", line 189, in infer_channel_dimensi
on_format
    raise ValueError("Unable to infer channel dimension format")
ValueError: Unable to infer channel dimension format

[UDF] Processing https://i.redd.it/g23lnsygj95c1.jpg
[UDF] Processing https://i.redd.it/2wxrfj4xd95c1.jpg
```

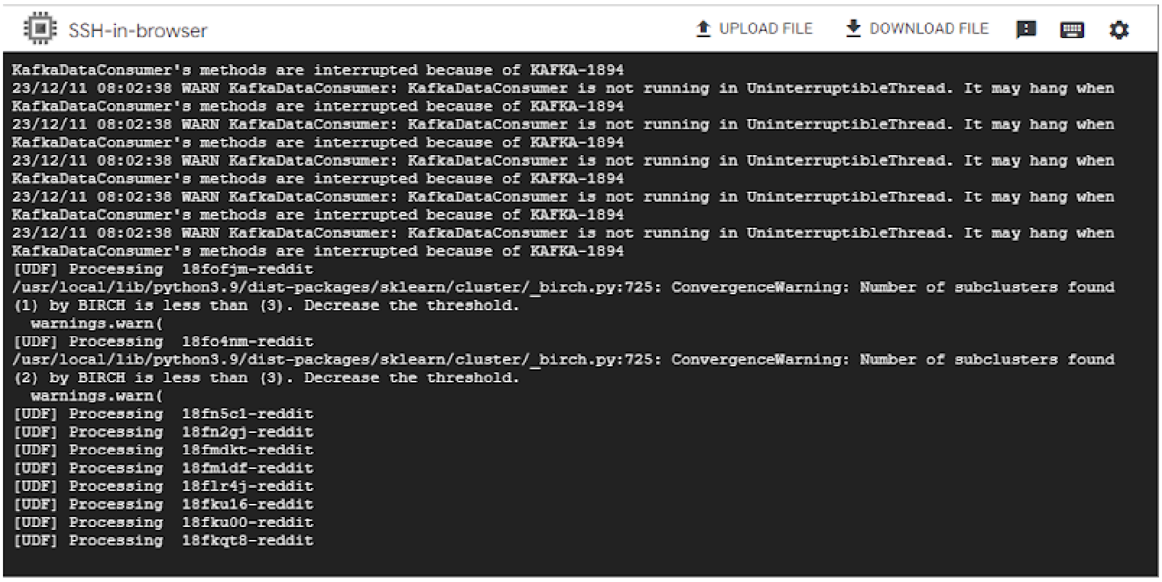
To verify that the module outputs the data correctly back on to Kafka we open a second SSH session and verify that the messages arrive:

```
kacper_grzymkowski00@processing:~$ kafka_2.13-3.6.0/bin/kafka-console-consumer.sh --bootstrap-server nifi:9092
--topic vit2analytics
{"global_id":"18egn0a-reddit","embeddings":""}
{"global_id":"18efl0v-reddit","embeddings":""}
{"global_id":"18ef6bg-reddit","embeddings":""}
{"global_id":"18eesbh-reddit","embeddings":"[-0.0001344276242889464, 5.8721670939121395e-05, -0.011037714779376
984, 0.00017634544929023832, -0.00010015119914896786, 0.15342006087303162, 0.00017552573990542442, 0.3912289142
6086426, 0.5732612609863281, -0.005986415781080723, -0.06410466134548187, 2.5796343834372237e-05, -0.3796417415
1420593, 9.045179467648268e-06, 4.695327515946701e-05, 6.812764331698418e-05, 0.09979596734046936, -0.217360034
58499908, -1.301190241065342e-05, -8.526384044671431e-05, -0.19885659217834473, 0.2250870168209076, -3.70769921
5738103e-05, 2.818646316882223e-05, -0.00010927546827588823, 6.543449853779748e-05, 0.00013226654846221209, 0.16
086433827877045, -1.0300899744033813, -0.03401597589254379, -0.9182242155075073, 0.1487763375043869, 0.36292356
25267029, -0.7703137397766113, 6.731135363224894e-05, 1.3938087224960327, 1.0703011751174927, 2.882828812289517
4e-05, -0.007153227925300598, 0.03991322964429855, 2.4782668333500624e-05, -0.00032861391082406044, -0.00070188
93375061452, -0.9315118193626404, -0.5202741622924805, -8.486994192935526e-05, 0.31911414861679077, -0.93870407
34291077, -0.015056824311614037, -0.08211596310138702, -3.336201189085841e-06, -0.00037676963256672025, -0.0600
24701058864594, -0.03854592144489288, 0.2941540479660034, -4.442226781975478e-05, 0.24105465412139893, -0.05298
139899969101, -3.2356794690713286e-05, -0.6562302112579346, 5.793822492705658e-05, -0.0005817145574837923, 7.35
9263690887019e-05, -0.00018779534730128944, -8.826454723021016e-05, 0.19539877772331238, -1.591230829944834e-05
, -6.433736416511238e-06, -0.4102936387062073, -0.0001336969726253301, 0.5022785067558289, -0.00012876493565272
54, -0.00017137386021204293, 0.00015302631072700024, 0.00017401107470504494, -2.323291846551001e-05, 8.207782957
470044e-05, 0.00012973409320693463, -0.00035189767368137836, -8.584761235397309e-05, 0.17410579323768616, -0.04
```

Due to the chosen encoding, the outputs are quite long. Some messages arrive with no embeddings - this is expected, as videos and animated GIFs are not supported.

Clustering module testing

Similarly to the ViT module, the clustering module was tested by turning on all upstream modules (Ingestion, ViT), as well as Kafka. If upstream modules are not available, then a Kafka console producer with appropriate messages can be used.
Another Spark job is submitted for the clustering job:



