



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687*

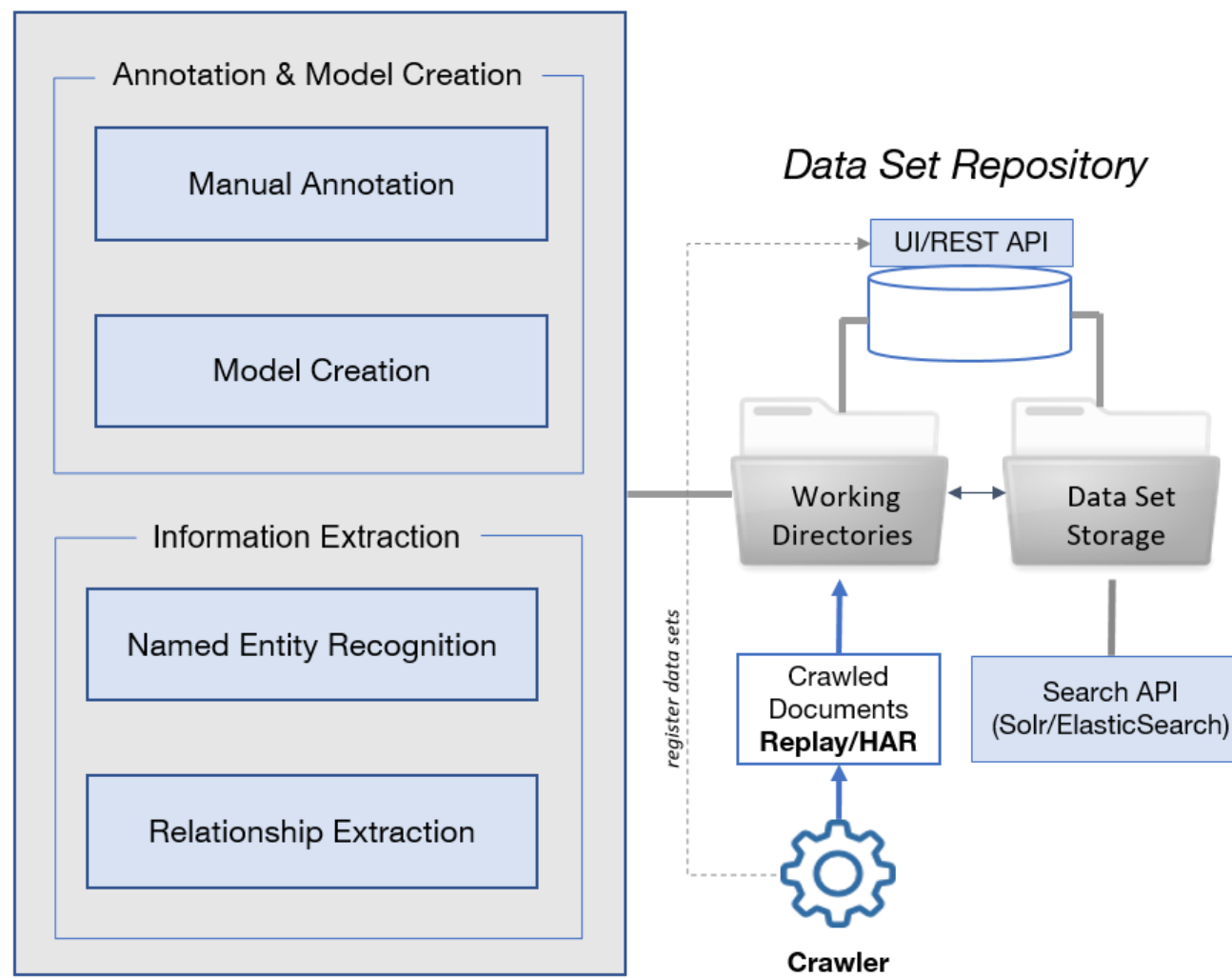
# Extracting information from darknet market advertisements and forums

- 
- Sven Schlarb [Sven.Schlarb@ait.ac.at](mailto:Sven.Schlarb@ait.ac.at), Mina Schütz [Mina.Schuetz@ait.ac.at](mailto:Mina.Schuetz@ait.ac.at),  
Clemens Heistracher [clemens.heistracher@ait.ac.at](mailto:clemens.heistracher@ait.ac.at)  
AIT Austrian Institute of Technology, Competence Unit Data Science & Artificial Intelligence
  - Faisal Ghaffar [FAISALGH@ie.ibm.com](mailto:FAISALGH@ie.ibm.com)  
IBM Innovation Exchange, Cognitive Computing Group

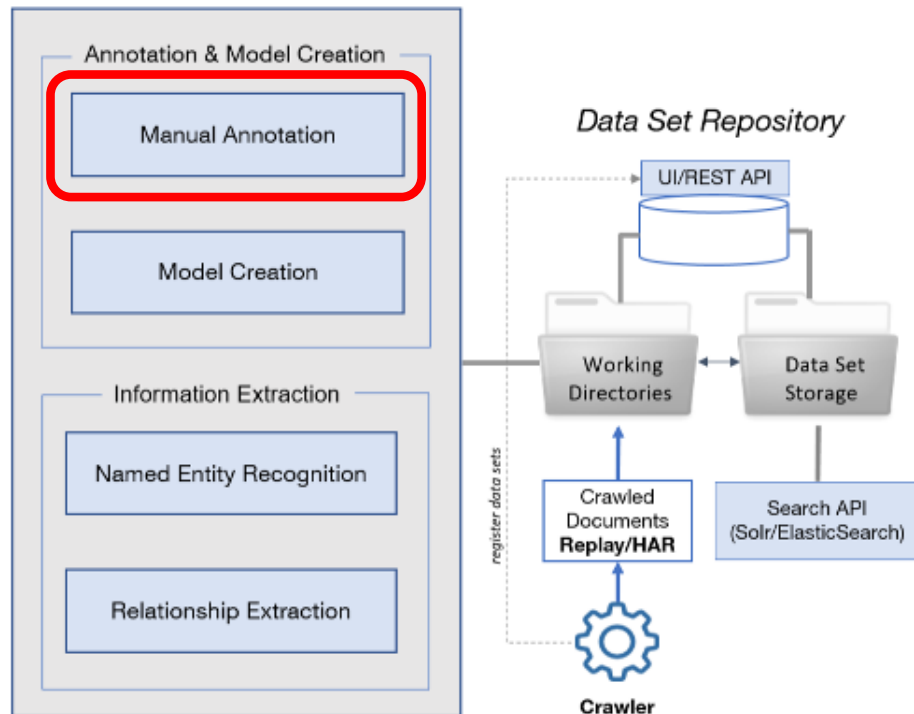
SECURWARE 2020 November 21, 2020 to November 25, 2020 - Valencia, Spain

# Overview

## Named entity recognition & Relationship extraction



# Manual annotation



The screenshot shows the **COPKIT Data Repository** interface. The top navigation bar includes links for **Administration**, **Data set creation**, **Data set management**, **Knowledge Extraction**, **Search & Access**, **Language**, and a user profile for **admin**.

### Annotation

The interface displays a list of data sets. The selected data set is **dnm-archives-2013-2015-grams-caas**. A **load** button is next to it.

Below the data set list, there is a text area containing a list of items. The selected item is **19 - Allright this is the first gun i listed ..**.

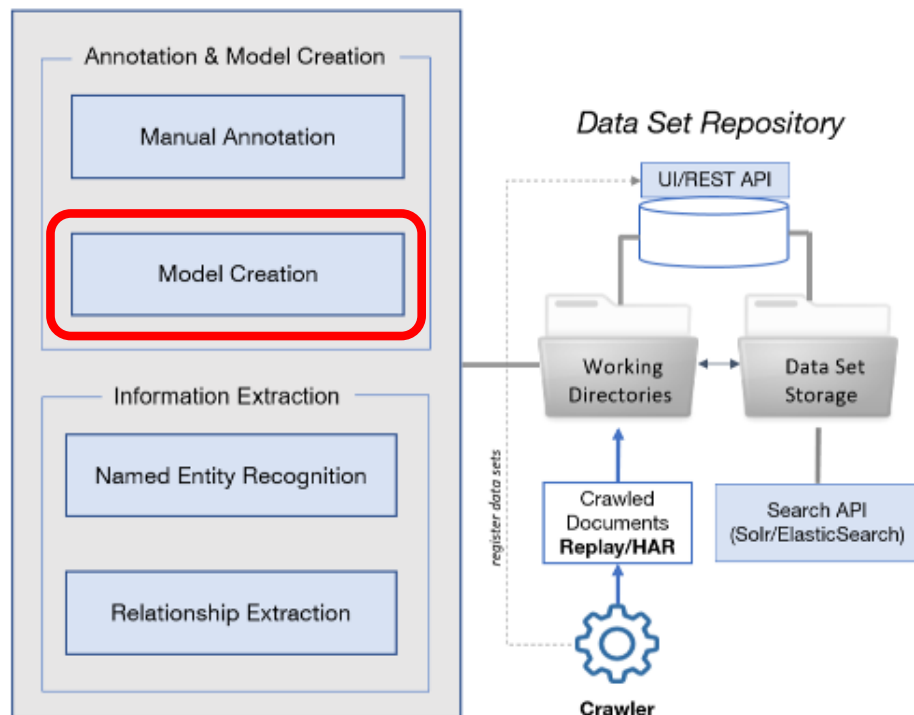
On the right side, there is a text area for manual annotation. The text being annotated is:

Allright this is the first gun i listed It is a **Remington** 1858 It got two drums one for ball and cap and one for 45 **Colt** cartridges The barrel is rifled and it s a great homedefence gun It is accurate But if your planning of hitting soda cans **200 meters** away i suggest you keep looking Includes besides the gun and the drums are Weaponcase i leather or leather like material I cant tell the difference 20 pcs of **Colt** 45 Ammo 50 pcs lead ball ammo If using ball and cap Black powder and measuring unit If using ball and cap Primers If using ball and cap The price than have i gone mad No i have not In fact it s a great price look it up The great thing is that this gun in legal to own in many **European** countries The drum that carries 45 cartridges however is not But here is the thing I do full escrow and i send in inside of **Sweden** No problems Do you want to take your car and collect it in **Sweden** No problems Full escrow But th

Below the text area, there is a **Store changes** button and a **Suggest annotations** button.

At the bottom, there is a **Persist annotations** button.

