

# IPTC EXTRA User Manual

Version 1.0 - 2017/07/14

<b>Overview</b>	<b>4</b>
<b>Taxonomy Manager</b>	<b>4</b>
Edit an Existing Taxonomy	5
Create a New Taxonomy	6
<b>Schemas Manager</b>	<b>8</b>
<b>Rule Editor &amp; EXTRA Query Language (EQL)</b>	<b>9</b>
EXTRA Query Language	9
Extended Operators	10
Search Clauses	12
Relations	12
Relation Modifiers	13
Boolean Operators	14
Boolean Modifiers	16
Text_Content	18
Rule Editor	18
Open the Rule Editor	18
Open a Rule	20
Test a Rule	21
Edit a Rule	22
Create a New Rule	24
Submit a Rule	25
<b>Document Search Tool</b>	<b>27</b>
<b>Document Tagging</b>	<b>27</b>
<b>Corpora</b>	<b>28</b>
<b>Logging Issues</b>	<b>28</b>
<b>Appendix: Sample Rules</b>	<b>29</b>
IPTC Media Topics - English	29
Advertising	29
Antitrust Issue	30
Bankruptcy	30
Cancer	30

Censorship	31
Children	32
Christianity	34
Cinema	35
Cycling	35
Data Protection Policy	36
Diseases and Conditions	36
Drug Trafficking	37
Economic Policy	38
Education	38
Employment Statistics	39
Fashion	39
Fire	40
Gays and Lesbians	40
Golf	41
Health Insurance	41
Homicide	42
Immigration	43
Islam	44
Medical Research	44
Music	45
National or Ethnic Minority	46
Nuclear Power	47
Rape	48
Real Estate	48
Retirement	49
Soccer	49
Summit Meetings	51
Taxation	52
Television	52
Tourism	53
IPTC Media Topics - German	54
Aktien (Stocks)	54
Arbeitsmarktdaten (Employment statistics)	54
Auszeichnung und Preis (Awards)	54
Betrug (Fraud)	55
Computerkriminalitaet (Computer Crime)	56
Datenschutz (Data Protection)	57
Drogenhandel (Drug trafficking)	57

Einwanderung (Immigration)	58
Essen und Trinken (Food)	58
Feuer (Fire)	59
Film (Cinema)	59
Fußball (Soccer)	59
Gewerkschaft (Unions)	60
Golf	60
Homosexualität (Gays and lesbians)	61
Immobilien (Real Estate)	61
Insolvenz (Bankruptcy)	63
Islam	63
Kartellfragen (Antitrust Issue)	64
Leichtathletik (Athletics)	64
Medizinische Forschung (Medical Research)	65
Militärische Ausrüstung (Military Equipment)	66
Mode (Fashion)	66
Parlamentswahlen (National Elections)	67
Pharmaindustrie (Pharmaceutical)	67
Radsport (Cycling)	68
Rente (Pension)	69
Sexuelle Gewalt (Sexual Assault)	69
Staatshaushalt (Government Budget)	70
Staatsverschuldung (Government Debt)	71
Steuer (Taxation)	71
Terroranschlag (Act of Terror)	73
Umwelt (Environment)	73
Universität (University)	74
Zusatzversicherung (Health Insurance)	75

# Overview

The IPTC EXTRA Toolkit is comprised of five tools, accessible from the menu in the upper lefthand corner (see Figure 1):

1. **Taxonomy Manager:** To define each taxonomy, or collection of rules, that will be called by the engine.
2. **Schemas Manager:** To define the schema of fields in your corpus that will be indexed, in order to write rules against them.
3. **Rules Editor:** To build the rules using the EXTRA Query Language (EQL)
4. **Document Search Tool:** To search or navigate (by topics), the indexed documents of your corpora.
5. **Document Tagging:** To submit individual documents to the rules engine and view the rule matches that are returned.

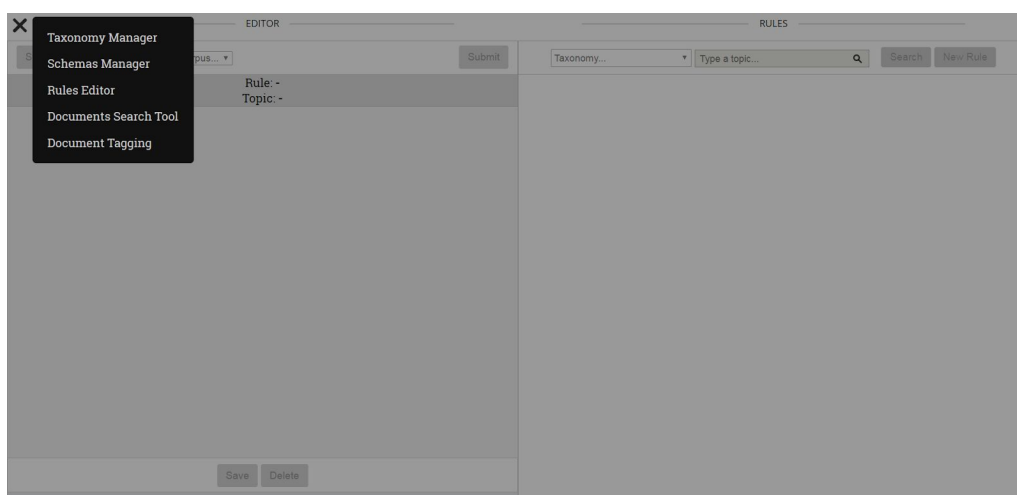


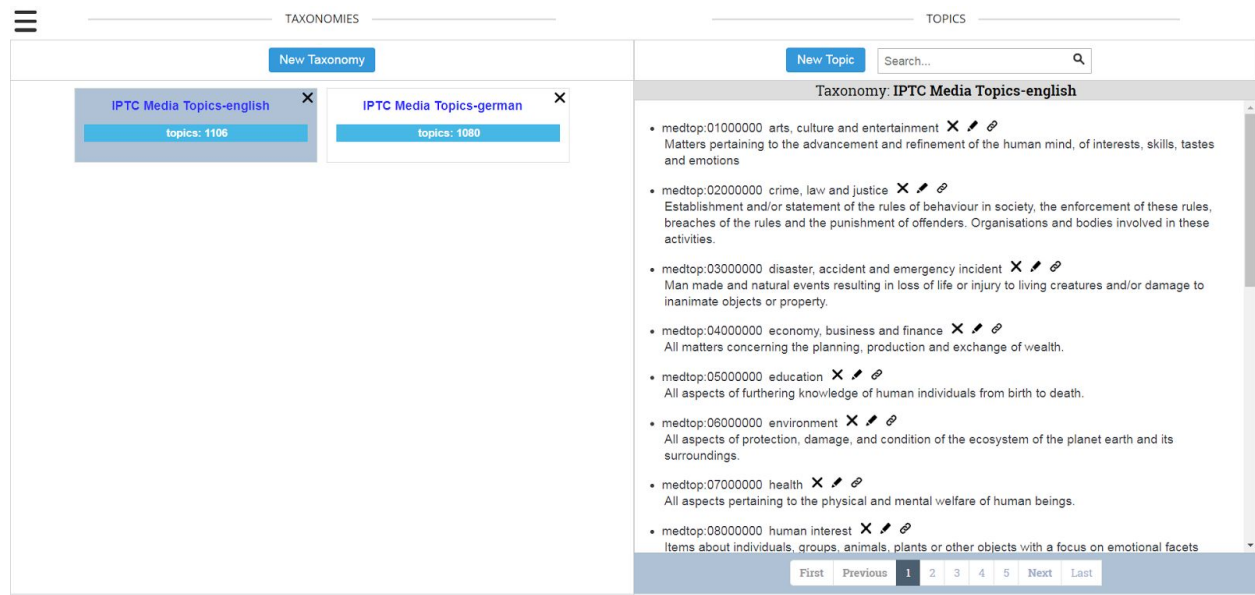
Figure 1: The EXTRA Toolkit Menu

## Taxonomy Manager

The Taxonomy Manager can be used to create, edit and delete taxonomies. Once a taxonomy has been created, the user may use this tool to retrieve, edit, delete or create new topics within the taxonomy.

## Edit an Existing Taxonomy

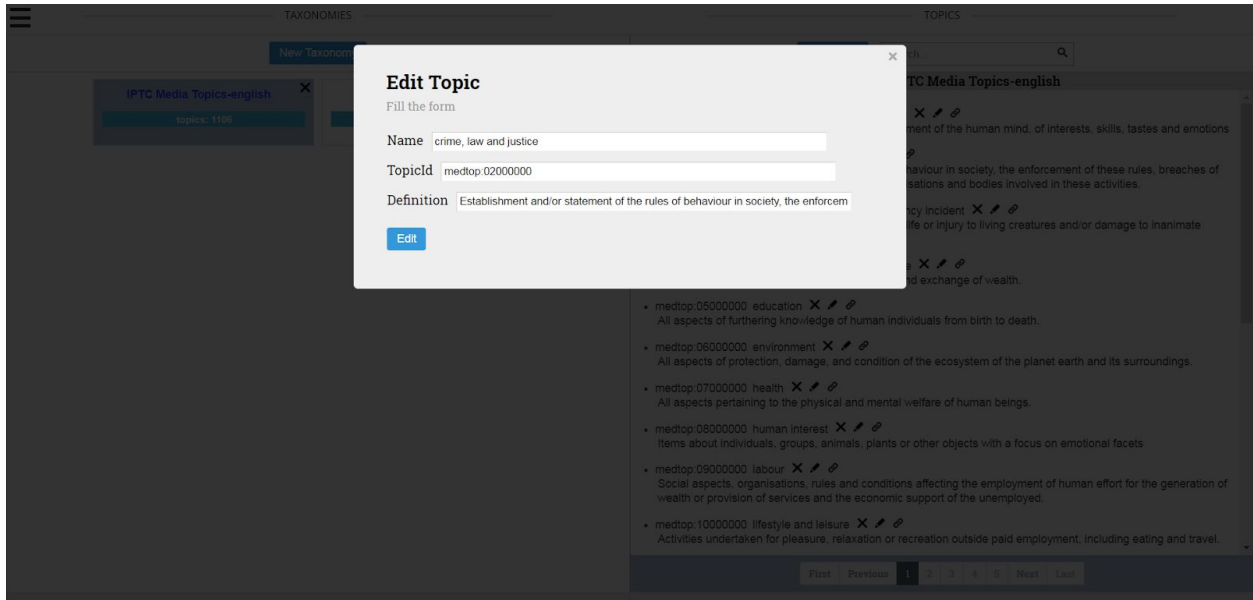
The EXTRA Toolkit comes preloaded with two sample taxonomies: 1.) IPTC Media Topics in English and 2.) IPTC Media Topics in German. Selecting a taxonomy from the left, will load the associated topics on the right. See Figure 2.



**Figure 2:** Load an existing taxonomy

To edit a taxonomy topic:

1. On the right pane, use the scroll bar, or the pagination buttons at the bottom of the screen, or the search bar at the top, to navigate the taxonomy.
2. Once you've found the desired topic, use the X icon to delete a topic, the pencil icon to edit a topic (see Figure 3), or the link icon to click through to the view/edit the rule associated with this topic in the Rule Editor (see the Rule Editor section)

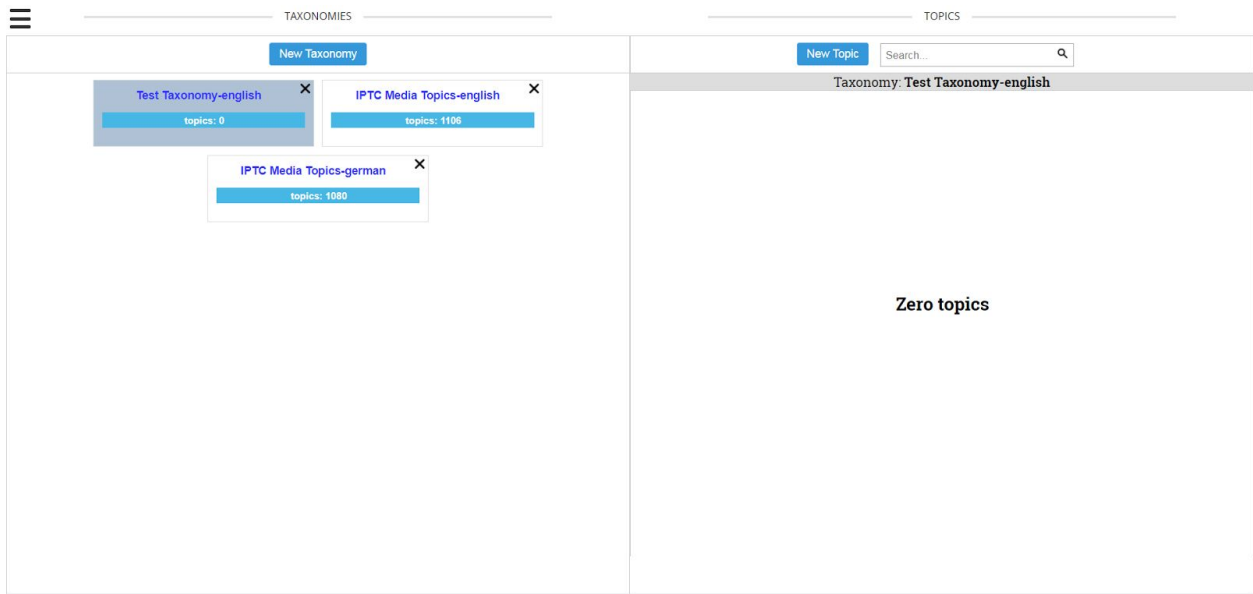


**Figure 3:** Edit a Topic

## Create a New Taxonomy

To bulk load an existing taxonomy, please use the API to script this process.

To create a new taxonomy from scratch, click the "New Taxonomy" button on the left, create a name and designate the appropriate language from the pulldown. You should see a new box on the left representing your taxonomy (See Figure 4).



**Figure 4:** Create a new taxonomy

Once you've created a new taxonomy, you may create the topics that comprise that taxonomy. On the right, click "New Topic" and enter the Name, Topic ID, and Definition (See Figure 5).