```
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
23/12/11 08:02:38 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
23/12/11 08:02:38 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
23/12/11 08:02:38 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
23/12/11 08:02:38 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
23/12/11 08:02:38 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when
KafkaDataConsumer's methods are interrupted because of KAFKA-1894
[UDF] Processing  18fofjm-reddit
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_birch.py:725: ConvergenceWarning: Number of subclusters found
(1) by BIRCH is less than (3). Decrease the threshold.
  warnings.warn(
[UDF] Processing  18fo4nm-reddit
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_birch.py:725: ConvergenceWarning: Number of subclusters found
(2) by BIRCH is less than (3). Decrease the threshold.
  warnings.warn(
[UDF] Processing  18fn5c1-reddit
[UDF] Processing  18fn2gj-reddit
[UDF] Processing  18fmdkt-reddit
[UDF] Processing  18fm1df-reddit
[UDF] Processing  18flr4j-reddit
[UDF] Processing  18fku16-reddit
[UDF] Processing  18fku00-reddit
[UDF] Processing  18fkqt8-reddit
```

As can be seen, [UDF] tagged logs are appearing, meaning the job is doing something.
By opening another session and starting a Kafka console consumer we can see the resulting clusters being emitted into the “cluster2analytics” topic:

SSH-in-browser UPLOAD FILE DOWNLOAD FILE


```
{"global_id":"18fmdkt-reddit","cluster":-2}
{"global_id":"18fm1df-reddit","cluster":-2}
{"global_id":"18flr4j-reddit","cluster":-2}
{"global_id":"18fku16-reddit","cluster":-2}
{"global_id":"18fku00-reddit","cluster":-2}
{"global_id":"18fkqt8-reddit","cluster":-2}
{"global_id":"18fkpts-reddit","cluster":-2}
{"global_id":"18fkn77-reddit","cluster":-1}
{"global_id":"18fk51h-reddit","cluster":-2}
{"global_id":"18fjx74-reddit","cluster":-2}
{"global_id":"18fjrav-reddit","cluster":-2}
{"global_id":"18foy6a-reddit","cluster":0}
{"global_id":"18foo6v-reddit","cluster":1}
{"global_id":"18foisl-reddit","cluster":0}
{"global_id":"18ford9-reddit","cluster":0}
{"global_id":"18folxm-reddit","cluster":1}
{"global_id":"18foj7z-reddit","cluster":0}
{"global_id":"18fnus4-reddit","cluster":0}
{"global_id":"18fnucg-reddit","cluster":1}
{"global_id":"18fnc31-reddit","cluster":1}
{"global_id":"18fn98p-reddit","cluster":-1}
{"global_id":"18fn679-reddit","cluster":0}
{"global_id":"18fmu3a-reddit","cluster":2}
{"global_id":"18fmo3i-reddit","cluster":0}
{"global_id":"18fmnt8-reddit","cluster":-2}
{"global_id":"18fmnjp-reddit","cluster":0}
{"global_id":"18fmn8s-reddit","cluster":0}
{"global_id":"18fmhd9-reddit","cluster":1}
{"global_id":"18fm2gw-reddit","cluster":0}
```


As can be seen, some non-negative results are appearing, which means the model is performing some clustering and inferring the cluster of new elements.