Version 1.0 (28.02.2020)



COPKIT Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Start processing

Processing: grams-entities-model

Selected data sources:

Data source label	Data source process id	Representations
ML1.1.3-dnm-archives-weaponsdrugs-subset	38f02113-9d42-4cf6-be09-517a0942a50f	<ul style="list-style-type: none"><li>mldata (364b54c0-fca5-4244-9eed-07625a03bce9)</li><li>grams (f11f0d1b-f099-4b92-bcd6-904b10e8f357)</li></ul>

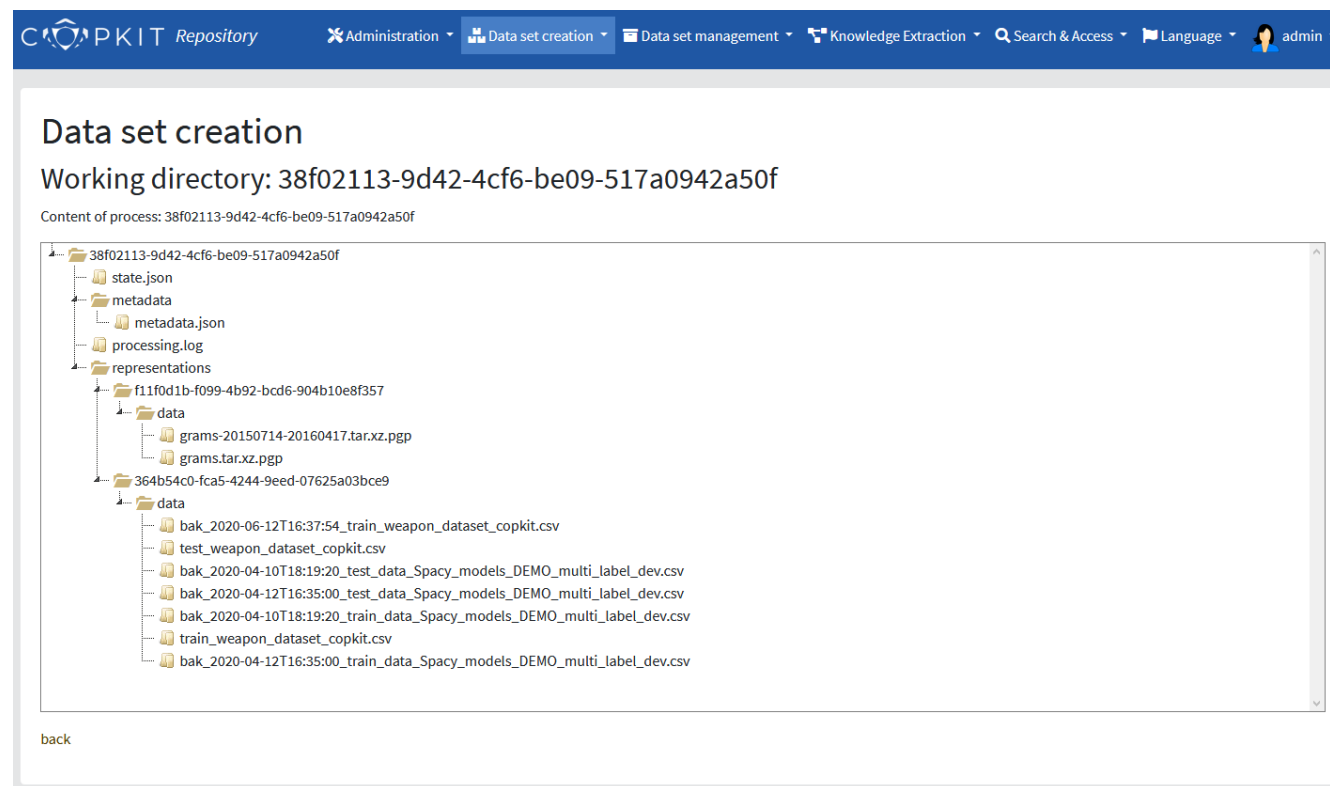
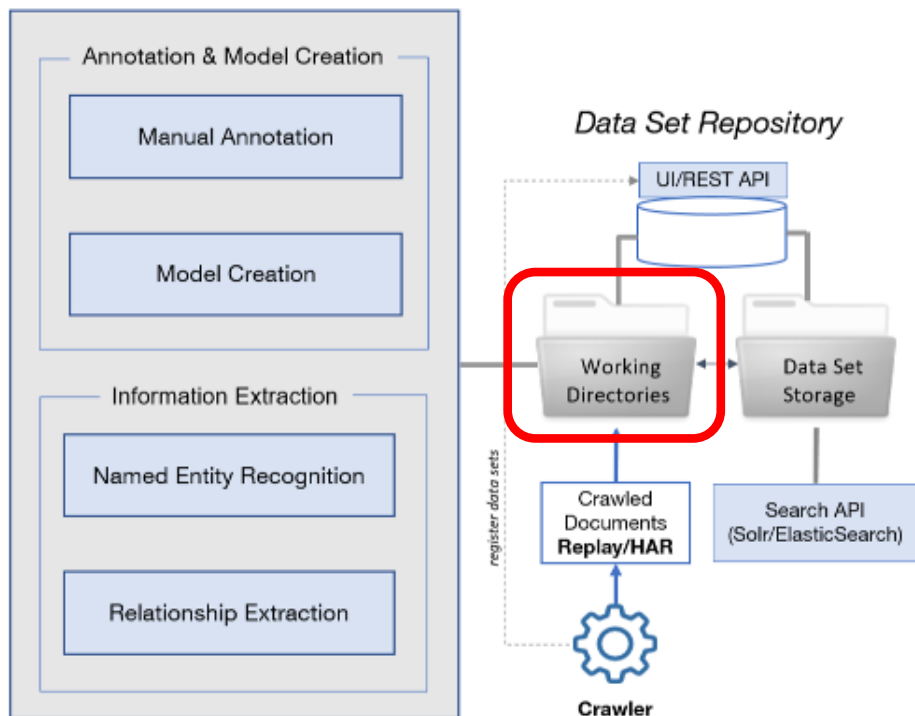
Processing pipeline

named\_entity\_extraction

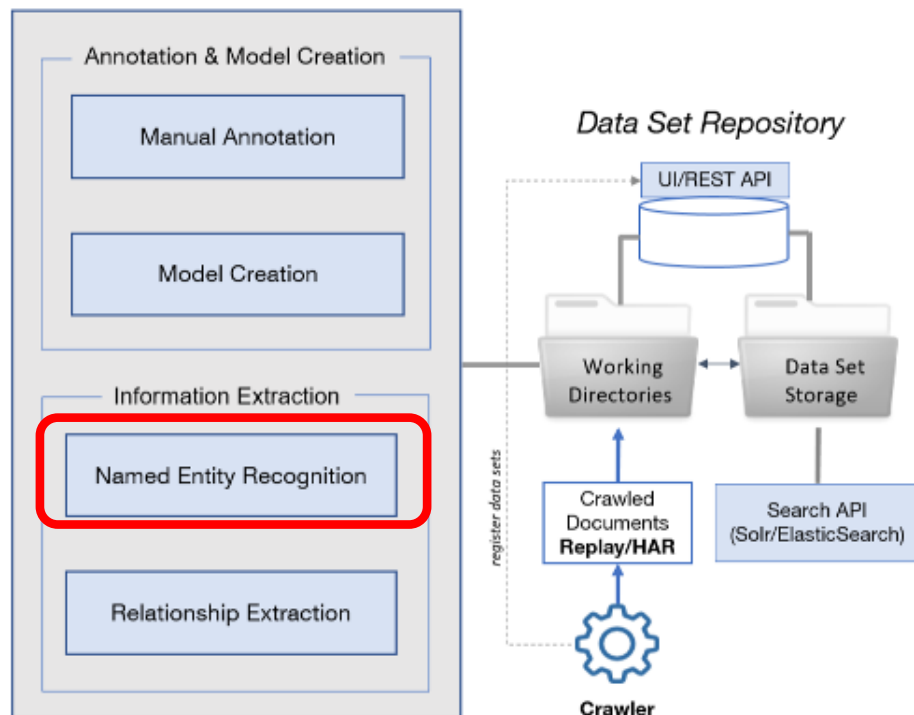
Start processing

Version 1.0 (30.04.2020)

# Working directories

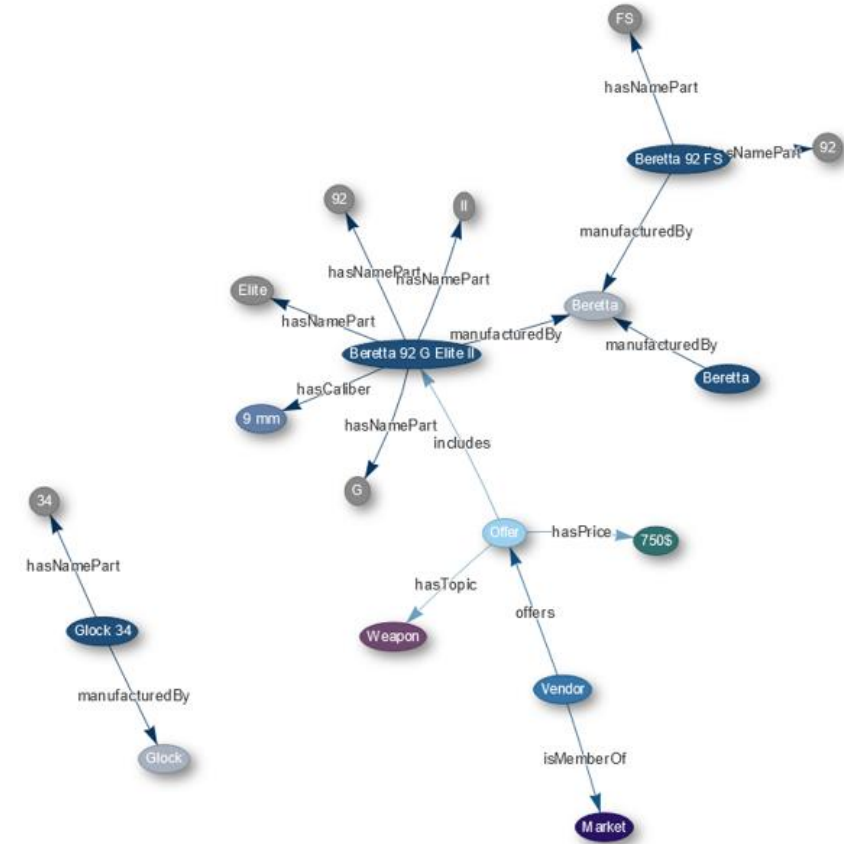
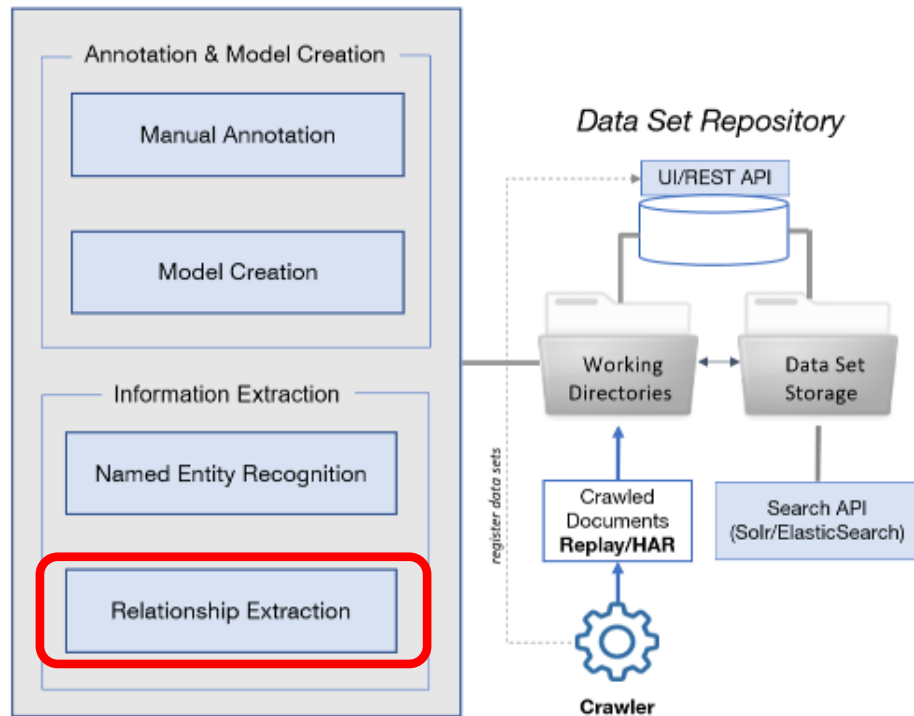