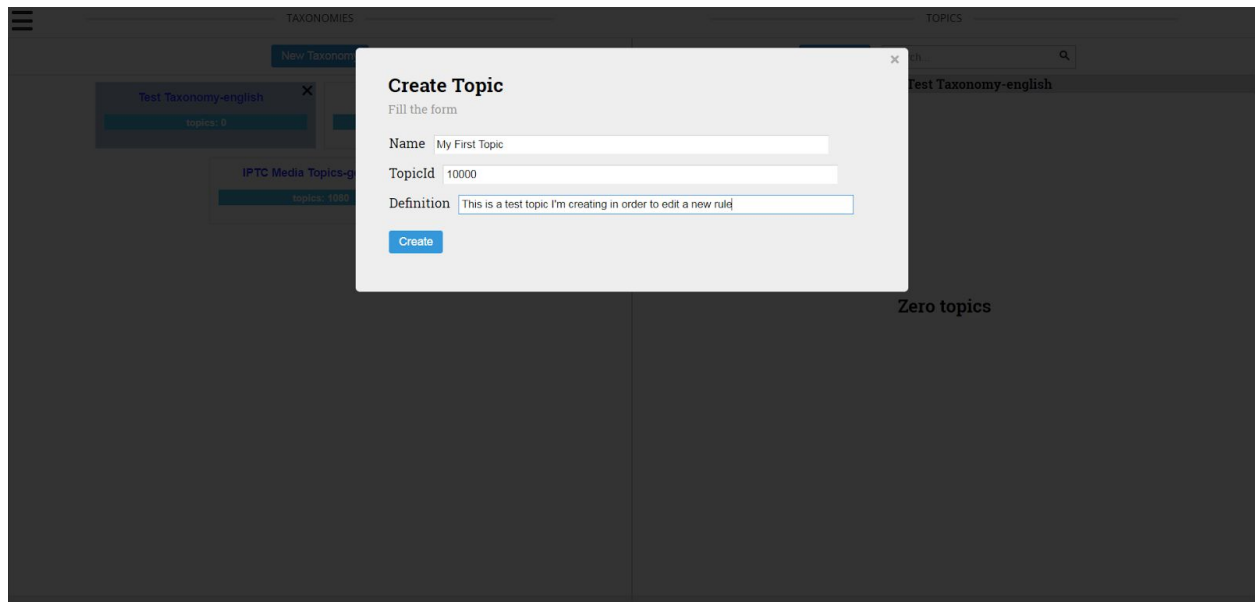


Figure 5: Create Topic Dialog

Any new topics you create will now appear in the right pane (See Figure 6).

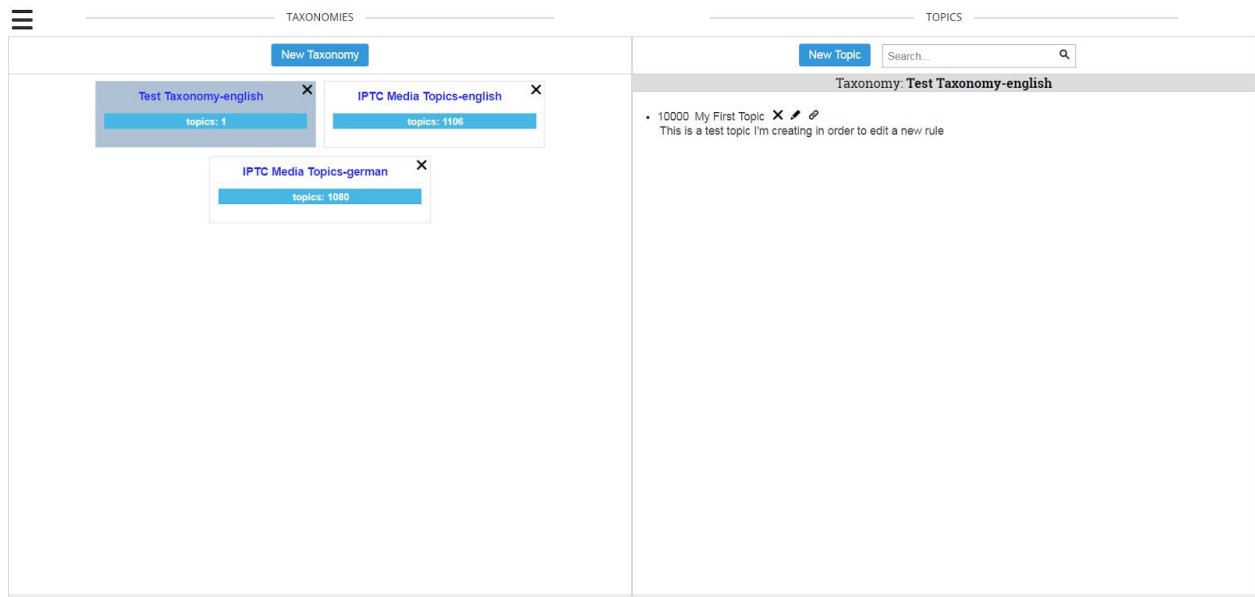


Figure 6: New Topic Created

# Schemas Manager

This tool may be used to create (or edit and delete) schemas. Schemas identify the fields in your corpus (e.g. headline, title, body, caption, slug, byline) that will be indexed for searching, and against which you will write your rules.

Schemas are specific to each corpus. For example, one news organization may use a "title" field, while another may use "headline." One organization may choose not index bylines, while another may find it useful, particularly if certain writers cover specific topics (in which case a rule may be written such as: if byline="Jane Doe", then topic="Sports")

The fields can be specified as either:

1. Text fields (e.g. title, body, etc.), where the user can indicate if the field contains sentences or paragraphs, or
2. Non-text fields, such as dates or IDs, which are not tokenized and used for exact match or, in the case of dates, range queries.

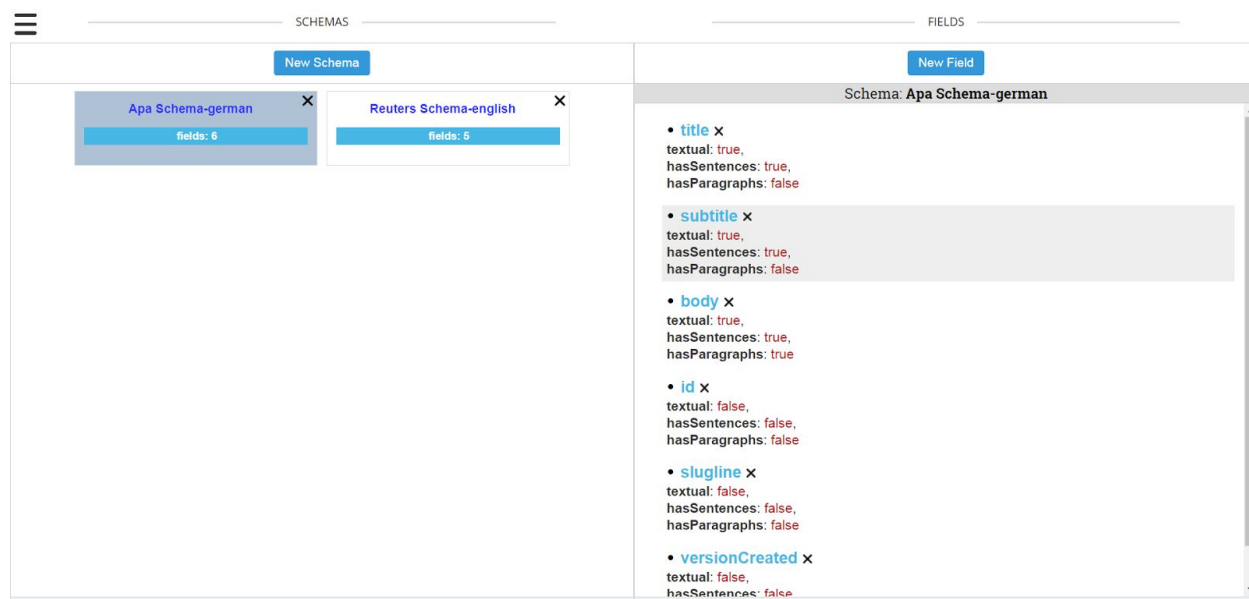
Two schemas (see Figure 7) are used to demonstrate the functionality of the IPTC EXTRA Toolkit in <http://github.com/ipetc/extra-examples>

1. Thomson Reuters (in English)
2. Austria Press Agency/APA (in German)

The corresponding corpora for these schemas are available upon request. Please see the Corpora section for details.

Please note that both of these schemas make use of a field called `text_content`, which is a concatenation of all text fields. This field is always automatically generated by the EXTRA engine, for all schemas, not just the two examples.





**Figure 7:** Schema for APA

## Rule Editor & EXTRA Query Language (EQL)

The Rule Editor is the tool used to build rules with the EXTRA Query Language. In order to begin writing rules, first review the basics of the EXTRA language.

### EXTRA Query Language

EQL is a robust, open source<sup>1</sup>, query language that enables you to write rules analyzing the text of a document with regard to proximity, frequency, and field (on structured documents). It also supports stemming, wildcards, literals and regular expressions.

While EQL supports four basic boolean operators (and, or, not and prox), these operators may be combined with modifiers to create fifteen "extended operators" as outlined in the table below (See Table 1). A more detailed discussion of the components of the language follows with a breakdown of relations, modifiers, boolean operators and boolean modifiers. However you may use this table for a quick start to get a sense of at the kinds of queries that are possible.

<sup>1</sup> Licensed under the MIT License, <https://opensource.org/licenses/MIT>

## Extended Operators

The following tables provides examples of extended operators available in EXTRA. Review the definitions and sample queries in order to begin constructing queries.

**Table 1:** The set of available extended operators to be used in EXTRA

#	EQL Extended Operator	Definition	Sample Query
1	and	AND: Takes two or more statements. Matches only if all statements are true.	(and (title any "star") (body any "movie") )
2	or	OR: Takes two or more statements. Matches if at least one statement is true.	(or ("soccer") (title ="fifa") )
3	not	NOT: Used in combination with AND. Matches if the statement does not appear in combination with statement under the AND.	(and (title = "entertainment") (not (title adj "Reuters Entertainment News Summary") ) )
4	or/countunique>=N	MINIMUM: Takes one or more statements. Matches if a minimum of N statements from the list appear in the text.	(or/countunique>=3 ("dollar") ("euro") ("renminbi") )
5	prox/unit=word /distance<=N	DISTANCE: Combined with a number. Takes two or more statements. Matches if statements are within N number of words from each other.	(prox/unit=word/distance<2 (body any "drug") (body any "prices") )
6	or/count>=N	MINIMUM OCCURRENCE: Combined with a number. Takes one or more statements. Matches if statement appears N amount of times in the text.  Takes only one-to-three word statements. Does not support nesting other operators.	(or/count>=3 (body any/stemming "bafta bollywood cinema documentary hollywood movie") ) (or/count>=3 (body adj/stemming "academy award") (body adj/stemming "box office") )
7	prox/ordered	ORDER: Takes two or more statements. Matches if statements appear in the text in the same order that they appear in the rule.	(prox/ordered (body any "police") (body any "shot") )

			Gets different results than:  (prox/ordered (body any "shot") (body any "police") )
8	prox/unit=sentence /distance=1	SENTENCE: Takes two or more statements. Matches if statements appear within the same sentence.	(prox/unit=sentence/distance=1 (body any/stemming "drug") (body any/stemming "price") )
9	prox/unit=paragraph /distance=1	PARAGRAPH: Takes two or more statements. Matches if all the statements occur in the same paragraph.	(prox/unit=paragraph/distance=1 (body = "Merkel") (body = "Obama") )
10	prox/unit=word/distance>0	NOT IN PHRASE: Takes two statements. Matches if the first statement occurs outside of the second statement.	(prox/unit=word/distance>0 (text_content any/stemming "baby") (text_content adj/stemming "baby boomer") )
11	prox/unit=word/distance>N	NOT WITHIN DISTANCE: Combined with a number. Takes two or more statements. Matches if the statements are not within N amount of words from each other.	(prox/unit=word/distance>8 (or ("birdie") ("bogey") ("eagle") ) (or ("putt") ) )
12	prox/unit=sentence /distance>1	NOT IN SENTENCE: Takes two or more statements. Matches if all the statements do appear in the same document, but not in the same sentence.	(prox/unit=sentence/distance>1 (body = "media") (body = "social") )
13	prox/unit=paragraph /distance>1	NOT IN PARAGRAPH: Takes two or more statements. Matches if all the statements do appear in the same document, but not in the same paragraph.	(prox/unit=paragraph/distance>1 (body = "media") (body = "social") )
14	prox/unit=word /distance<=N/ordered	ORDER AND DISTANCE: Combined with a number. Takes two or more statements. Category matches if both statements occur in the same order in which they are written in the rule and if both are within N amount of words to each other.	(prox/ordered/unit=word/distance<10 (body any "drug") (body any "prices") )  Gets different results than:

			(prox/ordered/unit=word/distance<10 (body any "prices") (body any "drug") )
15	or/count<N	MAXIMUM OCCURRENCE: Combined with a number. Takes at least one statement. Matches if the statement appears no more than N amount of times in the text.	(or/count<1 (body any/stemming "animal pet cat dog parrot hamster veterinary") )

## Search Clauses

The core building block of EQL is a search clause, which consists of an index, a relation and search terms. Search terms **MUST** be enclosed in double quotes if they contain any of the following characters: < > = / ( ) and whitespace. The index and relation are optional. If no index is included, then server choice is implied:

**searchClause = (index relation)? searchTerm**

For example:

(title == term)  
(Title adj "term1 term2")  
(body any "t1 t2 t3 ...tN")  
(slugline == "A/B/C")  
("this is a set of terms")

Modifiers can be applied on relations to adapt their meaning.

**relation = relationName (/modifierName comparator value)\***

For example:

(title any/stemming "t1 t2 t3")  
(title any/stemming/locale=en "t1 t2 t3")  
(body any/count>N "t1 ..... tN")

## Relations

**Table 2:** The set of available relations to be used in EQL

Relation	Description	Example
----------	-------------	---------

=	This relation is used for term matching	<b>(title = term1)</b> Search for a title contains term1
==	This relation is used for exact equality matching	<b>(identifier == "gb 141 staff a-m" )</b>  Search for the string 'gb 141 staff a-m' in the identifier index. This must be the full value of the field.
<>	This relation means 'not equal to' and matches anything which is not exactly equal to the search term.	<b>(date &lt;&gt; 2004-01-01)</b> Search for any date except the first of January, 2004  <b>(identifier &lt;&gt; "" )</b> Search for any identifier which is not the empty string.
<, >, <=,>=	These relations retain their regular meanings as pertaining to ordered terms (less than, greater than, less than or equal to, greater than or equal to).	<b>(date &gt; 2006-09-01)</b> Search for dates after the 1st of September, 2006
adj	Adjacency. Used for phrase searches. All of the words in the search term must appear, and must be adjacent to each other in the record in the order of the search term. Note: hyphenated words must use adj.	<b>(title adj "lord of the flies")</b> Search for the phrase 'lord of the flies' somewhere in the title.  <b>(text_content adj "anti-drug")</b> Search for the phrase "anti-drug" anywhere in the document.  <b>(description adj "blue shirt")</b> Search for 'blue' immediately followed by 'shirt' in the description.
all, any	These relations may be used when the term contains multiple items to indicate "all of these items" or "any of these items."  These queries could be expressed using boolean AND and OR respectively.  <b>Note:</b> If the search term consists not only of words but of phrases then OR/AND must be used.	<b>(title all "lord flies")</b> Search for both lord and flies in the title  <b>(description any "computer calculator")</b> Search for either computer or calculator in the description.
within	Within may be used with a search term that has multiple dimensions. It matches if the database's term falls completely within the range, area or volume described by the search term, including the given parameters.	<b>(date within "2002 2003")</b> Search for dates between 2002 and 2003 inclusive.