# Named Entity Recognition (CKNER)

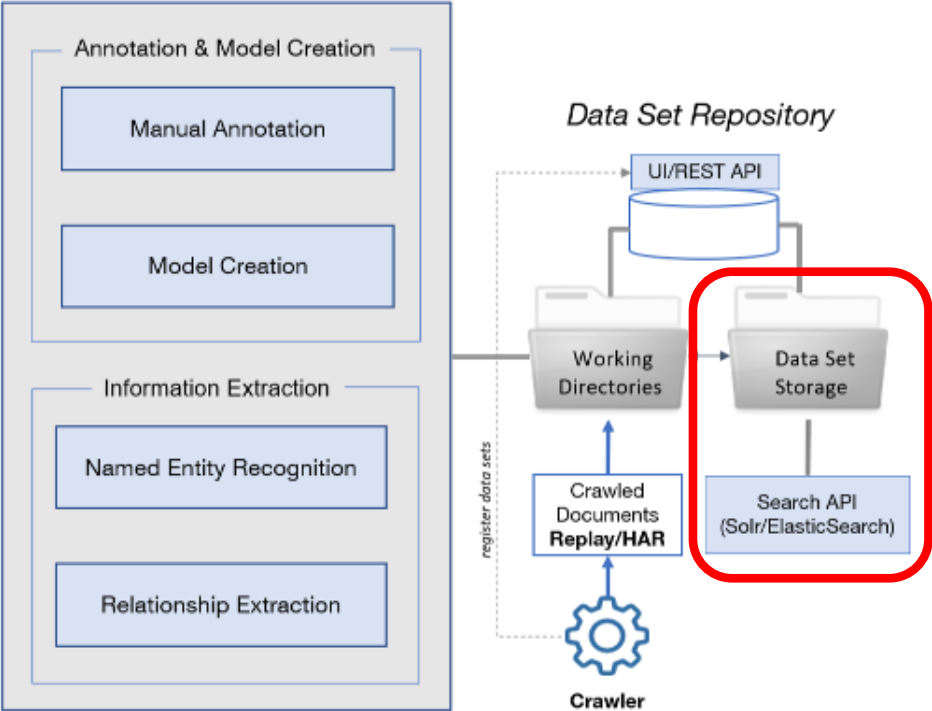


I have a Beretta WEAPON\_MANUFACTURER 92G Elite II I am looking to sell. It a 9mm QUANTITY designed for IDPA ORG or USPSA GPE shooting. It has the heavier Brigadier slide to help reduce felt recoil and muzzle flip. It has night sights. The G model means it is the de-cocker only model. No manual safety like a regular Beretta PRODUCT 92 FS. It operates more like a Sig 226 in that the de-cocker will drop the hammer safely, but the gun is still ready to fire in double action. These are very hard to find, Beretta PRODUCT no longer makes them. It is very similar in set up to the new Wilson Combat Custom Berettas. Comes with original case and two CARDINAL magazines. I got it intending to use it to shoot IDPA ORG, but I have a Glock WEAPON\_MANUFACTURER 34 that I am able to shoot better with. So the Beretta PRODUCT just sits in the safe. It is time it finds a new owner that will use it as intended! I have some bills I need to pay so I am motivated to sell. Looking for \$750 or best offer. I will also consider trades. Please contact for more pictures or questions.

# Relationship Extraction (RELEXT)



# Working directories



The screenshot shows the COPKIT Data Repository web interface. The top navigation bar includes links for Administration, Data set creation, Data set management, Knowledge Extraction, Search & Access, Language, and a user profile (admin).

The main section is titled "Search & Access" and "Full-text search". It contains a search form with the following fields:

- Search term**: A text input field.
- Information package**: A text input field with the placeholder "Identifier of the information package".
- Content type**: A dropdown menu with options: odt, doc, html, pdf.
- Sort**: A dropdown menu with the option "Relevance".
- Only data set**: A checkbox that is currently checked.
- Search**: A button to execute the search.

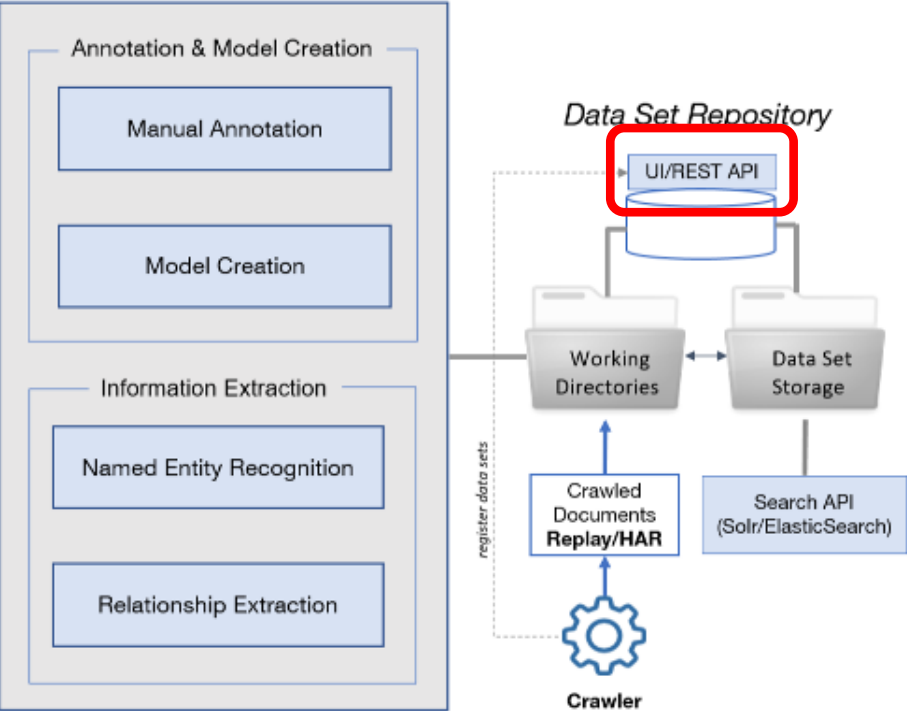
Below the search form, the version "Version 1.0 (28.02.2020)" is displayed.

The bottom section of the interface shows the Solr dashboard for the "storagecore1" instance. It includes a sidebar with navigation links (Dashboard, Logging, Core Admin, Java Properties, Thread Dump, Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query, Replication, Schema, Segments info) and a main content area with the following sections:

- Statistics**:
  - Last Modified: 3 months ago
  - Num Docs: 20
  - Max Doc: 20
  - Heap Memory Usage: -1
  - Deleted Docs: 0
  - Version: 175
  - Segment Count: 2
  - Current: ✓
- Instance**:
  - CWD: /opt/solr-8.5.2/server
  - Instance: /var/solr/data/storagecore1
  - Data: /var/solr/data/storagecore1/data
  - Index: /var/solr/data/storagecore1/data/index
  - Impl: org.apache.solr.core.NRTCachingDirectoryFactory
- Healthcheck**:
  - Ping request handler is not configured with a healthcheck file.
- Replication (Master)**:

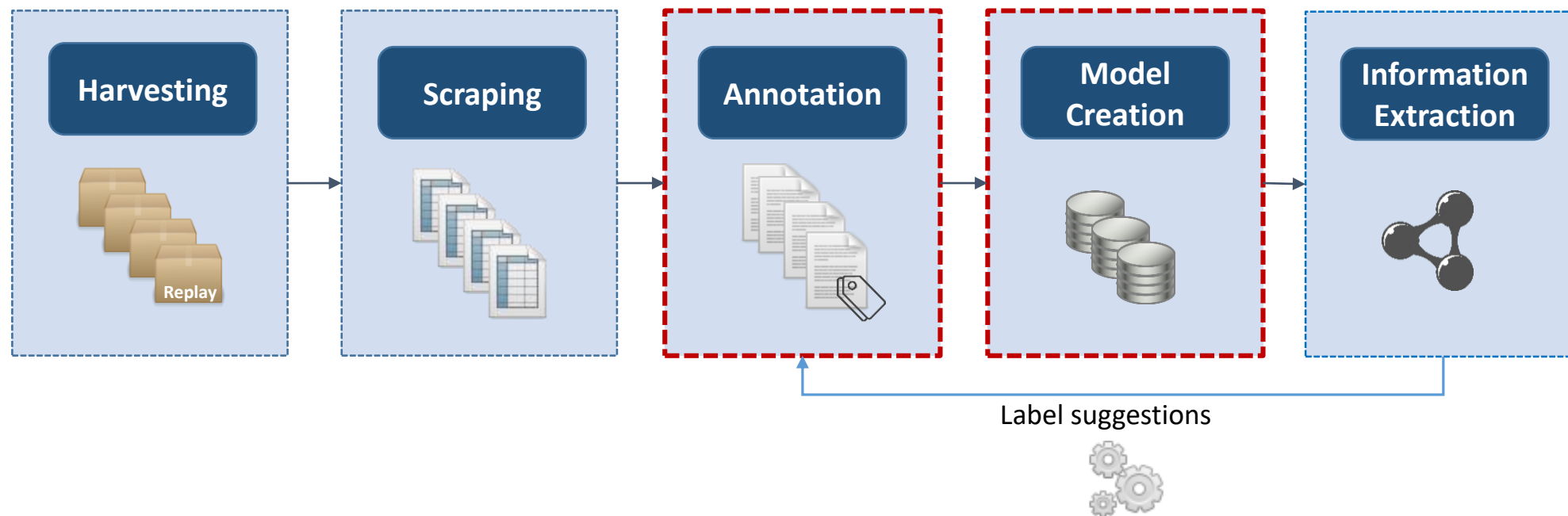
	Version	Gen	Size
Master (Searching)	1584546273240	43	204.96 KB
Master (Replicable)	-	-	-

At the bottom of the page, there are links for Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.



datasets			▼
GET	/datasets/	List information packages (database)	datasets_list 🔒
POST	/datasets/	Create an information package record (database)	datasets_create 🔒
GET	/datasets/representations/	List representations of authenticated user (database)	datasets_representations_list 🔒
GET	/datasets/status/	Status of all submissions (working area)	datasets_status_list 🔒
POST	/datasets/{identifier}/checkout-working-copy/	Checkout a working copy (database, file system)	datasets_checkout-working-copy_create 🔒
GET	/datasets/{identifier}/file-resource/{ip_sub_file_path}/	Retrieve file resource (file system)	datasets_file-resource_read 🔒
GET	/datasets/{identifier}/{entry}/stream/	Read package entry	datasets_stream_list 🔒
GET	/datasets/{process_id}/	Get selected information package (database)	datasets_read 🔒
PUT	/datasets/{process_id}/	Update information package record (database)	datasets_update 🔒
PATCH	/datasets/{process_id}/	Update information package record (database)	datasets_partial_update 🔒
DELETE	/datasets/{process_id}/	Delete registered information package (database)	datasets_delete 🔒
GET	/datasets/{process_id}/dir-json	List directory content as JSON (working area)	datasets_dir-json_list 🔒
GET	/datasets/{process_id}/file-resource/{ip_sub_file_path}/	Retrieve file resource (database, file system)	datasets_file-resource_read 🔒
DELETE	/datasets/{process_id}/file-resource/{ip_sub_file_path}/	Remove file resource (database, working area)	datasets_file-resource_delete 🔒
GET	/datasets/{process_id}/representation/{representation_label}/info/	Get representation ids by label (database, file system)	datasets_representation_info_list 🔒
GET	/datasets/{process_id}/representations/info/	Get representation ids by label (database, file system)	datasets_representations_info_list 🔒
GET	/datasets/{process_id}/status/	Status of selected submission (database, working area)	datasets_status_list 🔒
POST	/datasets/{process_id}/{datatype}/upload/	Upload file to a submission or working copy (database, working area)	datasets_upload_create 🔒

# Process Flow



# **Annotation interface (RecogitoJS integrated)**

# Annotation interface – Loading data



CPKIT Data Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Annotation

dnm-archives-2013-2015-grams-wd **load**

☒ Train ☐ Test ☐ changed

0 - ALL KINDS OF WEAPONS LONG GUNS TO S ^

1 - Model: AK-47 Fixed Stock Caliber: 7.62x..

2 - SWISS QUALITY WEED WHITE RUSSIAN Die

3 - Pistols Gun 6 35 I can put reality photo..

4 - Specifications: Manufacturer: Glock Mo..

5 - This is a custom listing for for L\*\*\*a ..

6 - used cobra .380. Firing conditions, sold..

7 - 8 GR Tutankhamon Genetics AK 47 selectio..

8 - Dinafem Seeds White Widow is the most po..

9 - NEW Glock 19 4th generation Only 1 avail..

10 - New Glock 17gen4 pistol never used i hav..

11 - ALL KINDS OF WEAPONS LONG GUNS TO

12 - Activate your Windows 10 with real key f..

13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..

14 - Cobra Big Bore Derringer .38 Special M..

15 - New in box SilencerCo SAKER 556 5.56 Sup..

16 - Let us introduce AK47 a staple in cannab..

17 - ALL KINDS OF WEAPONS LONG GUNS TO

18 - Taurus Model: 1911 Commander Caliber: ..

19 - Allright this is the first gun i listed ..

20 - This is a custom listing for djammo187 o..

21 - Heavy frame model 96,Crimson Trace Grips..

22 - This listing is for 1,2 or 14 grams of a

Version 1.0 (28.02.2020)

# Load grams weapons/drugs dataset



COPKIT Data Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Annotation

dnm-archives-2013-2015-grams-wd load

☒ Train ☐ Test ☐ changed

0 - ALL KINDS OF WEAPONS LONG GUNS TO S  
1 - Model: AK-47 Fixed Stock Caliber: 7.62x..  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - Specifications: Manufacturer: Glock Mo..  
5 - This is a custom listing for for L\*\*a..  
6 - used cobra .380. Firing conditions, sold..  
7 - 8 GR Tutankhamon Genetics AK 47 selectio..  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO  
12 - Activate your Windows 10 with real key f..  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Cobra Big Bore Derringer .38 Special M..  
15 - New in box SilencerCo SAKER 556 5.56 Sup..  
16 - Let us introduce AK47 a staple in cannab..  
17 - ALL KINDS OF WEAPONS LONG GUNS TO  
18 - Taurus Model: 1911 Commander Caliber:..  
19 - Allright this is the first gun i listed..  
20 - This is a custom listing for djammo187 o..  
21 - Heavy frame model 96,Crimson Trace Grips..  
22 - This listing is for 1.2 oz 14 grams of a

Glock's+d[1,2] <LABEL> apply

Select operation apply

Persist annotations

☒ You are annotating entities Weapons/drugs

LOCATION

Pistols Gun 6 35 I can put reality photo WORLD 3 16 business day Europe 2 10 business day

Store changes Suggest annotations

Version 1.0 (28.02.2020)

# Navigate through annotated postings



CPKIT Data Repository

Administration

Data set creation

Data set management

Knowledge Extraction

Search & Access

Language

admin

## Annotation

dnm-archives-2013-2015-grams-wd load

Train

Test

changed

0 - ALL KINDS OF WEAPONS LONG GUNS TO S  
1 - Model: AK-47 Fixed Stock Caliber: 7.62x..  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - Specifications: Manufacturer: Glock Mo..  
5 - This is a custom listing for L\*\*\*a..  
6 - used cobra .380. Firing conditions, sold..  
7 - 8 GR Tutankhamon Genetics AK 47 selectio  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO  
12 - Activate your Windows 10 with real key..  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Cobra Big Bore Derringer .38 Special M..  
15 - New in box SilencerCo SAKER 556 5.56 Sup..  
16 - Let us introduce AK47 a staple in cannab..  
17 - ALL KINDS OF WEAPONS LONG GUNS TO  
18 - Taurus Model: 1911 Commander Caliber: ..  
19 - Allright this is the first gun i listed..  
20 - This is a custom listing for djammo187 o..  
21 - Heavy frame model 96,Crimson Trace Grips..  
22 - This listing is for 1,2 or 14 grams of a

Glock\'+d{1,2}

<LABEL>

apply

Select operation

apply

Persist annotations

You are annotating entities

Weapons/drugs

8 GR Tutankhamon Genetics AK 47 selection Smell orange citrus super skunk THC 22 33  
Tutankhamon is another cannabis strain that deserves to be on everyones must grow list.  
Christmas tree type plant with a characteristic super stinky skunk smell. Tutankhamon has  
lots of sticky buds with an incredibly high THC level. Sehr starkes WEED super im Geschmack  
und Konsistens. Es ist auch bestens getrocknet.

Store changes

Suggest annotations

Version 1.0 (28.02.2020)



Version 1.0 (28.02.2020)

# Create regex pattern (e.g. Regex101.com)



The screenshot shows the Regex101.com interface. The main area displays the regular expression `mystring\s+[0-9]{2,2}` and the test string `mystring 92fs`. The pattern is highlighted in blue, and the test string is highlighted in light blue. The interface includes a sidebar with options for saving and sharing, flavor selection (PCRE, ECMAScript, Python, Golang), and tools. The right panel provides an explanation of the pattern, match information, and a quick reference for common tokens.

**REGULAR EXPRESSION** 1 match, 11 steps (~0ms)

`mystring\s+[0-9]{2,2}` "gm"

**TEST STRING** SWITCH TO UNIT TESTS

`mystring 92fs`

**EXPLANATION**

- ▼ `mystring\s+[0-9]{2,2}` "gm"  
mystring matches the characters `mystring` literally (case sensitive)
- ▼ `\s+` matches any whitespace character (equal to `[\r\n\t\f\v ]`)  
Quantifier — Matches between **one** and **unlimited** times, as many times as possible, giving back as needed (*greedy*)
- ▼ Match a single character present in the list below  
`[0-9]{2,2}`  
`{2,2}` Quantifier — Matches exactly **2** times  
`[0-9]` a single character in the range between **0** (index 48)

**MATCH INFORMATION**

Match 1

Full match 0-11 mystring 92

**QUICK REFERENCE**

Search reference

- All Tokens
- ★ Common Tokens ✓
- General Tokens
- Anchors
- Meta Sequences
- \* Quantifiers

A single character of: a, b or c `[abc]`  
A character except: a, b or c `^[abc]`  
A character in the range: a-z `[a-z]`  
A character not in the range: ... `^[a-z]`  
A character in the range: ... `[a-zA-Z]`  
Any single character `.`  
Any whitespace character `\s`

# Get named entity suggestions



COPKIT Data Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Annotation

dnm-archives-2013-2015-grams-wd load

Train Test changed

0 - ALL KINDS OF WEAPONS LONG GUNS TO S  
1 - Model: AK-47 Fixed Stock Caliber: 7.62x..  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - Specifications: Manufacturer: Glock Mo..  
5 - This is a custom listing for for L\*\*\*a..  
6 - used cobra .380. Firing conditions, sold..  
7 - 8 GR Tutankhamon Genetics AK 47 selectio..  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO  
12 - Activate your Windows 10 with real key f..  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Cobra Big Bore Derringer .38 Special M..  
15 - New in box SilencerCo SAKER 556 5.56 Sup..  
16 - Let us introduce AK47 a staple in cannab..  
17 - ALL KINDS OF WEAPONS LONG GUNS TO  
18 - Taurus Model: 1911 Commander Caliber: .45 ACP Capacity: 7  
19 - Allright this is the first gun i listed ..  
20 - This is a custom listing for djammo187 o..  
21 - Heavy frame model 96,Crimson Trace Grips..  
22 - This listing is for 1.2 oz 14 grams of a

Glock\s+\d{1,2} <LABEL> apply

Select operation apply

Persist annotations

You are annotating entities Weapons/drugs

Taurus  
Model: 1911 Commander  
Caliber: .45 ACP  
Capacity: 7

Send questions or custom orders via WICKR ID: scottbrown  
(download Wickr Me from play or app store)

Store changes

Suggest annotations

Version 1.0 (28.02.2020)

# Regular expression batch labelling



**COPKIT Data Repository** Administration

Annotation

dnm-archives-2013-2015-grams-wd **load**

☒ Train ☐ Test ☐ changed

0 - WE HAVE A WIDE RANGE OF FIREARMS ,AM  
1 - ALL KINDS OF WEAPONS LONG GUNS TO SI  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - I have a "like new" Glock 23 gen 4 with ..  
5 - Brand: Glock Model: 19 Caliber: 9mm Cond..  
6 - ALL KINDS OF WEAPONS LONG GUNS TO SI  
7 - 8 GR Tutankhamon Genetics AK 47 selectio..  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO !  
12 - Century RI3284-N VSKA AK47 7.62x39mm /  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Hello, are you in search of top quality..  
15 - This is a custom bulk listing for c\*\*\*\*\*..  
16 - Let us introduce AK47 a staple in cannab..  
17 - Good condition used 9mm makarov's Russia..  
18 - AK Pistol with 50 round magazine..  
19 - Allright this is the first gun i listed ..

Glock\s+\d{1,2} WEAPON **apply**

**Confirm batch annotation**

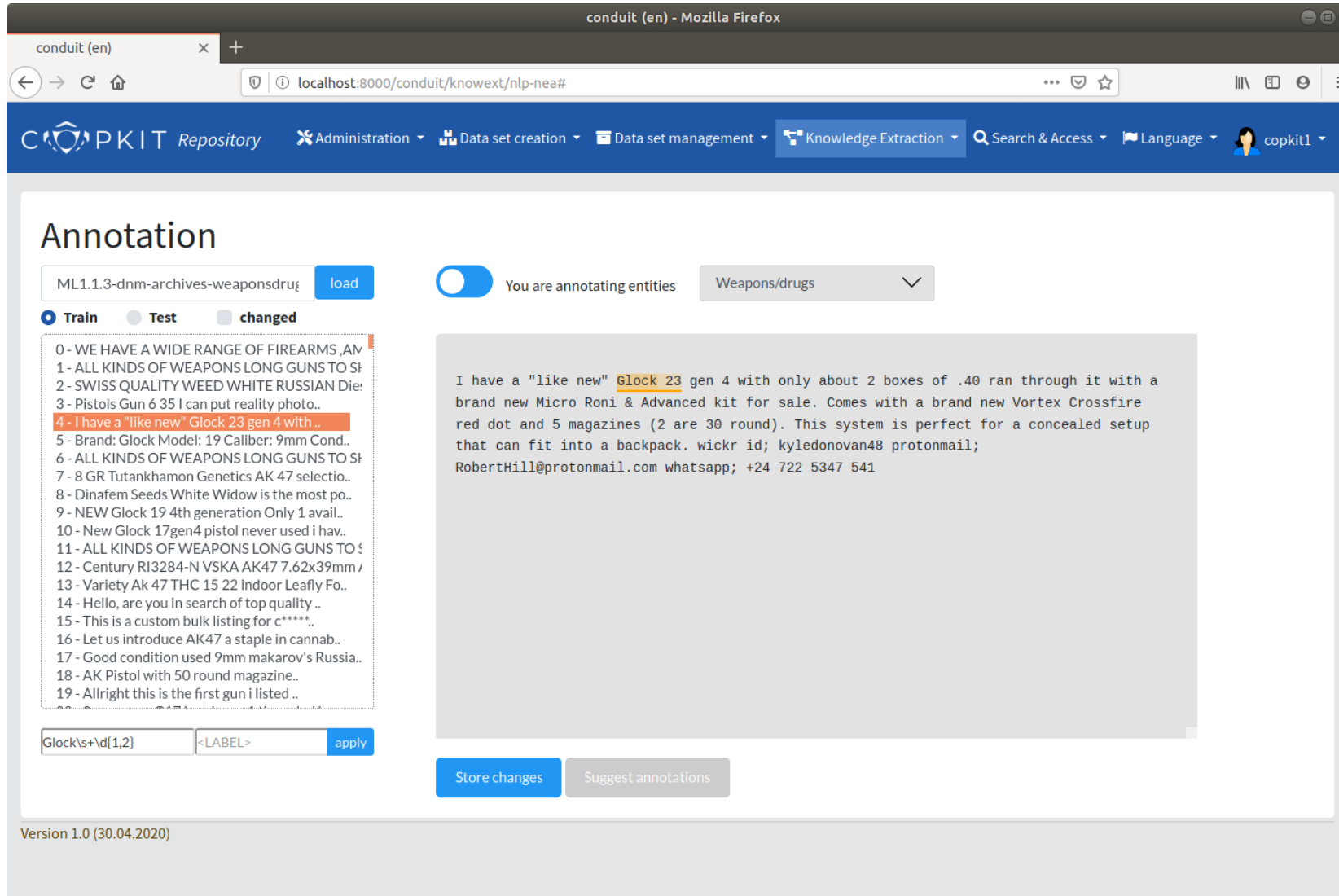
- 4: ... Glock 23 gen 4 with only about 2 boxes ...
- 9: ... Glock 19 4th generation Only 1 availib ...
- 10: ... Glock 17 gen4 pistol never used i have ...
- 24: ... Custom listing for dimit\*\*\*\*\* Glock 80 % + 9mm suppressor If you need ...
- 31: ... Glock 19 , new or used magazines availa ...
- 33: ... Glock 43 , excellent small concealable ...
- 37: ... Glock 23 gen 4 with only about 2 boxes ...
- 43: ... Glock 43 - 9 Round - TLR-6 - Trijicon ...
- 43: ... 125 Rounds Ammo - IWB Holster Glock 43 - 9 Round - TLR-6 - Trijicon ...