## Relation Modifiers

**Table 3:** The set of relation modifiers to be used in EQL

Modifier	Description	Example
<b>casesensitive</b>	The server is instructed to respect the case of the search term  <b>title any/casesensitive "Fish"</b>	(or (title adj/casesensitive "Gates charity") )  Gets different results than:  (or (title adj/casesensitive "gates charity") )
<b>word</b>	The term should be broken into words (default).	
<b>string</b>	The term is a single item, and should not be broken up.	Use string to target the untokenized field
<b>regexp</b>	The term should be treated as a regular expression.  <b>title adj/regexp "(lord king ruler) of th[ea] r.*s"</b>  Match lord or king or ruler, followed by of, followed by the or tha, followed by r plus zero or more characters plus s.	(or (title adj/regexp "d+?-?s+?year?-?s?olds?") (body adj/regexp "d+?-?s+?year?-?s?olds?") )  Matches the patterns "100-year-old man", "12 year olds", "50-year old custom" etc. in the title or body of an article.
<b>stemming</b>	The search term should be stemmed.	(and (body any/stemming "movie") (body any/stemming "star") (not (body adj "co-star") ) )
<b>literal</b>	Uses the whitespace tokenizer, no tokenized boundaries.	(or (title any/literal "\$5") )  (or (title any/literal "Backpage.com") )

## Boolean Operators

Search clauses can be combined with n-ary boolean operators in a prefix way. The set of supported operators is **or**, **and**, **not** and **prox**. Both lower and upper case is valid.

The basic way to way combine search clauses is:

**booleanClause = (booleanOperator searchClause+)**

For example:

```
(OR
  (searchClause1)
  (searchClause2)
  (searchClause3)
)
```

**Note: Parentheses in search clauses are required.**

Also boolean operators can be used in combination with boolean clauses, or searchClauses:

```
(booleanOperator1
  (searchClause1)
  (booleanOperator2
    (searchClause2)
    (searchClause3)
  )
  (searchClause4)
)
```

These boolean operators may be modified similar to way in which relations are modified:

**booleanOperator = booleanName (/modifierName comparator value)\***

For example:

```
(prox/unit=word/distance>4)
(prox/unit=sentence/distance<2)
(prox/unit=paragraph/distance<=1)
(or/count<4)
(or/count>3)
(and/occurrence>2)
(or/occurrence>2)
```

## Boolean Modifiers

**Table 4:** Boolean Modifiers to be used in EQL

Boolean Modifier	Description	Examples
<b>distance</b>	Modifies prox operator. The distance that the two clauses should be separated by.  Distance symbol value  Symbol can be =, >, <, >=, <= Default value 0	<p>(prox/unit=word/distance&gt;2/ordered (title = hat) (title any “cat dog”) )</p> <p>The first statement (title = hat) appears more than two words before the second.</p> <p>(prox/unit=paragraph ( “cat”) ( “hat”) )</p> <p>Find cat and hat appearing in the same paragraph (distance defaulting to 0) in either order (unordered default)</p>
<b>unit=value</b>	The type of unit for the distance  Proximity units: 'paragraph', 'sentence', 'word' and 'element'	
<b>ordered</b>	The order of the two clauses must be as per the query	
<b>unordered</b>	The order of the two clauses is unimportant. This is the default.	
<b>countunique</b>	Modifies OR/AND operators.  count symbol value E.g. countunique > n, countunique = n, countunique <= n  From the statements combined by OR/AND operators more, less or equal to n statements must appear according to count value.  In the case of OR, the modifier makes the operator more restrictive, while in the case of AND, the modifier relaxes it.	<p>(or/countunique&gt;2 (title = termX) (title any “t1 t2”) (body any “t1 t2 t3 t4”) (body any “t5 t6 t7 t8 t9”) )</p> <p>From the 4 statements at least two of them must be true (instead of one in the original version of or operator).</p>
<b>count</b>	Modifies OR/AND operators. Takes only one-to-three word statements. Does not support nesting other operators.  count symbol value E.g. count > n, count = n, count <= n  In the case of OR, any of the statements should be true. Also the total number of occurrences of the true statements must	<p>(or/count&gt;4 (body = obama) (body = trump) (“Clinton”) )</p> <p>The total number of occurrences must be more than 2 times.</p>



	<p>be less, equal or more than N times according to the value of count modifier.</p> <p>In the case of AND, all the statements must be true but also the total number of occurrences must be less, equal or more than N times according to the value of count modifier.</p>	
--	---	--

Consider the following example:

```
(and
  (title any "any of these should be appear in the title")
  (prox/unit=word/distance>=5
    (body any "term1 term2")
    (body any "these terms must have a distance of 5 words from term1 and term2")
  )
  (or/countunique>3
    (title adj "this is an exact title")
    ("list of terms that can match any field")
    (slugline == "eu/crisis/greece")
    (publicationDate < "8/4/2016")
    (body any/count>5 "election results")
    (prox/unit=sentence/distance<1/ordered
      (body any "X Y Z")
      (body any "K M L")
    )
  )
)
```

On the first level, the query consists of an and operator applied in three clauses. A search clause (red color) and two nested boolean clauses (blue and green respectively). All these must be true. The second boolean clause (blue color) is a prox operator that consists of two search clauses. These clauses must have a distance of five words as imposed by the prox modifiers. The third clause is an or operator modified by count boolean modifier. In other words, from the six statements (5 search clauses and one nested boolean clause) at least three must be true.

## Text\_Content

The `text_content` field is a concatenation of all text fields. This field is always automatically generated by the EXTRA engine, for all schemas, not just the two examples included in the toolkit.

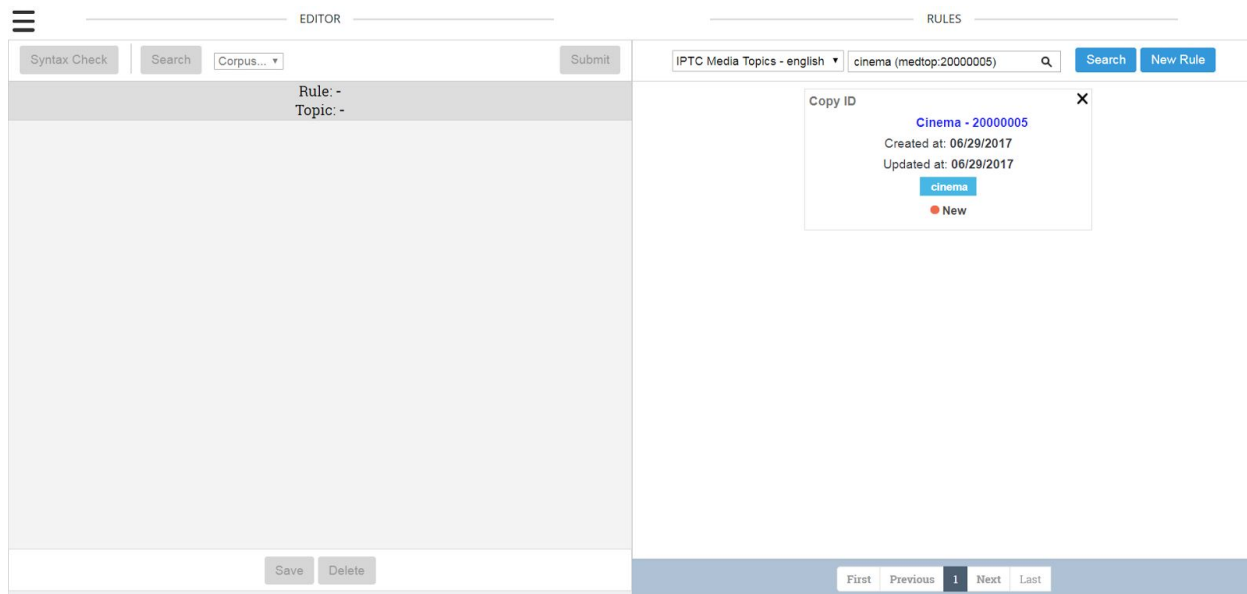
## Rule Editor

### Open the Rule Editor

Now you're ready to open the Rule Editor and begin writing rules. You may access it in two ways:

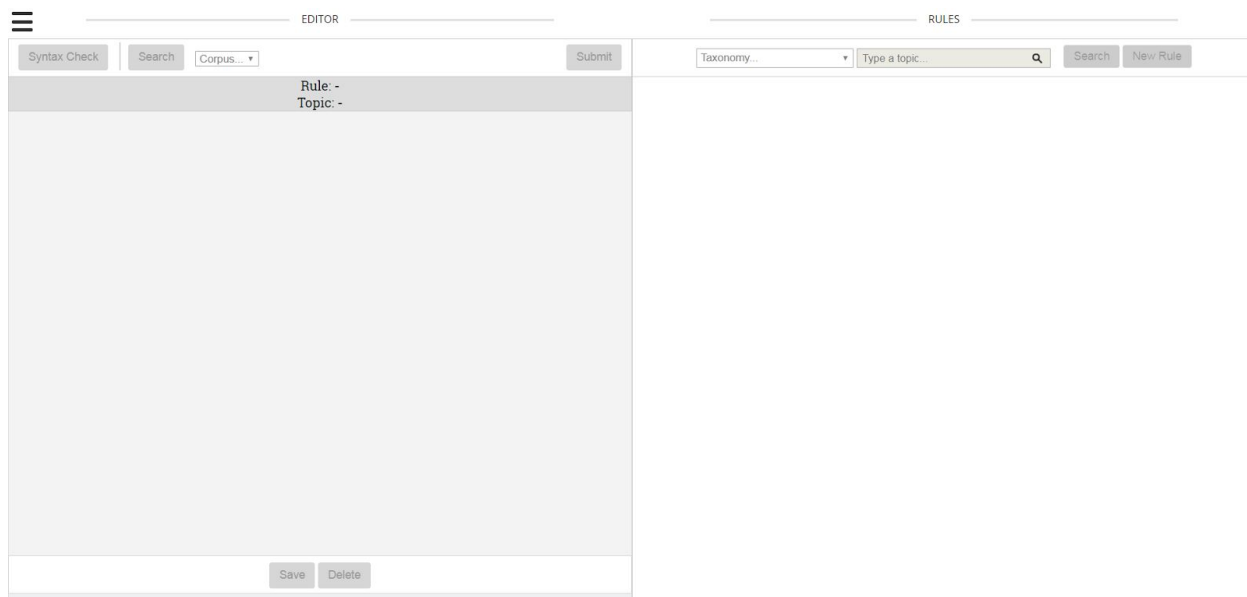
1. From within the Taxonomy Manager, by browsing to a specific topic, then clicking the link icon.
2. From the Main Menu in the top lefthand corner.

If accessing from the Taxonomy Manager, you will arrive directly at the rule for the topic which you have selected (See Figure 8). Note that there may be multiple rules associated with each topic. You may wish to use this feature to save different drafts of a rule, or different versions by different users. Ultimately however, it is recommended that only one rule per topic is used when classifying content.



**Figure 8:** Opening the Rule Editor via the Taxonomy Manager

If accessing from the main top left menu, then you will arrive at a blank screen (see Figure 9) and will need to select the relevant taxonomy from the pulldown in the top right pane, then either click Search to view all rules, or use the search bar to pull up a specific topic.



**Figure 9:** Opening the Rule Editor via the Main Menu

See Figure 10 for a screenshot of all rules that appear when selecting the "IPTC Media Topics - English" taxonomy and clicking "Search"

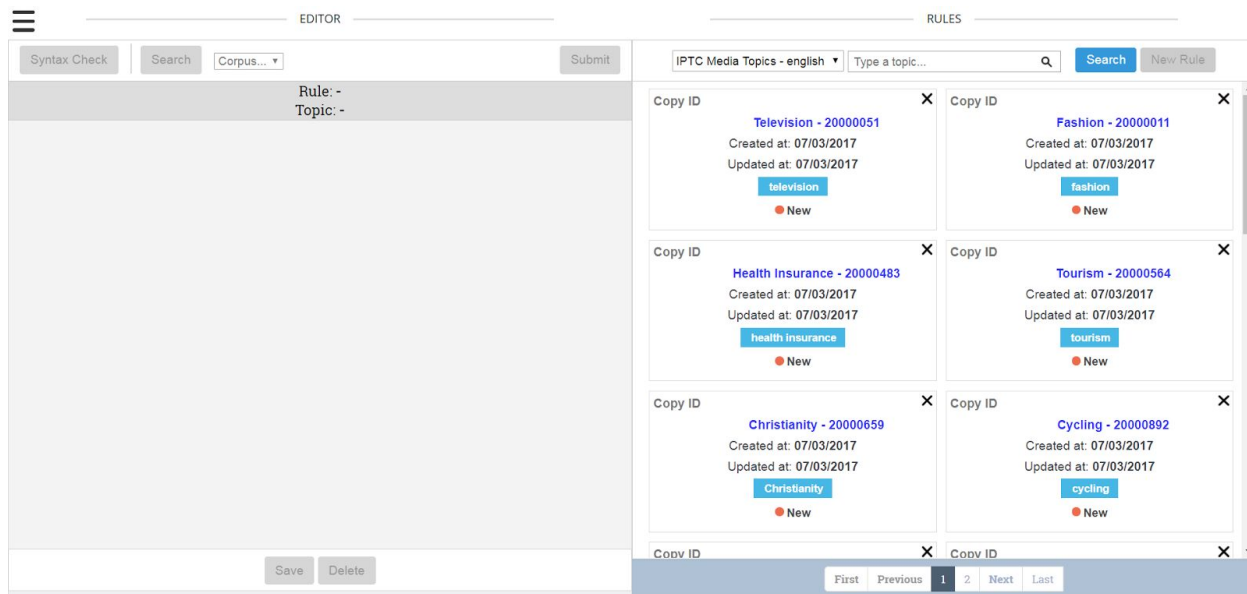


Figure 10: Selecting "IPTC Media Topics - English"

## Open a Rule

Once you've navigated to the rules and can see either all available rules in the taxonomy (Figure 10) or the just rules within a specific topic (Figure 8), simply click a rule to open it. It appears in the lefthand pane (See Figure 11).