**Close** **Apply**

**Store changes** **Suggest annotations**

Version 1.0 (28.02.2020)

# Multi-user



conduit (en) - Mozilla Firefox

conduit (en) x +

localhost:8000/conduit/knowext/nlp-nea#

COPKIT Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language copkit1

## Annotation

ML1.1.3-dnm-archives-weaponsdrug load

☒ Train ☐ Test ☐ changed

0 - WE HAVE A WIDE RANGE OF FIREARMS, AN  
1 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - I have a "like new" Glock 23 gen 4 with ..  
5 - Brand: Glock Model: 19 Caliber: 9mm Cond..  
6 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
7 - 8 GR Tutankhamon Genetics AK 47 selectio..  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
12 - Century RI3284-N VSKA AK47 7.62x39mm /  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Hello, are you in search of top quality..  
15 - This is a custom bulk listing for c\*\*\*\*\*..  
16 - Let us introduce AK47 a staple in cannab..  
17 - Good condition used 9mm makarov's Russia..  
18 - AK Pistol with 50 round magazine..  
19 - Allright this is the first gun i listed ..

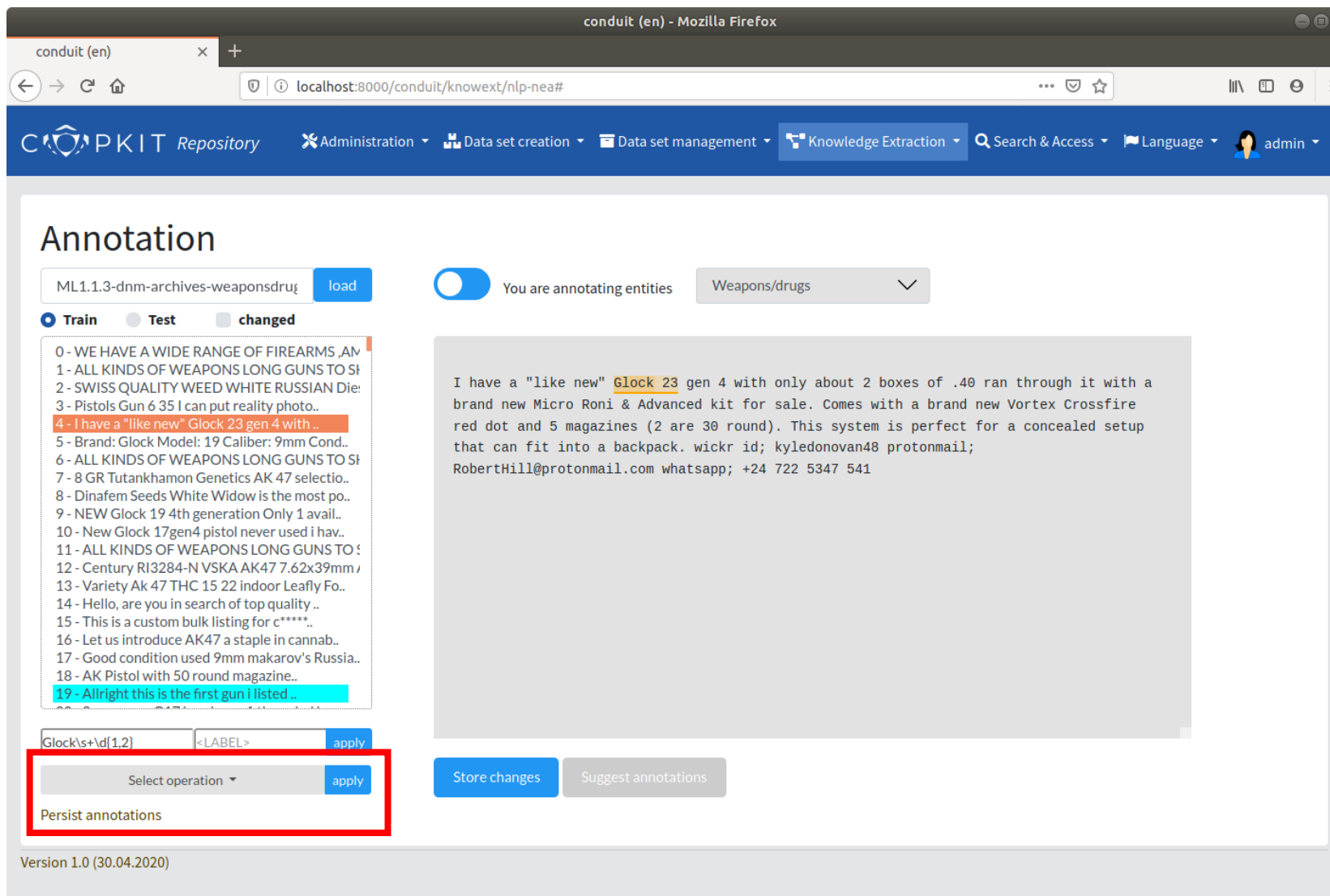
Glock\s+\d{1,2} <LABEL> apply

☒ You are annotating entities Weapons/drugs

I have a "like new" Glock 23 gen 4 with only about 2 boxes of .40 ran through it with a brand new Micro Roni & Advanced kit for sale. Comes with a brand new Vortex Crossfire red dot and 5 magazines (2 are 30 round). This system is perfect for a concealed setup that can fit into a backpack. wickr id; kyledonovan48 protonmail; RobertHill@protonmail.com whatsapp; +24 722 5347 541

Store changes Suggest annotations

Version 1.0 (30.04.2020)



conduit (en) - Mozilla Firefox

conduit (en) x +

localhost:8000/conduit/knowext/nlp-nea#

CPKIT Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Annotation

ML1.1.3-dnm-archives-weaponsdrug load

☒ Train ☐ Test ☐ changed

0 - WE HAVE A WIDE RANGE OF FIREARMS ,AN  
1 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
2 - SWISS QUALITY WEED WHITE RUSSIAN Die  
3 - Pistols Gun 6 35 I can put reality photo..  
4 - I have a "like new" Glock 23 gen 4 with ..  
5 - Brand: Glock Model: 19 Caliber: 9mm Cond..  
6 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
7 - 8 GR Tutankhamon Genetics AK 47 selectio..  
8 - Dinafem Seeds White Widow is the most po..  
9 - NEW Glock 19 4th generation Only 1 avail..  
10 - New Glock 17gen4 pistol never used i hav..  
11 - ALL KINDS OF WEAPONS LONG GUNS TO SH  
12 - Century RI3284-N VSKA AK47 7.62x39mm /  
13 - Variety Ak 47 THC 15 22 indoor Leafly Fo..  
14 - Hello, are you in search of top quality ..  
15 - This is a custom bulk listing for c\*\*\*\*\*..  
16 - Let us introduce AK47 a staple in cannab..  
17 - Good condition used 9mm makarov's Russia..  
18 - AK Pistol with 50 round magazine..  
19 - Allright this is the first gun i listed ..

<LABEL> apply

Select operation apply

Persist annotations

You are annotating entities Weapons/drugs

I have a "like new" Glock 23 gen 4 with only about 2 boxes of .40 ran through it with a brand new Micro Roni & Advanced kit for sale. Comes with a brand new Vortex Crossfire red dot and 5 magazines (2 are 30 round). This system is perfect for a concealed setup that can fit into a backpack. wickr id; kyledonovan48 protonmail; RobertHill@protonmail.com whatsapp; +24 722 5347 541

Store changes Suggest annotations

Version 1.0 (30.04.2020)

# Definition-based batch labelling



- Useful for initial labelling (before user annotation starts)
- Definition file in JSON format required
- Regex terms can be used instead of fixed strings to match tokens
- Filename convention: prelabel\_definitions\_\*.json

```
{
  "ACCOUNT": ["account", "address", "brute force", "brute force", "username"],
  "BOTNET_DDOS": ["botnet", "booter", "ddos", "denial of service"],
  "CREDIT_CARD": ["creditcard", "carding", "ccv", "fullz", "skimmer", ...],
  ...
  "VULNERABILITY": ["vulnerability", "exploit", "code signing", "windows shell"]
}
```

```
▼ ACCOUNT:
  0: "account"
  1: "address"
  2: "brute force"
  3: "brute force"
  4: "username"
▼ BOTNET_DDOS:
  0: "botnet"
  1: "booter"
  2: "ddos"
  3: "denial of service"
▼ CREDIT_CARD:
  0: "creditcard"
  1: "carding"
  2: "ccv"
  3: "fullz"
  4: "skimmer"
  5: "credit"
  6: "cvv"
  7: "fulls"
  8: "mastercard"
  9: "how to card"
  10: "credit card"
▼ PERSONAL_DOCUMENT:
  0: "Identity documents"
  1: "documents"
  2: "passport"
  3: "document"
  4: "drivers license"
  5: "driving license"
  6: "credit card statement"
  7: "electricity bill"
```

# Definition-based batch labelling



Repository Administration Data set creation Data set management Knowledge Extraction Search & Access Language admin

## Annotation

ML1.1.3-dnm-archives-caas-subset load

☒ Train ☐ Test ☐ changed

0 - USE PGP AT ALL TIMES This listing is for..  
1 - This listing is for 2 operators manuals ..  
2 - Use this program to create custom IDs yo..  
3 - The Antminer U3 as described below by it..  
4 - Anonymously Sourced Sim Cards ONLY Just..  
5 - MediSoft version 17 is used by students..  
6 - Digital download sent via PM THE LATEST..  
7 - Digital download sent via PM The Essenti..  
8 - A hard to find cult classic is now back..  
9 - digital download sent via PM The single..  
10 - Digital download sent via PM For more th..  
11 - Digital Download sent via PM Ever though..  
12 - Digital download sent via PM Men s Fitne..  
13 - Digital download sent via PM Included ar..  
14 - Digital download sent via PM What you ha..  
15 - A dazzling work of personal travelogue a..  
16 - Digital download sent via PM Discover wh..  
17 - Digital download sent via PM Johnny Long..  
18 - Digital download sent via PM The first b..  
19 - The world s most infamous hacker offers ..

Glock\s+\d{1,2} <LABEL> apply

Select operation apply

Persist annotations

☒ You are annotating entities Weapons/drugs

This listing is for 2 operators manuals for two different models of ATMs that are made by the same country These ATMs are sold all over the world but since I have never left the USA I can not guarantee that they are in your country as I have not been there to see them myself in person I do know they are in all 50 states even the shitty little town I grew up in The manuals contain the default master passcodes that if they remain unchanged by the installer most do for convenience then you can log in as the programmer and change the denomination of the bills that the machine believes it contains So for instance if you change the denomination 1 and pull out 20 then you will get 20 20 bills because it thinks it is spitting out 1 bills instead of 20 bills and 20 x 1 20 to the machine but you really get 200 for 20 Hope fully that makes sense If you buy this please have a pgp key posted so that I can encrypt the download instructions If you do not have one posted I will not complete the transaction This is not a hard thing to do nor do I think it is too much to ask

Store changes Suggest annotations

Version 1.0 (30.04.2020)

Two basic functions for handling train/test data are offered: 1) merge test into train, 2) split train into train and test



1.) Merge test set into training set  
→ Recommended before starting user annotation



2.) Split training set into training and test set  
→ Recommended before starting model creation



# Merge test set into training set



CPKIT Repository

Administration

Edge Extraction

Search & Access

Language

admin

## Annotation

ML1.1.3-dnm-archives-caas-subset

load

☒ Train ☐ Test ☐ changed

0 - USE PGP AT ALL TIMES This listing is for..  
1 - This listing is for 2 operators manuals ..  
2 - Use this program to create custom IDs yo..  
3 - The Antminer U3 as described below by it..  
4 - Anonymously Sourced Sim Cards ONLY Just..  
5 - MediSoft version 17 is used by students..  
6 - Digital download sent via PM THE LATEST..  
7 - Digital download sent via PM The Essenti..  
8 - A hard to find cult classic is now back..  
9 - digital download sent via PM The single..  
10 - Digital download sent via PM For more th..  
11 - Digital Download sent via PM Ever though..  
12 - Digital download sent via PM Men s Fitne..  
13 - Digital download sent via PM Included ar..  
14 - Digital download sent via PM What you ha..  
15 - A dazzling work of personal travelogue a..  
16 - Digital download sent via PM Discover wh..  
17 - Digital download sent via PM Johnny Long..  
18 - Digital download sent via PM The first b..  
19 - The world s most infamous hacker offers ..

A hard to find cult classic is now back ..

related Alchemist D Gold  
rets and modern  
ous editions sold 100 000  
erculture

Glock\s+\d{1,2} <LABEL> apply

Merge test into train apply

Persist annotations

Store changes

Suggest annotations

Version 1.0 (30.04.2020)

### Merge test into train set

This operation will merge all records from the test data set into the train data set. A backup file for the existing training and test data set will be created.

Close Apply

# Split training set into training and test set



CPKIT Repository

Administration

Annotation

ML1.1.3-dnm-archives-caas-subset

load

☒ Train ☐ Test ☐ changed

0 - USE PGP AT ALL TIMES This listing is for..  
1 - This listing is for 2 operators manuals..  
2 - Use this program to create custom IDs yo..  
3 - The Antminer U3 as described below by it..  
4 - Anonymously Sourced Sim Cards ONLY Just..  
5 - MediSoft version 17 is used by students..  
6 - Digital download sent via PM THE LATEST..  
7 - Digital download sent via PM The Essenti..  
8 - A hard to find cult classic is now back..  
9 - digital download sent via PM The single..  
10 - Digital download sent via PM For more th..  
11 - Digital Download sent via PM Ever though..  
12 - Digital download sent via PM Men s Fitne..  
13 - Digital download sent via PM Included ar..  
14 - Digital download sent via PM What you ha..  
15 - A dazzling work of personal travelogue a..  
16 - Digital download sent via PM Discover wh..  
17 - Digital download sent via PM Johnny Long..  
18 - Digital download sent via PM The first b..  
19 - The world s most infamous hacker offers ..

Glock\s\d{1,2} <LABEL> apply

Split train into train and test apply

Persist annotations

Version 1.0 (30.04.2020)

Select train/test distribution

Current training dataset

Number of rows in current training set: 584

Split training dataset

(80 % / 20 %)

Distribution after splitting:

- Number of rows in new training set: 467
- Number of rows in new test set: 117

Close Split

Store changes Suggest annotations

# Restore backups



CPKIT Repository

Administr

Edge Extraction

Search & Access

Language

admin

## Annotation

ML1.1.3-dnm-archives-caas-subset **load**

**Train** **Test** **changed**

0 - USE PGP AT ALL TIMES This listing is for..  
1 - This listing is for 2 operators manuals..  
2 - Use this program to create custom IDs yo..  
3 - The Antminer U3 as described below by it..  
4 - Anonymously Sourced Sim Cards ONLY Just..  
5 - MediSoft version 17 is used by students..  
6 - Digital download sent via PM THE LATEST..  
7 - Digital download sent via PM The Essenti..  
8 - A hard to find cult classic is now back..  
9 - digital download sent via PM The single..  
10 - Digital download sent via PM For more th..  
11 - Digital Download sent via PM Ever though..  
12 - Digital download sent via PM Men s Fitne..  
13 - Digital download sent via PM Included ar..  
14 - Digital download sent via PM What you ha..  
15 - A dazzling work of personal travelogue a..  
16 - Digital download sent via PM Discover wh..  
17 - Digital download sent via PM Johnny Long..  
18 - Digital download sent via PM The first b..  
19 - The world s most infamous hacker offers ..

Glock's+\\d{1,2} <LABEL> **apply**

Restore backup **apply**

Persist annotations

Version 1.0 (30.04.2020)

**Select backup**

2020-06-18T17:56:23\_train


Selected backup files:

- 2020-06-19T18:23:53\_test
- 2020-06-18T17:56:23\_train

**Close** **Restore**

**Store changes** **Suggest annotations**

- Update training/test data files with user annotations (user annotations are removed afterwards)

Administration ▾ Data set creation ▾ Data set management ▾ Knowledge Extraction ▾ Search & Access ▾ Language ▾ admin ▾

## Persist annotations

Data set	ML1.1.3-dnm-archives-weaponsdrugs-subset
Number of annotated documents	10

**Subset type:**

- ☐ train
- ☐ test

Submit

Version 1.0 (30.04.2020)

# Technical background: Named entity recognition

- Evaluate whether pre-trained embeddings improve the classification performance over simple vector space models, such as TF-IDF, when being transferred to the darknet domain.
- Best performance was achieved with Tensorflow Universal Sentence Encoder and a linear SVM

Heistracher, C. & Schlarb, S.

Machine Learning Techniques for the Classification of Product Descriptions from Darknet Marketplaces

Proceedings of the 11th International Conference on Applied Informatics, 2020

- SpaCy is an open-source library for NLP
- SpaCy's standard models can be used as a basis to build domain-specific models
- For example, SpaCy's "en\_core\_web\_lg" is a convolutional neural network trained for multiple tasks on the OntoNotes 5 dataset.
- 18 entity types that are available in the pretrained model.
- Words are represented as *GloVe* vectors, that were trained on the common crawl dataset.

Ghost **Glock 17 WEAPON** gen **3 CARDINAL** Info comes from the same factory only the ghost **glock WEAPON** is not registered the **glock WEAPON**  
is rather removed from the factory so that it can not be traced **9mm CALIBRE** weapon **2 CARDINAL** x magazine **50 CARDINAL** x **9mm CALIBRE**  
**ammunition WEAPON** included Sent only within **Europe LOC** I only trade on **Alphabay ORG** please contact if you want to buy Fixed price price for  
bulk deals is negotiable Dispatched from **Netherlands GPE**



- Dataset used for the evaluation was a subset of (Gwern 2015) that contained strings from a list of weapon related terms.
- This subset was manually categorized and the categories were used as labels related to weapons.
- The subset of the data contained 1580 product descriptions including: 1582 Drugs, 769 Weapons, 319 ebooks, 96 3D-printing.

Precision	76.97
Recall	70.88

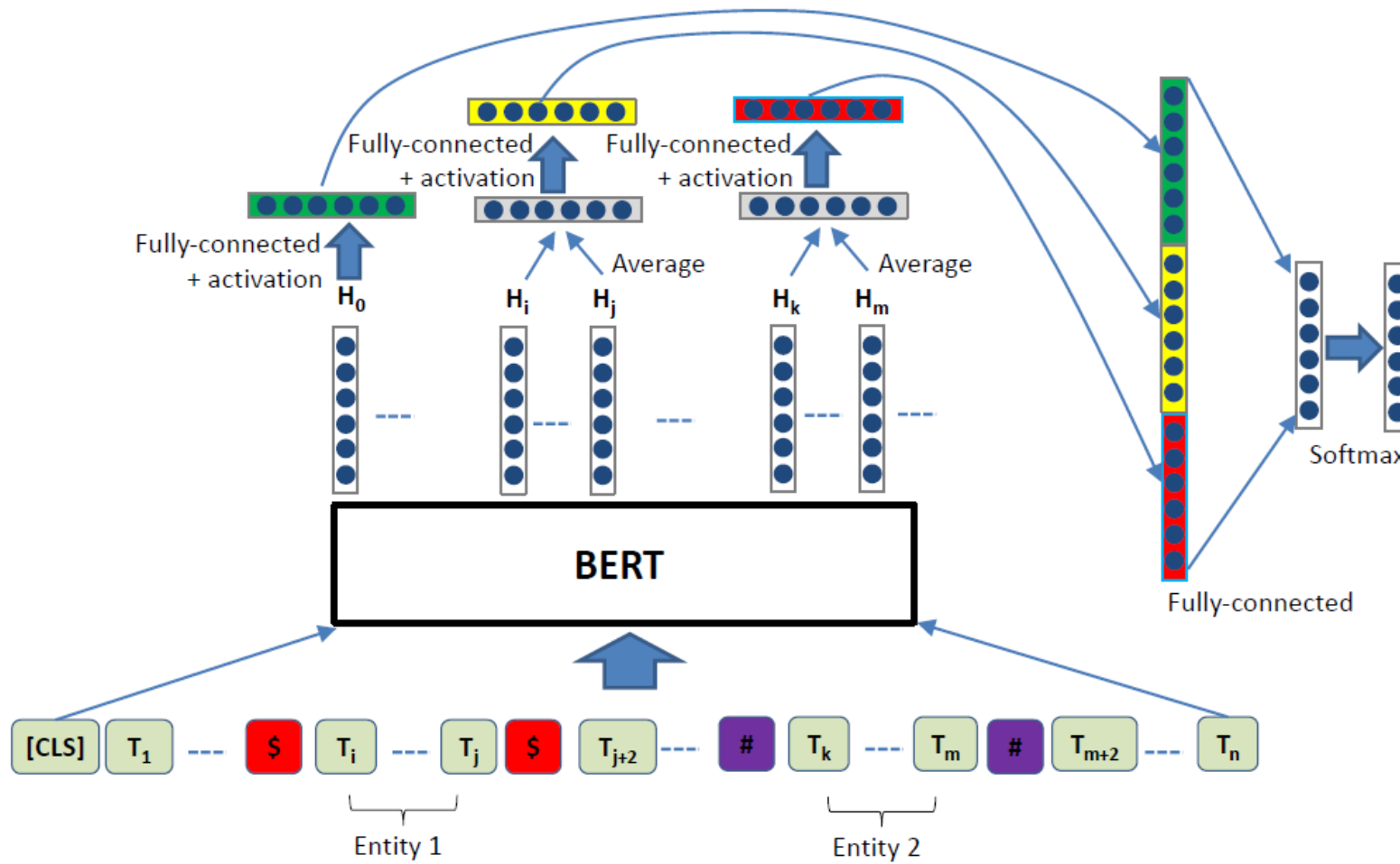
# Technical background: Relation extraction

- **Focus on Transformer models**

- Can be pre-trained on smaller datasets
- Model can be used for NLP downstream tasks
- Several papers already published with BERT, ALBERT, RoBERTA, GPT..
- Mostly trained and evaluated on SemEval 2010 Task 8 & TACRED datasets
- Various approaches
  - Most commonly used: sentence-level
  - Semantic Role Labeling
  - Joint Approaches (NER + RE)
  - Multi-Relation-Extraction

- Entities are already annotated
- No ground truth for relations for our dataset annotated
  1. Unsupervised relation extraction
  2. Using an already labeled dataset with pre-defined labels such as SemEval 2010 Task 8
- Two models are fitting for our approach:
  - R-BERT
    - BERT, RoBERTa and ALBERT are available
    - Sentence-level
  - OpenNRE
    - BERT
    - Sentence-, bag-, document-level and few-shot

# Relation extraction – R-BERT



## NER

*Ernest Miller Hemingway was an American journalist*  $\Rightarrow$  *Ernest Miller Hemingway was an American journalist*

B I E O O S S

## Sentence-level RE

*Ernest Hemingway was raised in Oak Park, Illinois*  $\Rightarrow$  *[Ernest Hemingway]*  $\xrightarrow{\text{place of birth}}$  *[Oak Park, Illinois]*

## Bag-level RE

In 1921, *Ernest Hemingway* married *Hadley Richardson*, the first of his four wives

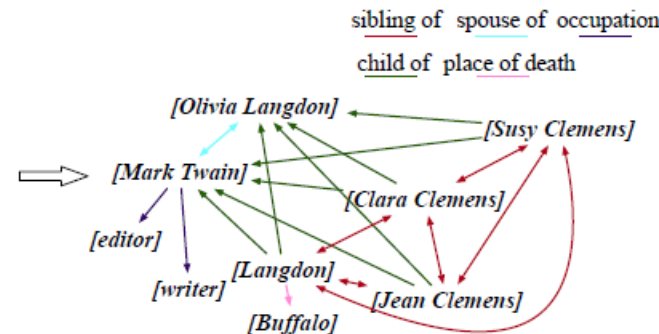
*Hadley Richardson* was the first wife of American author *Ernest Hemingway*

... ...