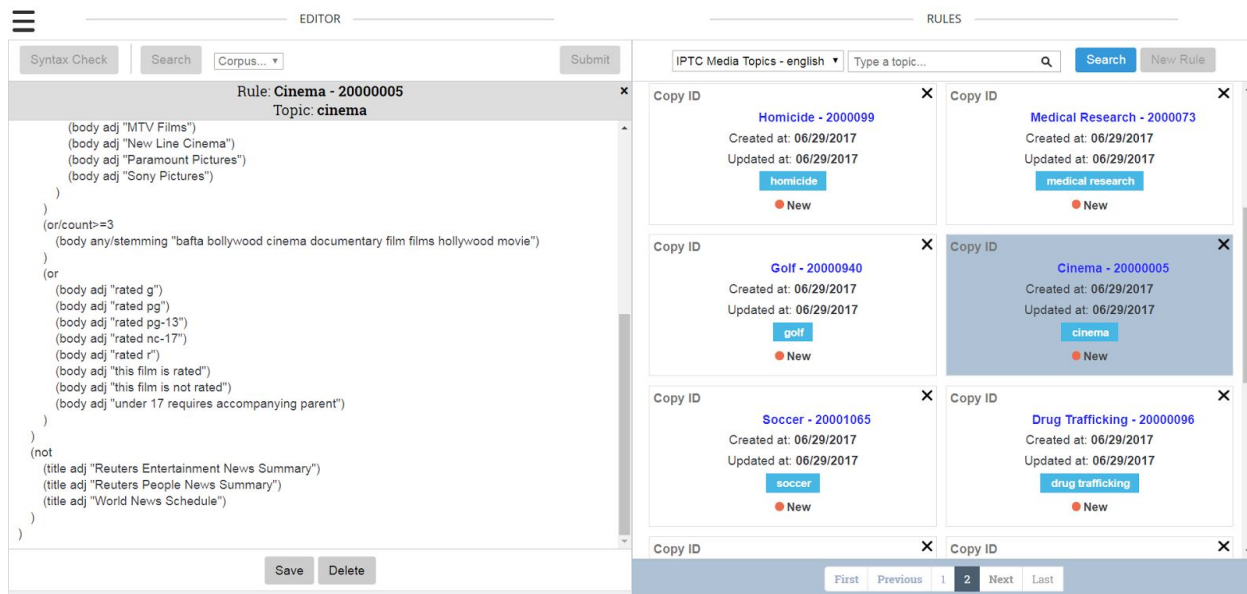


Figure 11: Open a rule

## Test a Rule

To test the performance of a rule, you must first choose a corpus. To request a sample corpus, please see the Corpora section. To load your own corpus, see:

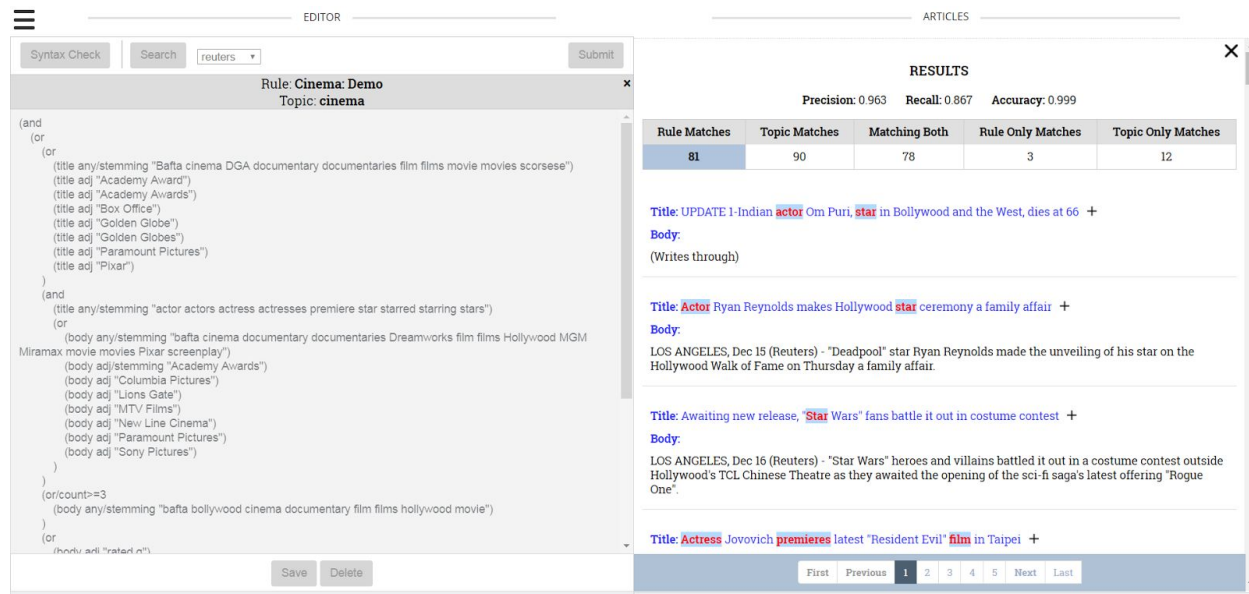
<http://github.com/iptc/extra-examples>

Once loaded, a user may select the relevant corpus from the upper lefthand pane, then click Search.

You will see your results broken out into 5 sets, along with precision and recall scores (see Figure 12).

### Results Sets:

- **Rule Matches:** The results in the corpus that match the rule. Specific parts of the document that match are called out in red.
- **Topic Matches:** The results in the corpus that are tagged with the topic to which this rule belongs.
- **Matching Both:** True positives. The results in the corpus that match both the rule and the tagged documents.
- **Rule Only Matches:** False positives. The results in the corpus that match the rule, but not the tagged documents.
- **Topic Only Matches:** False Negatives. The results in the corpus that match the tagged documents, but not the rule.
- **Precision:**  $\text{True Positives} / (\text{True Positives} + \text{False Positives})$
- **Recall:**  $\text{True Positives} / (\text{True Positives} + \text{False Negatives})$



**Figure 12:** Test a rule and view search results with precision and recall scores

## Edit a Rule

To edit a rule, simply place your cursor in the lefthand pane and begin typing. To check that the rule you've written contains the correct syntax, click the "Syntax Check" button on the top left.

Note: if you have not yet selected a Corpus from the pulldown to the right of the "Search" button, you will need to make that selection first in order to make the "Syntax Check" button active.

Once you have clicked "Syntax Check," use the righthand pane to toggle between three views:

1. Parsed Rule (Figure 13)
2. Elastic Search Query (Figure 14)
3. Syntax Tree (Figure 15)

The views are not editable and are for informational purposes only to help you gain a better understanding of how your rules are structured and the underlying Elastic Search translations.

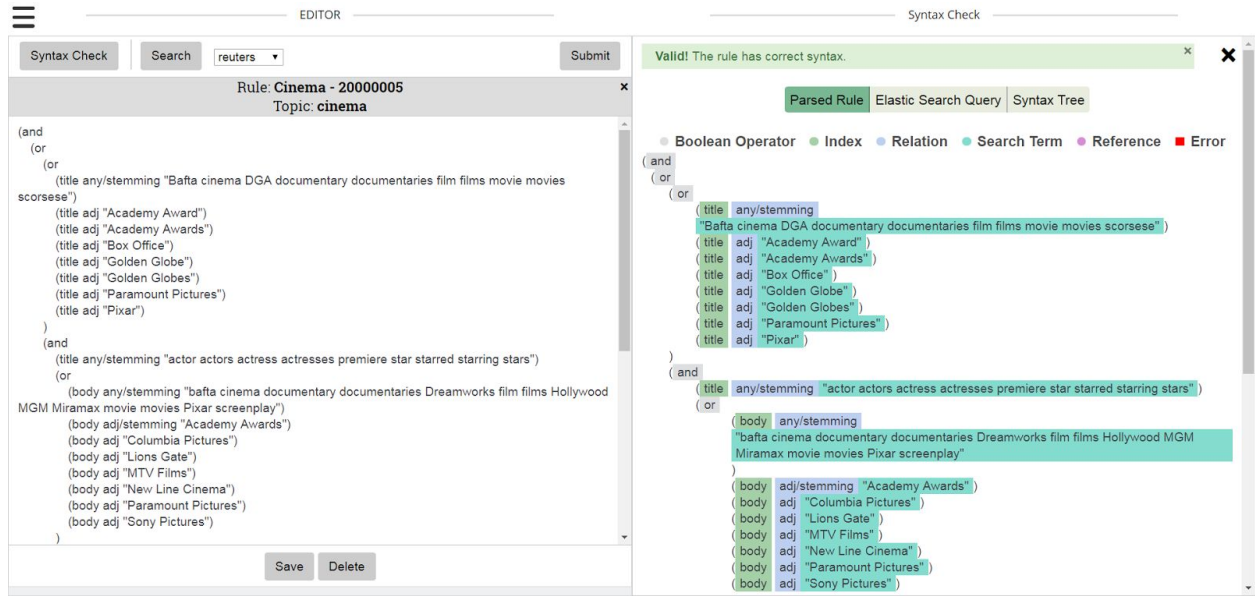


Figure 13: Syntax Check, Parsed Rule

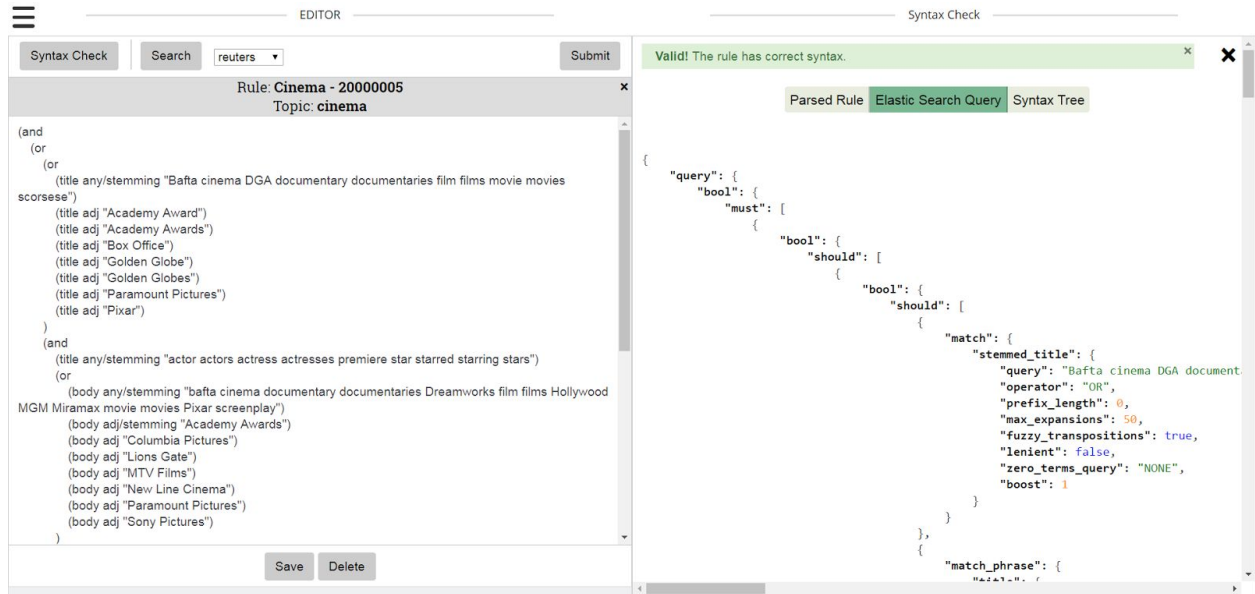


Figure 14: Syntax Check, Elastic Search Query translation

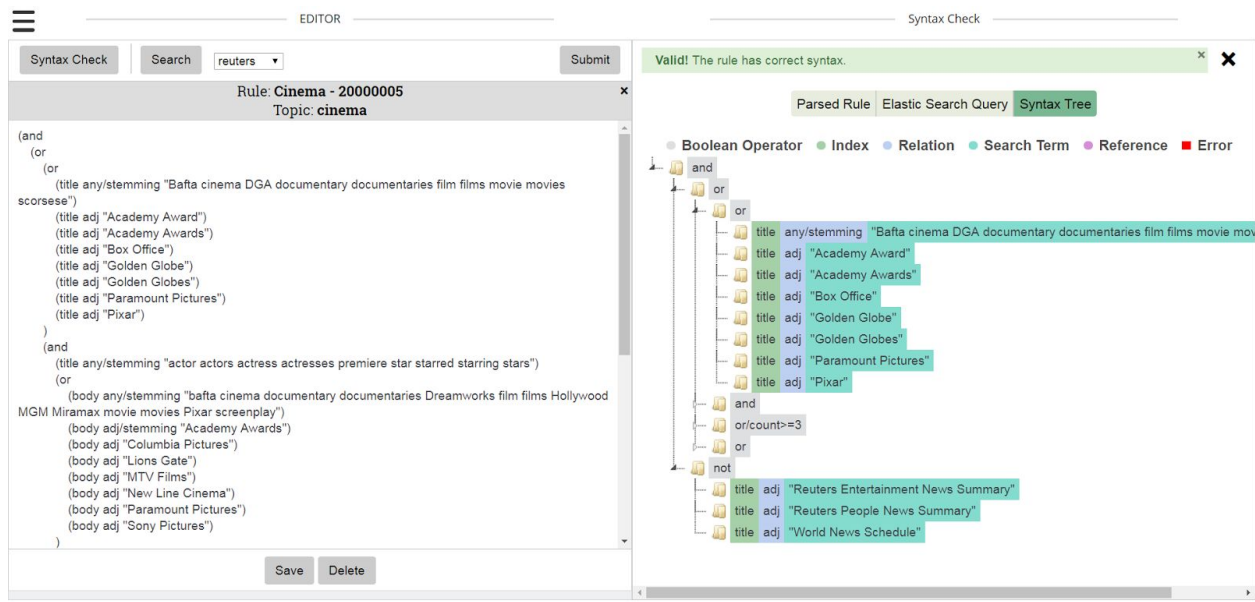


Figure 15: Syntax Check, Syntax Tree

## Create a New Rule

To create a new rule, you must navigate first to the relevant topic with which this rule will be associated. Select the taxonomy and then use the search bar to select the topic within that taxonomy (See Figure 16).