$\Rightarrow$  *[Ernest Hemingway]*  $\xrightarrow{\text{spouse}}$  *[Hadley Richardson]*

## Document-level RE

*Mark Twain* and *Olivia Langdon* corresponded throughout 1868. She rejected his first marriage proposal, but they were married in Elmira, New York in February 1870. Then, Twain owned a stake in the Buffalo Express newspaper and worked as an *editor* and *writer*. While they were living in *Buffalo*, their son *Langdon* died of diphtheria at the age of 19 months. They had three daughters: *Susy Clemens*, *Clara Clemens*, and *Jean Clemens*.



## Few-shot RE

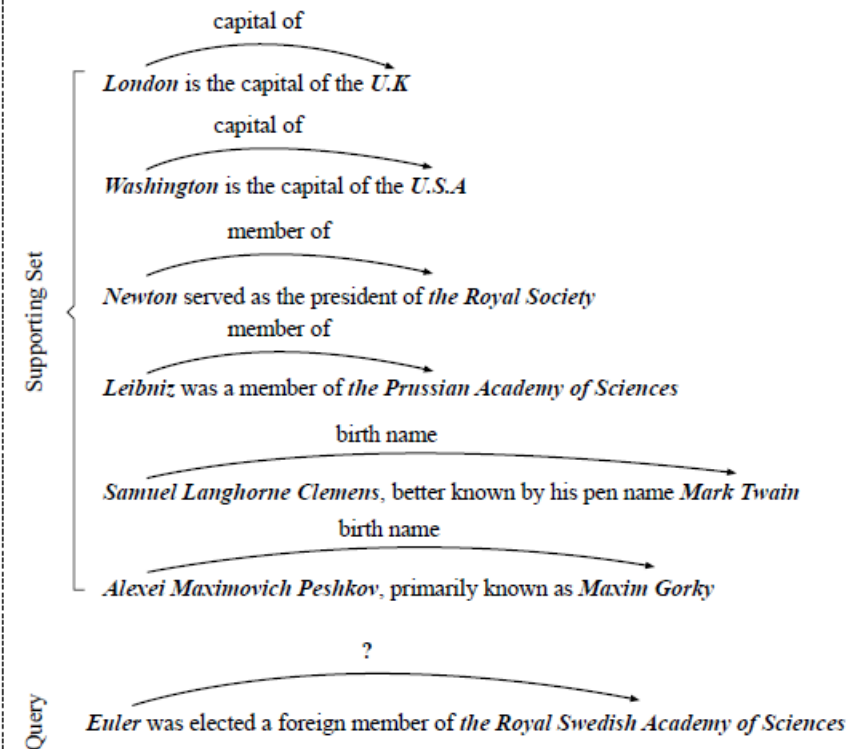


Figure 1: The examples of all application scenarios in OpenNRE.

- Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 169–174, Hong Kong, China, November. Association for Computational Linguistics. <https://github.com/thunlp/OpenNRE>
- Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. <https://github.com/monologg/R-BERT>

# Thank you !

---

# IBM's Moment Recognizer (MoRec)

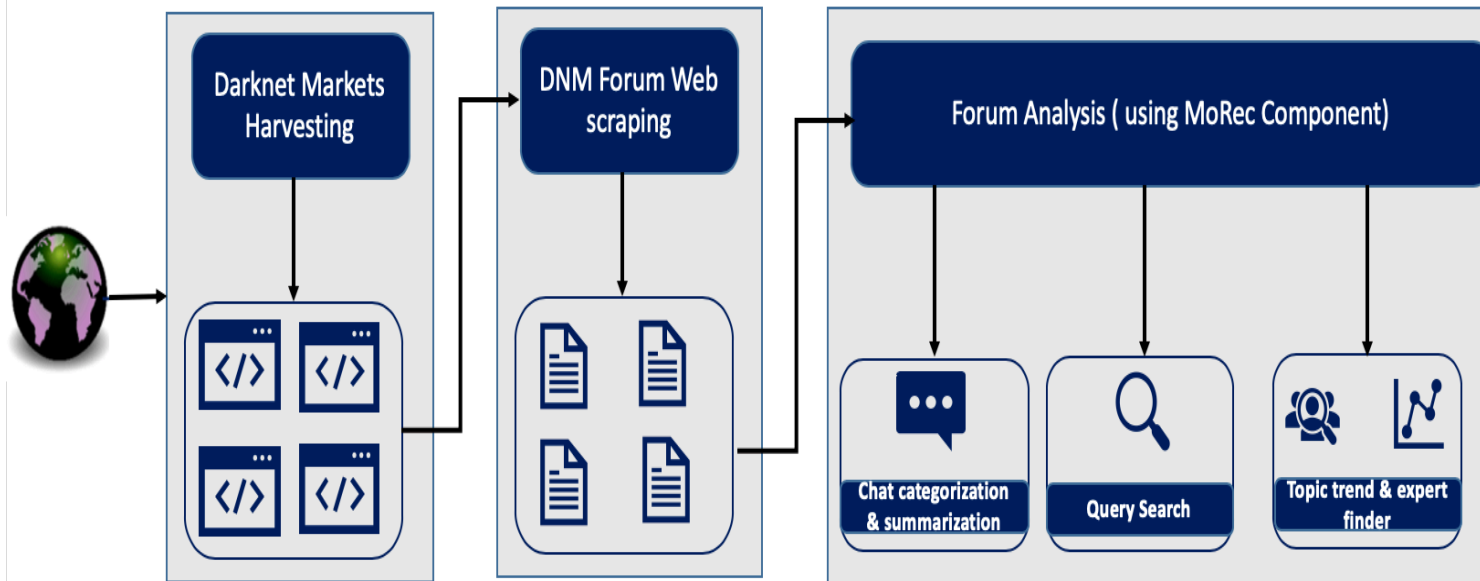
Faisal Ghaffar (IBM), Brian Maguire (IBM)

Wednesday 19<sup>th</sup> August 2020

---

## 2. Component Overview

- AS an (LEA) end-user, I would like to extract knowledge from DNM forums to
  - understand thread discussions
  - determine topics of discussion
  - determine trends in topics over time
  - identify experts on specific topic(s)



hi i don't have a headphone switch in the volume control, it means that the headphones on my laptop and the speakers work at the same time. I didn't have this problem with ubuntu 8.10. I really hope there is a solution to the problem because the many threads i've seen on the subject didn't have a solution, thanks

anyone????? i've posted it a while ago but still have not found an answer

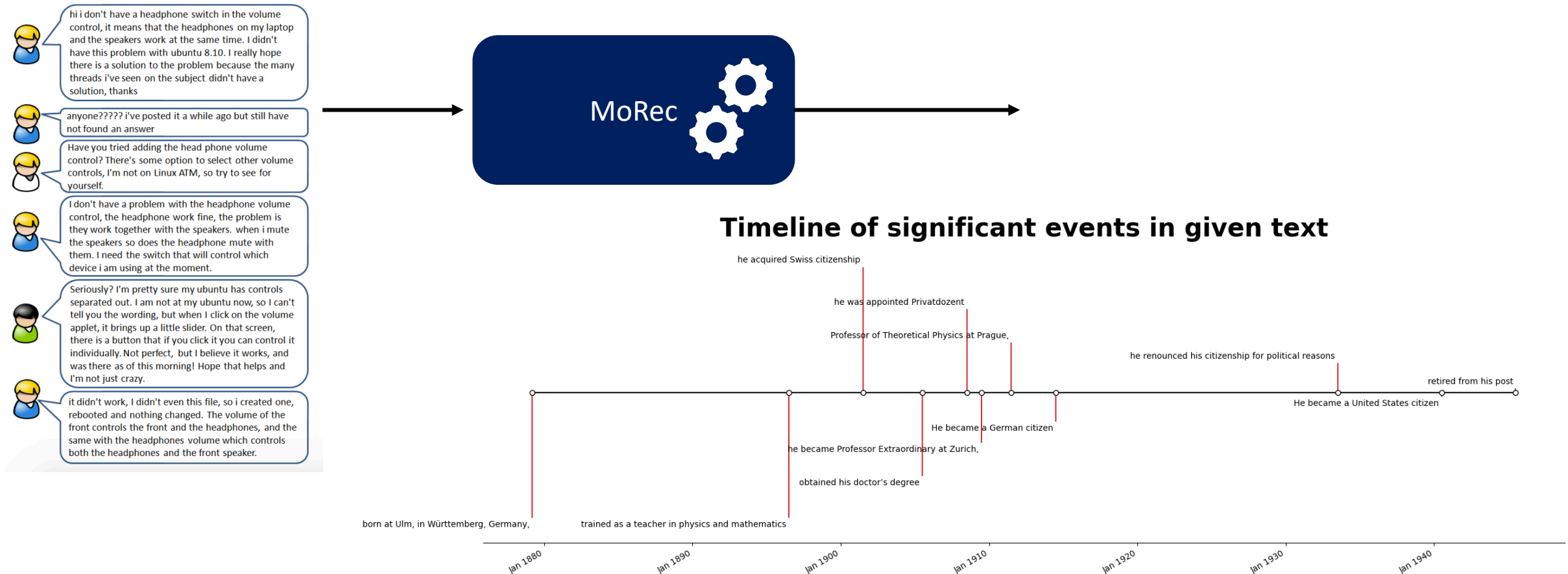
Have you tried adding the head phone volume control? There's some option to select other volume controls, I'm not on Linux ATM, so try to see for yourself.

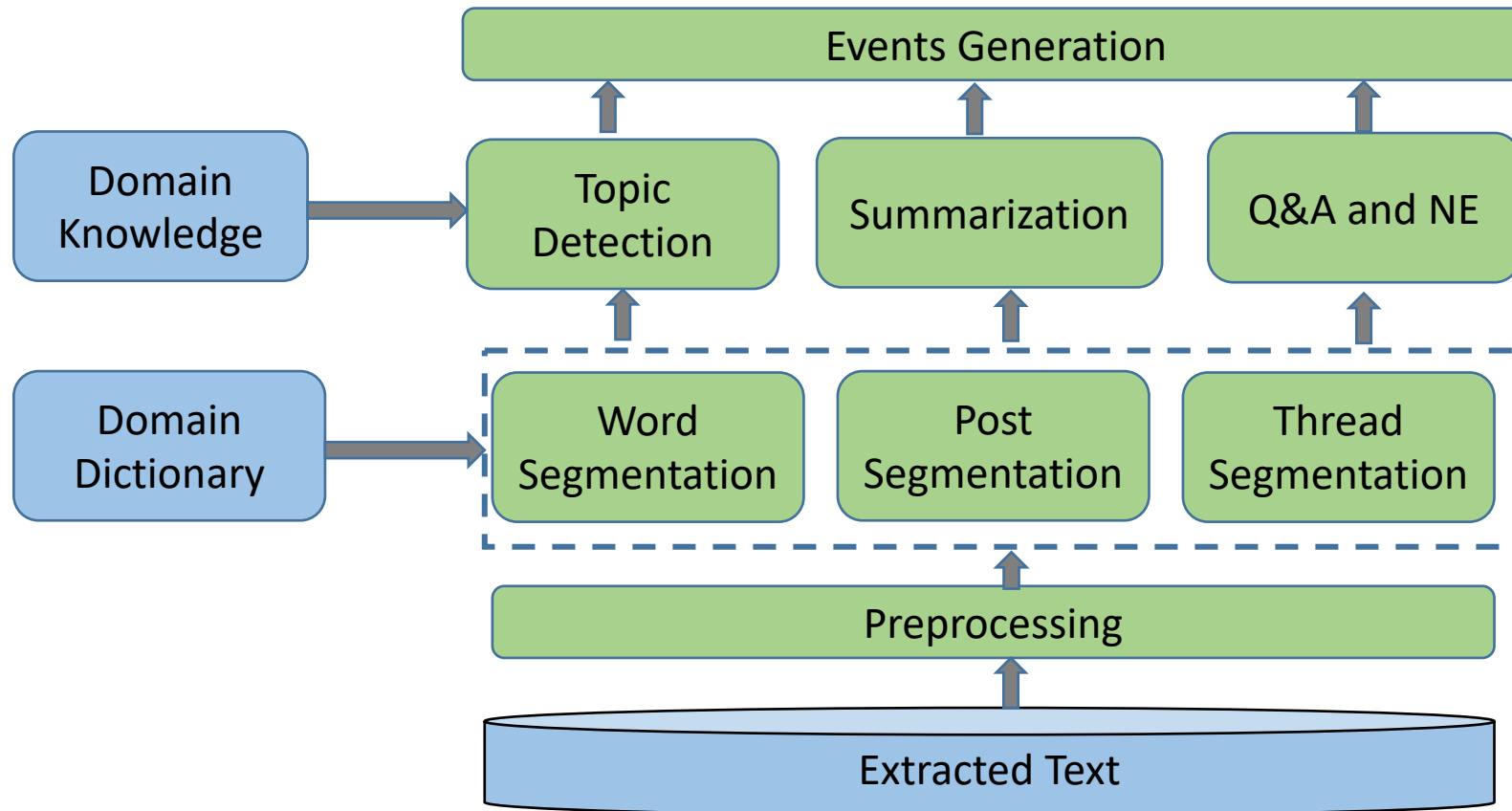
I don't have a problem with the headphone volume control, the headphone work fine, the problem is they work together with the speakers. when i mute the speakers so does the headphone mute with them. I need the switch that will control which device i am using at the moment.

Seriously? I'm pretty sure my ubuntu has controls separated out. I am not at my ubuntu now, so I can't tell you the wording, but when I click on the volume applet, it brings up a little slider. On that screen, there is a button that if you click it you can control it individually. Not perfect, but I believe it works, and was there as of this morning! Hope that helps and I'm not just crazy.

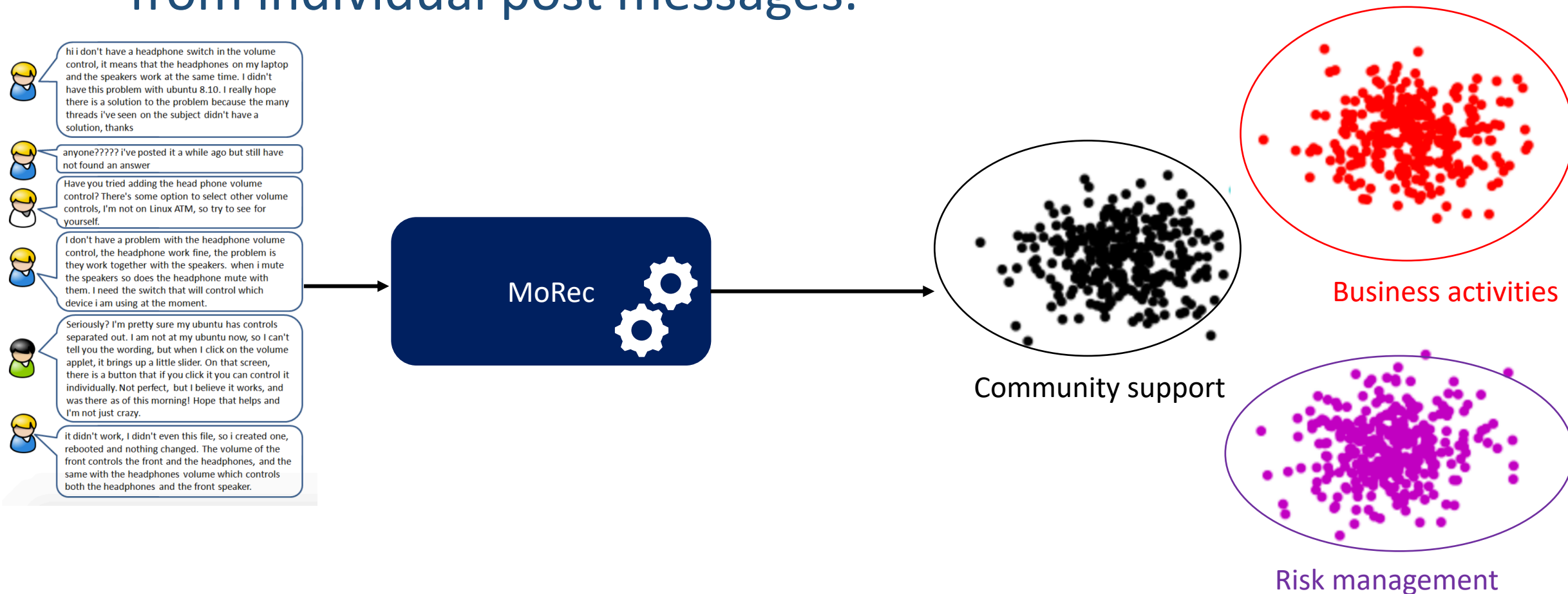
it didn't work, I didn't even this file, so i created one, rebooted and nothing changed. The volume of the front controls the front and the headphones, and the same with the headphones volume which controls both the headphones and the front speaker.

## Given a thread, extract who said what over a daily timeline

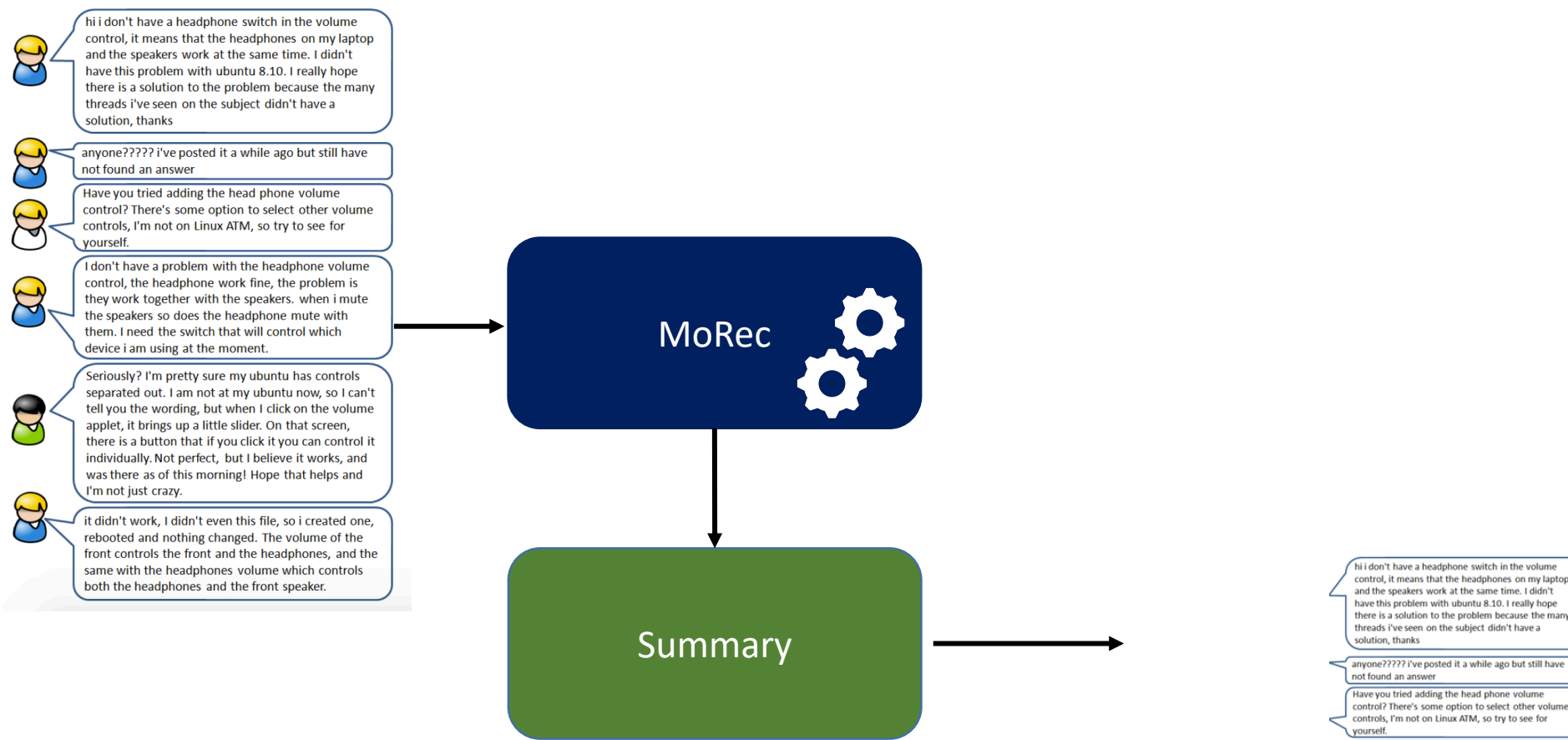




## 1. Given a darknet forum, determine topics of discussion from individual post messages.



## Provide 'extractive' and 'abstractive' summary



Given a forum, retrieve answers to questions related to text

## Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

## Question

What causes precipitation to fall?

## Answer Candidate

gravity

- Between question and answer

cause---gravity

precipitation---gravity

fall---gravity

what---gravity

Example: question-answer from a context text. Image Source: <https://rajpurkar.github.io/mlx/qa-and-squad/>

## RoBERTa

- Based on BERT with following differences
- Dynamic word-masking instead of static
- Removed NSP training requirement
- Trained on 8k sequences compared to 256 of BERT

## Evaluation Metric

- Label Ranking Average Precision
- Final Model : {'LRAP': 0.967, 'eval loss': 0.185}

- Multi-Label Categorization based on trained LM
  - Is forum post belong to CaaS and carding, hacking, trafficking, drugs
- Trained Language models (LM) (RoBERTa ) on dark net data
- Summarization updated based on LM
- Topic detection based on embeddings
- similarity search function
- Timeline creation from forum threads
- Question-answering (TBD)
- Trade Network extraction and analysis (TBD)

## COPKIT MoRec APIs 0.1.2 OAS3

</openapi.json>

MoRec component APIs for NLP on DarkNet Forums.

### default



GET

/ Root

POST

/api/token\_tagger Token Tagger

POST

/api/label\_posts Label Post

POST

/api/tag-all-labels Token Tagger And Sequence Classifier

POST

/api/text-summarization Text Summary

GET

/api/extract-metadata Extract Metadata

POST

/api/plot\_events\_timeline Plot Timeline

**Thank you !**  
**Time for Questions and  
Discussion**

---