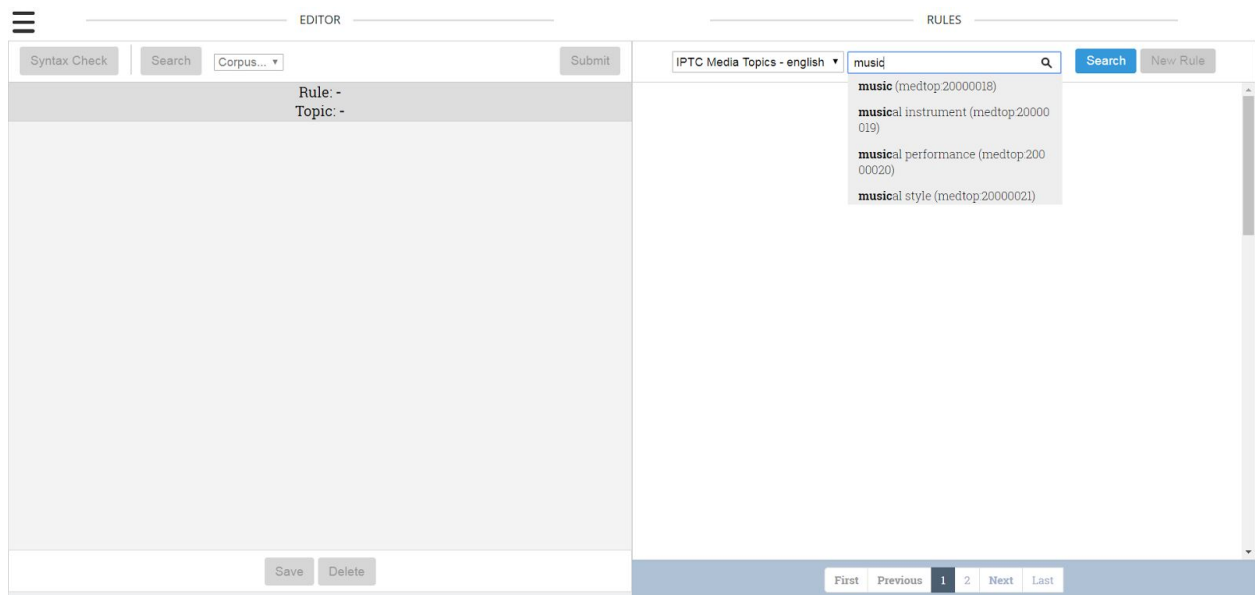
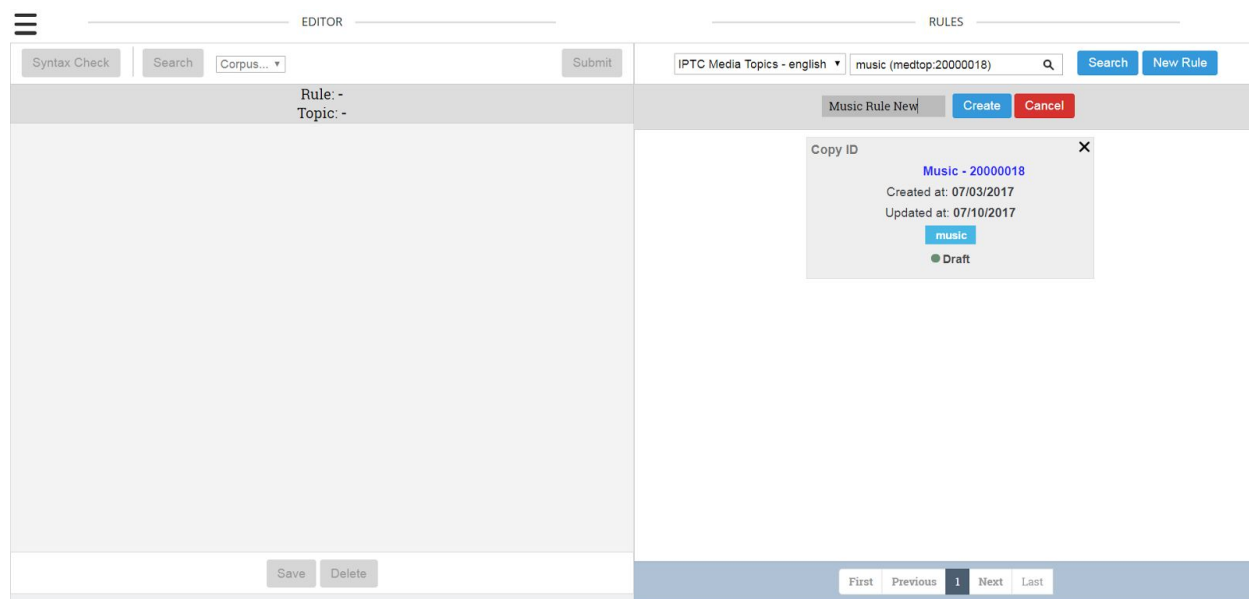
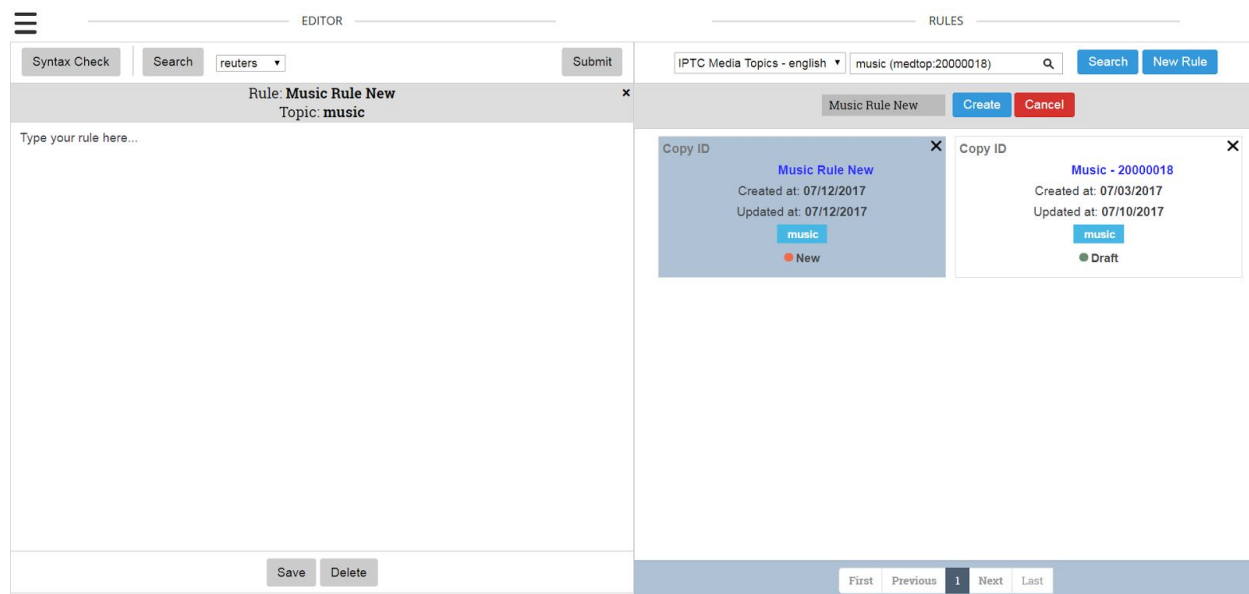


Figure 16: Create a new rule, select topic





**Figure 17:** Create a new rule, name the rule



**Figure 18:** Create a new rule, begin entering your rule

## Submit a Rule

To submit a rule, click the "Submit" button on the upper righthand corner of the left pane. When you submit a rule, you make it available for classification. You will be prompted for to assign a "Group Name" but this field is optional. To skip it, leave the field blank and click "Submit" (See

Figure 19). Once the rule is successfully submitted, you will receive a confirmation message (See Figure 20). See the Document Tagging section of this document to learn more about classifying content. Note: only rules that have been submitted will be returned in the Document Tagging module.

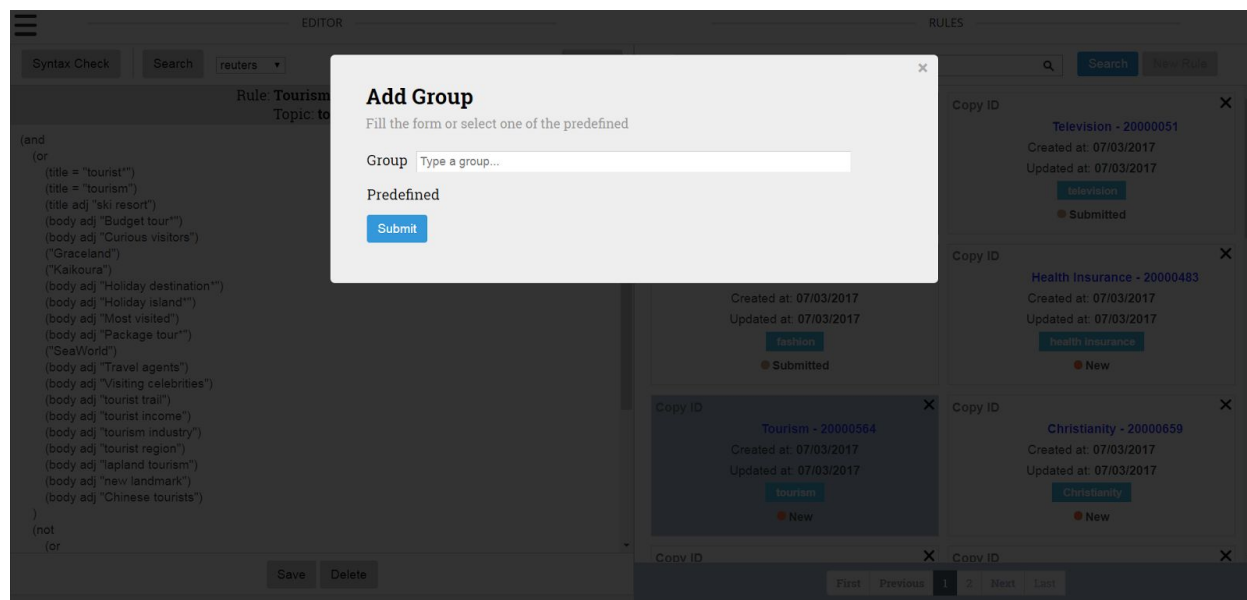


Figure 19: Add Group is optional

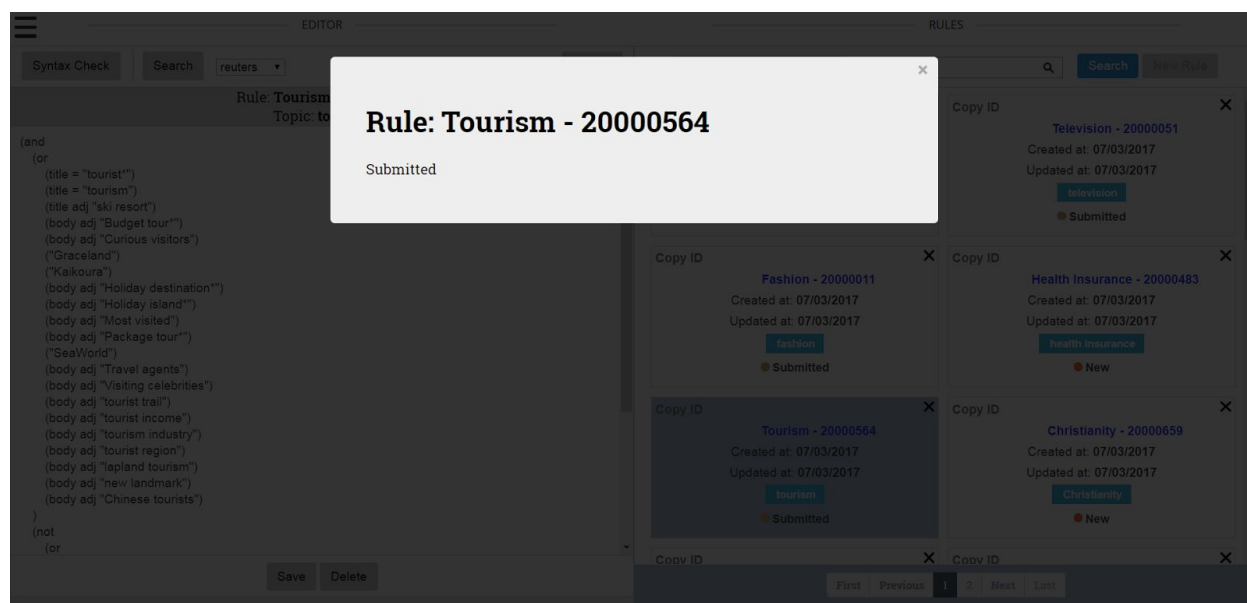


Figure 20: Rule submitted

# Document Search Tool

The Document Search Tool (Figure 21) enables you to navigate (by topics), the indexed documents of the sample corpora. If you would like to use this tool on the sample corpora, please see the Corpora section about obtaining permission.

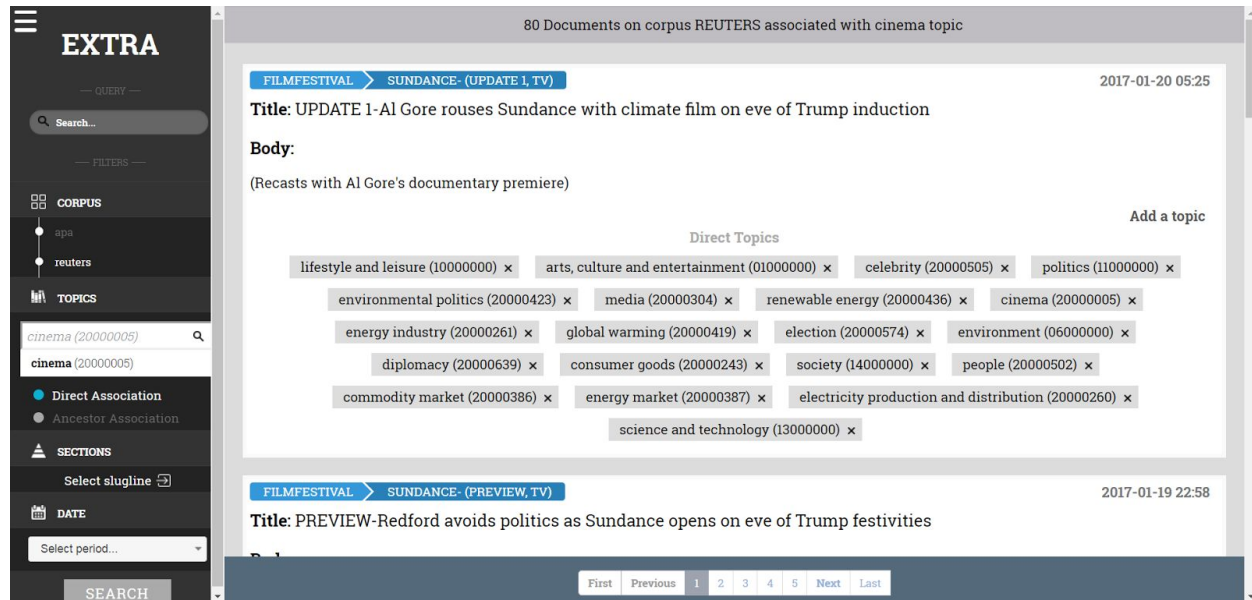


Figure 21: Document Search results

## Document Tagging

Use the Document Tagging tool to submit documents from your corpus to the rules engine and see which topics are returned. The document submitted must be in json format. If your document contains special characters or quotation marks, you'll want to first escape those characters (search online for "json escape" to find freely available tools).

For example:

```
{
  "title":"This is the title text",
  "subtitle": "This is the subtitle text",
  "body" : "This is the body text"
}
```

If a field contains paragraphs, these should be defined by using <p> html tags e.g.

"body": "<p>.....</p><p>.....</p><p>.....</p><p>.....</p>"

See Figure 22 for an example of a movie review submitted, with the rule "Cinema" returned.

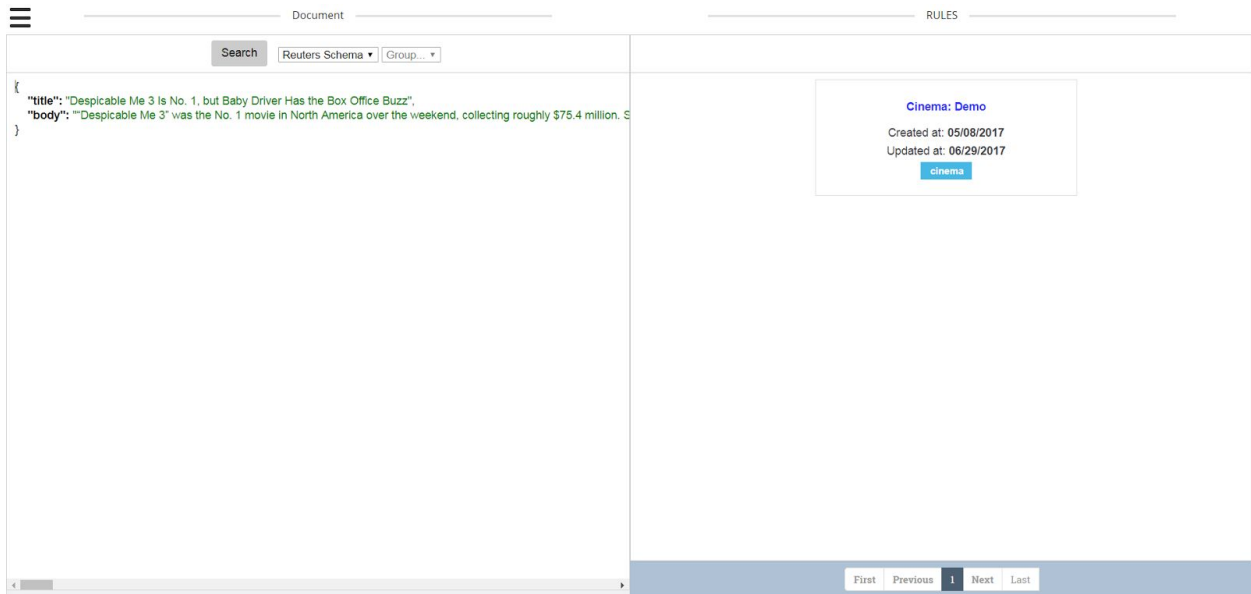


Figure 22: Document Tagging, submitting a movie review and getting the topic "Cinema" returned

## Corpora

If you are interested in obtaining one of the sample corpora for testing purposes, please complete this form: <https://www.iptc.org/get/extra-corpus-downloadform>

Upon clicking Submit, you'll be sent to a page from where both corpora can be downloaded. The download URL is valid for 24 hours.

## Logging Issues

Should you find any issues when testing the software, please use the GitHub Issues system of [extra-ext](https://github.com/iptc/extra-ext) repository to report bugs: <https://github.com/iptc/extra-ext/issues>

# Appendix: Sample Rules

## IPTC Media Topics - English

### Advertising

```
(and
  (or
    (title any "ad ads advertise advertisement advertisements advertising")
    (or/count >= 3
      (text_content any "ad ads advertise advertisement advertisements advertising")
    )
    (title = "marketing")
    (title adj "ad spend")
    (title adj "ad package*")
    (body adj "Ad revenue*")
    (body adj "Ad site")
    (body adj "Ad spend")
    (body adj "Advertising agencies")
    (body adj "Advertising agency")
    (body adj "Advertising campaign*")
    (body adj "Advertising client")
    (body adj "Advertising clients")
    (body adj "Advertising firm")
    (body adj "Advertising firms")
    (body adj "Advertising rule")
    (body adj "Advertising rules")
    (body adj "Advertising website")
    (body adj "Company to sever ties with the show")
    (body adj "Firm cut ties to show")
    (title adj "Kawasaki dumps apprentice")
    (body adj "Online ad")
    (body adj "Online ads")
    (body adj "Personalised ads")
    (body adj "Personalized ads")
    (body adj "Targeted ads")
    (body adj "Targeted advertising")
  )
  (not
    (title = "summary")
    (title = "schedule")
    (title = "timeline")
    (title adj "sumner redstone")
    (title adj "black market")
    (title = "politics")
    (body adj "advertising site")
  )
)
```

## Antitrust Issue

```
(or
  (title = "antitrust")
  (title adj "anti-trust")
  (body adj "competition and markets")
  (body adj "competition watchdog")
  (text_content adj "price fixing")
  (text_content adj "price-fixing")
  (or/count>=3
    (text_content adj "antitrust")
    (text_content adj "anti-trust")
    (text_content adj "price fixing")
    (text_content adj "price-fixing")
  )
)
```

## Bankruptcy

```
(and
  (or
    (title = "bankruptcy")
    (or/count>=3
      ("bankruptcy")
    )
    (body adj "Administration process")
    (body adj "bankruptcy filing")
    (body adj "Goes into administration")
    (body adj "Gone into administration")
    ("Insolvency")
    (body adj "Side goes bust")
    (body adj "Went into administration")
  )
  (not
    (title = "summary")
    (title adj "to invest")
    (title adj "special report")
    (title adj "wilbur ross")
    (title = "iceland")
    (body adj "bankruptcy attorney")
    (title adj "stimulus plan")
  )
)
```

## Cancer

```
(and
  (or
    (title any "cancer cancerous Hodgkins hodgkin's leukemia melanoma melanomas oncologist oncologists oncology")
    (or/count>=4
      (text_content any "cancer oncology")
    )
  )
  (and
    (text_content adj "National Cancer Institute")
    (text_content any "death deaths")
  )
)
```

```

    )
  )
  (not
    (title adj "Reuters Health News Summary")
  )
)

```

## Censorship

```

(and
  (or
    (title = "censor*")
    (title = "censorship")
    (or/count>=3
      (body = "censor*")
    )
  )
  (and
    (text_content adj "Communications Decency Act")
    (text_content adj "legal shield")
  )
  (and
    (text_content any "blacklist")
    (text_content any "censor")
  )
  (and
    (body any/literal "Backpage.com")
    (or
      (text_content adj "supreme court")
      (text_content adj "Senate subcommittee")
    )
    (or
      (text_content any "censor censored censoring censorship")
      (text_content adj "free speech")
    )
  )
  (and
    (or
      (text_content any "china iran korea korean russia")
      (text_content adj "north korea")
    )
    (or
      (prox/unit=sentence/distance=1
        (body any "blocked censor* restrict*")
        (body any "cyberauthority internet web")
      )
      (prox/unit=sentence/distance=1
        (body any "blacklist*")
        (body any "author* writer*")
      )
    )
  )
)
(not
  (or
    (title = "summary")
    (title = "schedule")
    (title = "breivik")
  )
)

```

```

    (text_content adj "foreign investment")
  )
)
)

```

## Children

```

(and
  (or
    (title any "babies child children infant infants kids newborn newborns pediatric pediatrician pediatricians toddler toddlers")
    (prox/unit=word/distance>0
      (title any "kid")
      (or
        (title adj "comeback kid")
        (title adj "tennis-comeback kid")
      )
    )
  )
  (and
    (title any "baby boy boys girl girls")
    (or
      (body any "boyhood child children circumcised circumcision girlhood infant infants kid kids pediatric pediatrics pediatrician pediatricians toddler toddlers vaccine vaccinate vaccination")
      (body adj "boys-and-girls")
    )
  )
  (and
    (or
      (title adj/regex "\d+\.?\s+?year\-\?\s?old")
      (title adj/regex "\d+\.?\s+?month\-\?\s?old")
    )
    (body any/stemming "boy child children girl infant juvenile kid newborn schoolboy schoolgirl toddler")
  )
  (or/count>=3
    (body any "child children newborn newborns infant infants kids pediatric pediatrician pediatricians schoolboy schoolboys schoolgirl schoolgirls toddler toddlers")
    (text_content adj "1-year-old")
    (text_content adj "2-year-old")
    (text_content adj "3-year-old")
    (text_content adj "4-year-old")
    (text_content adj "5-year-old")
    (text_content adj "6-year-old")
    (text_content adj "7-year-old")
    (text_content adj "8-year-old")
    (text_content adj "9-year-old")
    (text_content adj "10-year-old")
    (text_content adj "11-year-old")
    (text_content adj "12-year-old")
    (text_content adj "1-year old")
    (text_content adj "2-year old")
    (text_content adj "3-year old")
    (text_content adj "4-year old")
    (text_content adj "5-year old")
    (text_content adj "6-year old")
    (text_content adj "7-year old")
    (text_content adj "8-year old")
    (text_content adj "9-year old")
    (text_content adj "10-year old")
  )
)

```



```

(text_content adj "11-year old")
(text_content adj "12-year old")
(text_content adj "one-year-old")
(text_content adj "two-year-old")
(text_content adj "three-year-old")
(text_content adj "four-year-old")
(text_content adj "five-year-old")
(text_content adj "six-year-old")
(text_content adj "seven-year-old")
(text_content adj "eight-year-old")
(text_content adj "nine-year-old")
(text_content adj "ten-year-old")
(text_content adj "eleven-year-old")
(text_content adj "twelve-year-old")
(text_content adj "one-year old")
(text_content adj "two-year old")
(text_content adj "three-year old")
(text_content adj "four-year old")
(text_content adj "five-year old")
(text_content adj "six-year old")
(text_content adj "seven-year old")
(text_content adj "eight-year old")
(text_content adj "nine-year old")
(text_content adj "ten-year old")
(text_content adj "eleven-year old")
(text_content adj "twelve-year old")
(text_content adj "age of one")
(text_content adj "age of two")
(text_content adj "age of three")
(text_content adj "age of four")
(text_content adj "age of five")
(text_content adj "age of six")
(text_content adj "age of seven")
(text_content adj "age of eight")
(text_content adj "age of nine")
(text_content adj "age of ten")
(text_content adj "age of eleven")
(text_content adj "age of twelve")
(text_content adj "was just six")
(text_content adj "Aylan Kurdi")
(text_content adj "boys-and-girls")
(text_content adj "childhood stunting")
(text_content adj "child beauty")
(text_content adj "child development")
(text_content adj "children tested")
(text_content adj "create babies")
(text_content adj "commercial surrogacy")
(text_content adj "cub scouts")
(text_content adj "katie rough")
(text_content adj "one-child policy")
(text_content adj "Sharbat Gula")
(text_content adj "underprivileged children")
(text_content adj "young children")
)
(and
  (prox/unit=sentence/distance=1
    (text_content any/stemming "boy boys child children girl girls youth youths")
  )
)

```

```

        (or
          (text_content any/stemming "abuse circumcised circumcision detention homeless kidnap kidnapped malnutrition
malnourished molest molestation sex trafficker traffickers trafficking underprivileged vaccinate vaccination")
          (text_content adj/stemming "female genital mutilation")
        )
      )
    (or/count>=2
      (text_content any/stemming "boy boys child children girl girls youth")
    )
  )
)
(not
  (title adj "baby orangutan")
  (text_content adj "Edgar Maddison Welch")
  (title any "Summary")
  (title adj "World-News-Schedule")
  (title adj "World News Schedule")
)
)

```

## Christianity

```

(and
  (or
    (title = "christian*")
    (title = "vatican")
    (title = "epiphany")
    (title = "nazarene")
    (title adj "christmas poster")
    (title = "pope")
    (body adj "jesuit")
    (body adj "catholic core")
    (body adj "infant christ")
    (body adj "episcopal church")
    (body adj "christmas day service")
    (body adj "christian nonprofit")
    (body adj "christian anti")
    (body adj "christian town")
    (body adj "christian pilgrims")
    (body adj "christian farmers")
    (body adj "senior christian")
    (body adj "christian missionaries")
    (body adj "knights of malta")
    (body adj "coptic christian")
    (body adj "muslims and christians")
  )
  (not
    (title = "summary")
    (title = "schedule")
    (title = "update")
    (title = "refile")
    (title = "diary")
  )
)
)

```

## Cinema

```
(and
  (or
    (or
      (title any/stemming "Bafta cinema DGA documentary documentaries film films movie movies scorsese")
      (title adj "Academy Award")
      (title adj "Academy Awards")
      (title adj "Box Office")
      (title adj "Golden Globe")
      (title adj "Golden Globes")
      (title adj "Paramount Pictures")
      (title adj "Pixar")
    )
    (and
      (title any/stemming "actor actors actress actresses premiere star starred starring stars")
      (or
        (body any/stemming "bafta cinema documentary documentaries Dreamworks film films Hollywood MGM Miramax movie
movies Pixar screenplay")
        (body adj/stemming "Academy Awards")
        (body adj "Columbia Pictures")
        (body adj "Lions Gate")
        (body adj "MTV Films")
        (body adj "New Line Cinema")
        (body adj "Paramount Pictures")
        (body adj "Sony Pictures")
      )
    )
    (or/count>=3
      (body any/stemming "bafta bollywood cinema documentary film films hollywood movie")
    )
    (or
      (body adj "rated g")
      (body adj "rated pg")
      (body adj "rated pg-13")
      (body adj "rated nc-17")
      (body adj "rated r")
      (body adj "this film is rated")
      (body adj "this film is not rated")
      (body adj "under 17 requires accompanying parent")
    )
  )
  (not
    (title adj "Reuters Entertainment News Summary")
    (title adj "Reuters People News Summary")
    (title adj "World News Schedule")
  )
)
```

## Cycling

```
(and
  (or
    (body adj "Bicycle lane*")
    (body adj "Bike lane*")
  )
)
```

```

    (body adj "Cycle path*")
    (body adj "Cycle track*")
    ("Cycling")
    ("Cyclists")
    (body adj "Riding bicycle*")
    (body adj "Tour de France")
    (body adj "Tour down under")
    (body adj "Use bike*")
    (body adj "Using bike*")
  )
  (not
    (title = "summary")
    (title = "schedule")
  )
)

```

## Data Protection Policy

```

(and
  (or
    (title = "manning")
    (title = "snowden")
    (title = "wikileaks")
    (title adj "anti-hacking")
    (title adj "reality winner")
    (title adj "cyber campaign")
    (title adj "cyber controls")
    (title adj "massive hack")
    (title adj "yahoo hack")
    (body adj "data protection")
    (body adj "protect data")
    ("hackers")
    ("encryption")
  )
  (not
    (title = "summary")
    (title = "schedule")
  )
)

```

## Diseases and Conditions

```

(and
  (or
    (title any "alzheimer's bubonic cancer disease diseases diabetes flu hiv influenza superbug superbugs zika")
    (title any/literal "AIDS")
    (title adj "avian flu")
    (title adj "avian influenza flu")
    (title adj "binge-eating")
    (title adj "bird flu")
    (title adj "heart disease")
    (body adj "avian flu")
    (body adj "the plague")
  )
  (not
    (title = "summary")
  )
)

```

)  
)

## Drug Trafficking

```
(and
  (or
    (title any "narcotic narcotics")
    (title adj "El Chapo")
    (title adj/stemming "drug crime")
    (title adj "drug crisis")
    (title adj/stemming "drug gang")
    (title adj/stemming "drug lord")
    (title adj/stemming "drug offender")
    (title adj/stemming "drug sentence")
    (title adj "drug traffick*")
    (title adj/stemming "drug war")
    (title adj "drug trafficking")
    (title adj/stemming "illegal drug")
    (title adj "war on drugs")
  )
  (or/count>3
    (body adj "Drug Enforcement Administration")
    (body adj "D.E.A.")
    (body adj/stemming "drug killing")
    (body adj/stemming "drug traffick")
    (body adj "drug trafficking")
    (body adj/stemming "drug crime")
    (body adj/stemming "drug gang")
    (body adj/stemming "drug offender")
  )
)
(and
  (title any "Duterte")
  (or
    (body any "drug drugs narcotic narcotics")
    (body adj "anti-drug")
    (body adj "anti-narcotic")
    (body adj "anti-drugs")
    (body adj "anti-narcotics")
  )
)
)
(and
  (title any "cannabis cocaine crack drug drugs heroin marijuana meth PCP pot narcotic narcotics")
  (text_content any/stemming "abuse cartel crime dealer gang overdose rehab prison seize seizure sell sentence sentencing sold smuggle")
)
)
(prox/unit=sentence/distance=1
  (body any "Duterte Duterte's")
  (body any "anti-drug anti-drugs anti-narcotic anti-narcotics drug drug narcotic narcotics")
)
)
(prox/unit=sentence/distance=1
  (body any "drug drugs")
  (body any "sentence sentencing")
)
)
(prox/unit=sentence/distance=1
  (body any "marijuana")
  (body any "illegal legalization legalize")
)
```

```

    )
  )
(not
  (text_content any "biotech cancer healthcare medical medicine Pfizer pharma pharmaceutical")
  // excluding news summaries and schedules from reuters corpus //
  (title adj "News Summary")
  (title adj "News Schedule")
)
)

```

## Economic Policy

```

(or
  (title adj "economic policy")
  (title adj "anti-trade")
  (title adj "retaliatory trade")
  (title adj "trade priority")
  (title adj "free trade deal")
  (title = "subsidies")
  (title adj "most protectionist")
  (body adj "decision to float")
  (body adj "freely floated")
)

```

## Education

```

(and
  (or
    (title any "education colleges professor university universities")
    (prox/unit=word/distance>0
      (title any "college")
      (title adj "electoral college")
    )
    (body adj "Disrupted education")
    (body adj "Educate and care for")
    (body adj "Education and opportunities")
    (body adj "Education firm*")
    (body adj "Education sector")
    (body adj "Educational program*")
    (body adj "Educational support")
    (body adj "Home schooling")
    (body adj "Lack of schools")
    (body adj "Lack of teachers")
    (body adj "Language tutor*")
    (body adj "Middle school")
    (body adj "On their way to school")
    (body adj "Patriotic education")
    (body adj "Primary school*")
    (body adj "Private school*")
    (body adj "Public education")
    (body adj "Religious school*")
    (body adj "Rural school*")
    (body adj "School board*")
    (body adj "School district*")
    (body adj "School rules")
    (body adj "School standards")
  )
)

```

```

        (body adj "Teacher salaries")
        (body adj "University lecturer*")
    )
    (not
        (title = "summary")
        (title = "schedule")
        (title = "poll*")
        (title = "diary")
        (title adj "top jobs")
        (title = "election")
        (title = "golf")
        (title = "nfl")
    )
)

```

## Employment Statistics

```

(and
    (or
        (title adj "employment statistics")
        (title adj "job growth")
        (title adj "employment growth")
        (title adj "jobless claims")
        (title adj "jobs report")
        (title adj "unemployment rate")
        (title adj "jobs data")
        (body adj "employment statistics")
        (body adj "job growth")
        (body adj "employment growth")
        (body adj "jobless claims")
        (body adj "jobs report")
        (body adj "unemployment rate")
    )
    (not
        (or
            (title = "summary")
            (title = "schedule")
            (title = "bipartisan")
        )
    )
)

```

## Fashion

```

(and
    (or
        (body adj "American Apparel")
        ("catwalk")
        ("catwalks")
        (body adj "clothing design*")
        (body adj "clothing industry")
        ("couture")
        ("Lanvin")
        (body adj "fashion consultant*")
        (body adj "fashion designer*")
        (body adj "fashion magazine*")
    )
)

```

```

        (body adj "fashion model*")
        (body adj "fashion show*")
        (body adj "fashion week")
        (body adj "French fashion brand*")
        (body adj "high fashion")
        (body adj "low fashion")
        (body adj "model* agenc*")
        (body adj "street fashion")
    )
    (not
        (title = "summary")
        (title = "general")
        (title = "schedule")
        (title = "NHL")
        (title = "suspects")
        (title = "accuses")
        (title = "dossier")
        (title = "chief")
        (title = "driver")
    )
)

```

## Fire

```

    (and
        (or
            (title any "firefighter firefighters")
            (body adj "fire swept")
            (body adj "battle the blaze")
            (body adj "blazing high-rise")
            (body adj "house fire")
            (text_content any/stemming "arson arsonist arsonists")
        )
        (not
            (title = "summary")
            (title = "schedule")
        )
    )
)

```

## Gays and Lesbians

```

    (and
        (or
            (title any "homosexual homosexuality homosexuals lesbian lesbians lgbt")
            (title adj "bathroom bill")
            (title adj "gay pride")
            (title adj "gay rights")
            (title adj/literal "l.g.b.t.")
            (title adj "same-sex")
            (or/count>=3
                (text_content any "gay lgbt homosexual homosexuality homosexuals trans transgender transsexual transsexuals")
            )
            (text_content adj "gay equality")
            (text_content adj "gay rights")
            (text_content adj "gender dysphoria")
            (text_content adj "gender parity")
        )
    )
)

```



```

    (and
      (text_content any "Orlando")
      (text_content adj "Noor Salman")
    )
    (and
      (title any "manning")
      (body any "trans transgender")
    )
  )
  (not
    (title = "summary")
    (title = "schedule")
    (title = "golf")
    (title adj "cash-for-ash")
    (body adj "esteban santiago")
  )
)

```

## Golf

```

    (and
      (or
        (or
          (title any "golf pga p.g.a.")
          (title adj "abu dhabi championship")
          (title adj "Ryder Cup")
          (title adj "South African Open")
          (or/countunique>3
            (body adj "1-iron")
            (body adj "2-iron")
            (body adj "3-iron")
            (body adj "4-iron")
            (body adj "5-iron")
            (body adj "6-iron")
            (body adj "7-iron")
            (body adj "8-iron")
            (body adj "9-iron")
            (body adj "Jordan Spieth")
            (body adj "over-par")
            (body adj "Rory McIlroy")
            (body adj "under-par")
          )
        )
      )
    )
    (not
      (body adj "Billy Walters")
      (title adj "Reuters Sports News Summary")
      (title adj "Reuters Sports Schedule")
      (title adj "Reuters U.S. Sports Schedule")
    )
  )

```

## Health Insurance

```

    (and
      (or
        (title any "ACA AXA healthcare Medicaid Medicare Obamacare")
      )
    )
  )

```

```

(title adj "Health coverage")
(title adj "Health insur*")
(or/count>=2
  (text_content any "ACA Aetna BlueCross GHI Medicaid Medicare Obamacare UnitedHealth")
  (text_content adj "Blue Cross")
  (body adj "health coverage")
  (text_content adj "health insur*")
  (body adj "single payer")
  (text_content adj "United Health")
)
)
(not
  (or
    (title = "summary")
    (title = "schedule")
  )
)
)

```

## Homicide

```

(and
  (or
    (title = "homicide*")
    (title = "murder*")
    (title = "gunman")
    (title = "gunmen")
    (title = "breivik")
    (title = "rampage")
    (title = "executed")
    (title adj "nightclub shoot*")
    (title adj "lauderdale shoot*")
    (title adj "florida shooting")
    (title adj "truck attack*")
    (title adj "dylann roof")
    (title adj "shooter kill*")
    (title adj "massacre prison")
    (title adj "church shoot*")
    (title adj "shooting suspect")
    (title adj "shooter open* fire")
    (title adj "people shot")
    (title adj "airport shooting")
    (title adj "nightclub attack*")
    (title adj "convicted of killing")
    (title adj "berlin attack")
    (title adj "shot dead")
    (title adj "fatally shot")
    (body adj "extrajudicial killings")
    (body = "homicide*")
    (body = "murder")
    (body = "murdered")
    (body adj "arrested following the death")
    (body adj "personally killing")
    (body adj "execution-style")
    (body adj "shooting rampage")
  )
)
(not

```

```

(or
  (title = "schedule")
  (title = "summary")
  (title = "kidnap*")
  (title = "aleppo")
  (title = "abortion*")
  (title = "militant*")
  (title adj "very good")
  (title adj "civil rights")
  (title adj "getting away with murder")
  (title adj "massacre denial")
  (title adj "fake news")
  (body adj "getting away with murder")
)
)
)

```

## Immigration

```

(and
  (or
    (title any "deportation deported deporting deports emigration immigration immigrant immigrants migrant migrants refugee
refugees visa visas")
    (title adj "border wall")
    (title adj "eu passport")
    (title adj "entry visa")
    (title adj "migrant crisis")
    (title adj "migration deal")
    (title adj "curb migration")
    (title adj "asylum cap")
    (title adj "stem migration")
    (title adj "round up kids")
    (body adj "border with mexico")
    (body adj "border wall")
    (body adj "cuban immigrants")
    (body adj "eb-5")
    (body adj "h1-b")
    (body adj "high-skilled visa")
    (body adj "high-skilled visas")
    (body adj "immigration policy")
    (body adj "immigration policies")
    (body adj "illegal immigration")
    (body adj "immigration skills")
    (body any "emigration")
    (body adj "migrant worker")
    (body adj "migrant workers")
    (body adj "rescind visas")
    (body adj "transit country")
    (body adj "without visa")
    (body adj "without visas")
    (body adj "work visa")
    (body adj "work visas")
    (prox/unit=sentence/distance=1
      (body adj "anti-immigrant")
      (body adj "alternative for germany")
    )
  )
)

```

```

(text_content any "immigrant immigration immigrants migrant migrants refugee refugees")
(prox/unit=word/distance<2
  (body any "asylum residency")
  (body any "grant grated granting seek seeking sought")
)
)
)
(not
  (title = "schedule")
  (title = "summary")
  (title = "inauguration")
  (title = "protest*")
  (title = "forex")
  (title = "fifa")
  (title = "factbox")
  (title = "poll")
)
)

```

## Islam

```

(or
  (title any "burqa islam hijab islamist islamophobia mosque muslim salafi salafis salafist salafists salafism shariah wahhabism
wahhabist wahhabists")
  (prox/unit=word/distance>0
    (or
      (title any "Islamic")
      (title any "Islamist")
    )
    (or
      (title adj "Islamic militants")
      (title adj "Islamic state")
      (title adj "Islamist group")
      (title adj "Islamic militant")
      (title adj "Islamic militants")
    )
  )
  (title adj "antimuslim")
  (title adj "anti-muslim")
  (and
    (or
      (text_content adj/stemming "face veil")
      (text_content adj/stemming "full-face veil")
      (text_content adj/stemming "head scarf")
      (text_content adj/stemming "head scarves")
    )
    (or
      (text_content any/stemming "burqa islam hijab islamist islamophobia mosque muslim salafi salafis salafist salafists salafism
shariah wahhabism wahhabist wahhabists")
    )
  )
  (or/count>=2
    (text_content any "burqa hijab islam islamist islamophobia mosque muslim salafi salafis salafist salafists salafism shariah
wahhabism wahhabist wahhabists")
  )
  (or/count>=2

```

```

    (text_content any "abaya abayas burka burkas burqa burqas burqini burqinis burquini burquinis chador chadors hijab hijabs
    khimar khimars niqab niqabs shayla shaylas")
  )
)

```

## Medical Research

```

(and
  (or
    ("amgen")
    ("biosimilar*")
    (body adj "biotech cancer")
    (body adj "biotech drug*")
    (body adj "biotech firm*")
    (body adj "cancer drug*")
    ("esmo")
    ("Glaxosmithkline")
    (body adj "higher risk of developing")
    ("immunotherapeutic*")
    ("Lancet")
    (body adj "medical journal*")
    (body adj "medical research")
    (body adj "pharmaceutical industry")
    (body adj "tamiflu treatment")
  )
  (not
    (title = "summary")
    (title = "schedule")
  )
)
)

```

## Music

```

(or
  (title any "beatles coldplay flamenco lumineers shakira u2")
  (title adj "Andrea Bocelli")
  (title adj "Bob Dylan")
  (title adj "Bruce Springsteen")
  (title adj "David Bowie")
  (title adj "Elton John")
  (title adj "Elvis Presley")
  (title adj "George Michael")
  (title adj "Grateful Dead")
  (title adj "Katy Perry")
  (title adj "Michael Jackson")
  (title adj "music festival")
  (title adj "Pearl Jam")
  (title adj "Peter Sarstedt")
  (title adj "Pink Floyd")
  (title adj "Rolling Stones")
  (title adj "Taylor Swift")
)
  (and
    (title any "conductor conductors singer singers")
    (text_content any "carols music musician musicians orchestra orchestras sing song songs sung")
  )
)

```

```

(text_content adj "The Slants")
(or
  (text_content adj "rock band")
  (text_content adj "Simon Tam")
)
)
(prox/unit=word/distance<=3
  (text_content any/stemming "play")
  (text_content any/stemming "bass drum guitar piano")
)
(prox/unit=word/distance<=3
  (text_content any/stemming "perform write")
  (text_content any/stemming "song")
)
(and
  (or/count>=3
    (text_content any/stemming "album autotune choir concert Graceland musician orchestra percussionist recital rapper song
sing")
    (text_content adj "best original song")
    (text_content adj "best song")
    (text_content adj/stemming "boy band")
    (text_content adj "Bruce Springsteen")
    (text_content adj "country and western")
    (text_content adj "early music")
    (text_content adj "Elvis Presley")
    (text_content adj "favorite group")
    (text_content adj "featured music")
    (text_content adj/stemming "girl group")
    (text_content adj/stemming "Grammy Award*")
    (text_content adj "heavy metal")
    (text_content adj "movie and music")
    (text_content adj/stemming "music festival")
    (text_content adj "music business")
    (text_content adj/stemming "music star")
    (text_content adj/stemming "pop group")
    (text_content adj/stemming "pop music")
    (text_content adj/stemming "pop star")
    (text_content adj/stemming "punk rock")
    (text_content adj/stemming "rap star")
    (text_content adj/stemming "recording artist")
    (text_content adj "rock and roll")
    (text_content adj/stemming "rock band")
    (text_content adj/stemming "rock star")
    (text_content adj "singing concert")
    (text_content adj "song-writer*")
  )
)
(not
  (title = "summary")
  (title = "diary")
  (title = "schedule")
  (title = "fortress")
  (title = "soccer")
  (title adj "Saudi prince")
  (title adj "top jobs")
  (title = "BREXIT")
  (title = "trauma")
)

```

)  
)

## National or Ethnic Minority

(and

(or

(title = "minorit\*")  
(title = "jewish")  
(title = "discrimination")  
(title = "bias\*")  
(title adj "civil rights")  
(title adj "lgbt rights")  
(title adj "same-sex")  
(body = "minorities")  
("Rohingya")  
("KKK")  
(body adj "ku klux klan")  
(body adj "issues of race")  
(body adj "religious minorit\*")  
(body adj "religious group\*")  
(body adj "ethnic minorit\*")  
(body adj "minority group\*")  
(body adj "latino communit\*")  
(body adj "black communit\*")  
(body adj "hate crime\*")  
(body adj "lavender scare")  
(body adj "racial discrimination")  
(body adj "racial slur\*")  
(body adj "target\* minorit\*")  
(body adj "ban on muslims")  
(body adj "muslim ban")  
(body adj "white suprem\*")  
(body adj "white actor")  
(body adj "black lives matter")  
(body adj "martin luther king's")

)

(not

(or

(title = "summary")  
(title = "schedule")  
(title = "factbox")  
(title = "manning")  
(title = "brexit")

)

)

)

## Nuclear Power

(and

(or

(title = "reactors")  
(title adj "nuclear power")  
(title adj "nuclear deal")  
("fukushima")

```

("stuxnet")
(body = "AREVA")
(body adj "nuclear power")
(body adj "nuclear waste")
(body adj "nuclear deal")
(body adj "nuclear market")
(body adj "nuclear agreement")
(body adj "atomic energy")
(body adj "nuclear energy")
(body adj "nuclear reactor")
(body adj "nuclear output")
)
(not
(or
(title = "summary")
(title = "schedule")
(title = "manning")
(title = "aseman")
(title = "neutrality")
(title adj "first jet")
(title adj "western plane")
(title adj "top jobs")
(body adj "power strategies")
)
)
)

```

## Rape

```

(and
(or
(title any "rape raped rapes raping")
(title adj "comfort woman")
(title adj "sex nightmare")
(title adj "sex abuse")
(title adj "sexual assault")
(title adj "child porn")
(body adj "Child sex trafficking")
(body adj "Child porn*")
(body adj "Comfort women")
(body adj "Crimes against women")
("molest*")
("rape")
("rapes")
("raped")
(body adj "Sex abuse")
(body adj "Sexually abusing")
(body adj "sexual assault*")
(body adj "sex* attack*")
(body adj "Sex slave*")
(body adj "Sex traffic*")
(body adj "sexual violence")
)
(not
(body adj "nez rape")
(title = "summary")
(title = "schedule")
)
)

```



```
(title = "shortens")
(title = "sweetheart")
)
)
```

## Real Estate

```
(and
  (or
    (title = "eviction")
    (title = "housing")
    (title = "HUD")
    (title = "mortgage")
    (title = "property")
    (title = "rent")
    (title adj "childhood home")
    (title adj "estate agent")
    (title adj "home price*")
    (title adj "new homes*")
    (title adj "office boom")
    (title adj "real estate")
    (title adj "flats boom")
    (or/count>3
      (text_content adj "real estate")
      (text_content adj "property market")
    )
    (body = "housebuilder*")
    (text_content adj "home prices")
    (text_content adj "home sales")
    (text_content adj "housing market")
    (text_content adj "new-home sales")
    (body adj "over land use")
    (body adj "hot property market")
    (body adj "long-term accommodation")
  )
  (not
    (title adj "wrapup")
  )
)
```

## Retirement

```
(and
  (or
    (title = "retirement")
    (title = "pension")
    (title = "pensioners")
    (title = "elderly")
    (body = "aarp")
    (body adj "ageing population")
    (body adj "pension system")
    (or/count>3
      (text_content = "retirement")
    )
  )
)
```

```

(not
  (or
    (title = "summary")
    (title = "cricket")
    (title = "athletics")
  )
)
)

```

## Soccer

```

(and
  (or
    (title adj "african nations cup")
    (title adj "cup of nations")
    (title any "bundesliga concacaf fifa midfielder soccer striker")
    (title adj/literal "m.l.s")
    (title adj "champions league")
    (title adj "confederations cup")
    (title adj "copa america")
    (title adj "copa libertadores")
    (title adj "english football league")
    (title adj "fa cup")
    (title adj "football association")
    (and
      (title adj "gold cup")
      (not
        (text_content any "horse horseracing")
        (text_content adj "horse racing")
      )
    )
  )
  (title adj "la liga")
  (title adj "ligue 1")
  (title adj "major league soccer")
  (title adj "premier league")
  (title adj "serie a")
  (and
    (title adj "world cup")
    (body = "soccer")
  )
)
(title adj "women's world cup")
(or/countunique>2
  (body adj "african footballer of the year")
  (body adj "african nations cup")
  (body any "midfielder qualifier qualifying soccer striker")
  (body adj/literal "m.l.s.")
  (body adj "athletic bilbao")
  (body adj "atletico madrid")
  (body = "arsenal")
  (body adj "bayern munich")
  (body adj "borussia dortmund")
  (body = "bundesliga")
  (body adj "cambridge united")
  (body adj "champions league")
  (body = "chelsea")
)

```

```

(body adj "clint dempsey")
(body = "concacaf")
(body adj "confederations cup")
(body adj "copa america")
(body adj "copa libertadores")
(body adj "cristiano ronaldo")
(body adj "english football league")
(body = "everton")
(body adj "fa cup")
(body = "fifa")
(body adj "football association")
(body = "juventus")
(body adj "la liga")
(body adj "leeds united")
(body adj "ligue 1")
(body adj "leicester city")
(body adj "lionel messi")
(body = "liverpool")
(body adj "major league soccer")
(body adj "manchester city")
(body adj "manchester united")
(body adj "national team")
(body adj "olympique lyonnais")
(body adj "paris st germain")
(and
  (body adj "premier league")
  (not
    (body = "cricket")
  )
)
)
(body adj "real madrid")
(body adj "riyad mahrez")
(body adj "serie a")
(body = "villarreal")
(body adj "west ham united")
(and
  (body adj "world cup")
  (not
    (text_content any "cricket golf hockey rugby ski skier skiing")
  )
)
)
)
// below is specific to reuters corpora //
(not
  (title adj "reuters sports news summary")
  (title adj "reuters sports schedule")
  (title adj "world news schedule")
)
)

```

## Summit Meetings

```

(and
  (or
    (title adj "Top economic events")
    (title = "summit")
  )
)

```

```

    (body adj "Abe meeting Putin")
    (body adj "Annual summit")
    (body adj "G20 conference")
    (body adj "G20 summit")
    (body adj "G7 nations")
    (body adj "G7 partners")
    (body adj "Hot spring summit")
    (body adj "Putin Abe")
    (body adj "Summit at")
    (body adj "summit convened")
    (body adj "Summit meeting")
    (body adj "Summit preparations")
  )
  (not
    (title = "summary")
    (title = "schedule")
    (title = "tennis")
    (title = "cycling")
  )
)

```

## Taxation

```

(and
  (or
    (title = "tax*")
    (title = "tariff*")
    (title = "levy")
    ("supertax*")
    (body adj "stamp duty")
    (body adj "border tax")
    (body adj "tax reform")
    (body adj "lower taxes")
    (body adj "tax disputes")
    (body adj "tax payments")
    (body adj "tax and regulation")
  )
  (not
    (title = "summary")
    (title = "schedule")
    (title adj "tax return*")
  )
)

```

## Television

```

(and
  (or
    (body adj "ABC show")
    (body adj "airs on NBC")
    (body adj "celebrity apprentice")
    (body adj "comedy series")
    (body adj "coming to television")
    (body adj "drama series")
  )
)

```

```

(body adj "gets on tv")
(body adj "new episodes")
(body adj "reality tv")
(body adj "returning to tv")
(body adj "show had aired")
("Sitcom")
(body adj "sky arts")
(body adj "sports channel")
(body adj "television personality")
(body adj "television sports")
(body adj "television star")
(body adj "the x factor")
(body adj "tv comedy")
(body adj "tv drama")
(body adj "tv show")
(body adj "tv star")
(body adj "urban myths series")
(body adj "U.S.television")
(body adj "youtube views")
)
(not
  (title = "summary")
  (title adj "world news schedule")
  (body adj "star donald trump")
  (and
    (title = "Trump")
    (or
      (body adj "tv star")
      (body adj "television star")
    )
  )
)
)
)
)

```

## Tourism

```

(and
  (or
    (title = "tourist*")
    (title = "tourism")
    (title adj "ski resort")
    (body adj "Budget tour*")
    (body adj "Curious visitors")
    ("Graceland")
    ("Kaikoura")
    (body adj "Holiday destination*")
    (body adj "Holiday island*")
    (body adj "Most visited")
    (body adj "Package tour*")
    ("SeaWorld")
    (body adj "Travel agents")
    (body adj "Visiting celebrities")
    (body adj "tourist trail")
    (body adj "tourist income")
    (body adj "tourism industry")
    (body adj "tourist region")
    (body adj "lapland tourism")
  )
)

```

```

    (body adj "new landmark")
    (body adj "Chinese tourists")
  )
  (not
    (or
      (body adj "tourist resort town")
      (title = "summary")
      (title = "schedule")
      (title = "gambia*")
      (title adj "expenses scandal")
      (title = "kardashian")
      (title adj "bpm music")
      (title adj "bpm festival")
      (title adj "honey trap*")
    )
  )
)

```

## IPTC Media Topics - German

### Aktien (Stocks)

```

(or
  (text_content any/stemming "Börse")
  (title = "Leitbörsen")
  (title adj "Euro-Stoxx-50")
  (title adj "Die zehn größten Kursgewinner und -verlierer des Tages")
  (title adj "Die zehn größten Kursgewinner und -verlierer der Woche")
)

```

### Arbeitsmarktdaten (Employment statistics)

```

(and
  (or/countunique>3
    (title any/stemming "Arbeitslose")
    (title = "Arbeitslosigkeit")
    (title any "Arbeitsmarkt")
    (text_content any/stemming "Arbeitslosenzahl")
    (text_content any/stemming "Ausbildungserhebung")
    (text_content any/stemming "Jobchancen")
    (text_content any/stemming "Wachstumsperspektiven")
    (text_content any/stemming "Wirtschaftsforschungsinstitut")
    ("AMS")
    ("Arbeitskräfteerhebung")
    ("Arbeitslosenquote")
    ("Arbeitslosenrate")
    ("Arbeitslosigkeit")
    ("Arbeitsmarktdaten")
    ("Arbeitsmarktintegration")
    ("Arbeitsmarktintegration")
    ("Arbeitsmarktservice")
  )
  (or
    ("Statistik")
    ("Prozent")
  )
)
(prox/unit=sentence/distance=1
  ("steigt")
  ("Arbeitslosigkeit")
)

```

```
)  
)  
)
```

## Auszeichnung und Preis (Awards)

```
(or  
  // examples of awards //  
  (text_content adj "Amadeus Austrian Music Awards")  
  (text_content adj "Kunstpreis des Stifts Klosterneuburg")  
  (text_content adj "Presidential Early Career Awards for Scientists and Engineers")  
  (text_content adj "Sony World Photography Awards")  
  (text_content adj "Statuette des Heiligen Leopold in Bronze")  
  (text_content adj "Wappenmedaille in Gold")  
  (text_content adj/stemming "Goldenen Ehrenzeichen für Verdienste um die Republik")  
  (title adj "Austrian Fashion Awards")  
  (title adj "Autorenpreis der Theaterallianz")  
  (title adj "Ballon d'Or")  
  (title adj "Bank Austria Kunstpreis")  
  (title adj "Deutscher Medienpreis")  
  (title adj "Ehrenzeichen des Landes Tirol")  
  (title adj "ERC-Preisträger")  
  (title adj "Europasportler des Jahres")  
  (title adj "FIFA-Gala")  
  (title adj "Forbes-30 unter 30")  
  (title adj "Gatterer-Preis")  
  (title adj "Heimo-Erbse-Preis")  
  (title adj "Heiner-Carow-Preis")  
  (title adj "Herbert-Walther-Preis")  
  (title adj "Hörspiel des Jahres")  
  (title adj "Kreisky-Preis")  
  (title adj "Oscar-Verleihung")  
  (title adj "Pressefotos des Jahres")  
  (title adj "Pritzker-Preis")  
  (title adj "Rothschild-Preis für Wirtschaftspublizistik")  
  (title adj "Schandfleck des Jahres")  
  (title adj "Schauspieler des Jahres")  
  (title adj "Schweizer Filmpreis")  
  (title adj "Schwimmer des Jahres")  
  (title adj "Teddy Award")  
  (title adj "Turnsportler des Jahres")  
  (title adj "UEFA-Team des Jahres")  
  (title adj "UNESCO-Welterbe")  
  (title adj "Vogelsang-Staatspreis")  
  (title adj "Weltfußballer des Jahres")  
  (title adj "Weltfußballer-Gala")  
  (title adj "Wissenschaftler des Jahres")  
  (title adj/stemming "Athleten des Jahres")  
  (title adj/stemming "Consolidator-Preis")  
  (title adj/stemming "Deutscher Krebspreis")  
  (title adj/stemming "Ehrenmitglied der Kunstuni Graz")  
  (title adj/stemming "Golden Globes")  
  (title adj/stemming "Großen Schauspielpreis")  
  (title adj/stemming "internationaler Wettbewerb zur politischen Bildung")  
  (title adj/stemming "Kinder- und Jugendbuchpreise")  
  (title adj/stemming "Oscar-Nominierungen")  
  (title adj/stemming "Österreichische Kunstpreise")  
  (title adj/stemming "Österreichischer Filmpreis")  
  (title adj/stemming "Radiopreise der Erwachsenenbildung")  
  (title any "Friedensnobelpreis Grammy Grammys Mozartwettbewerb Oscars Weltfußballer")  
)
```

## Betrug (Fraud)

```
(or
  // sufficiently specific terms in title //
  (title any/stemming "Betrug Betrugsfall Betrugsprozess Betrugsverdacht Betrugsversuch Identitätsdieb Kreditkartenbetrug
  Neffentrickbetrüger Telefonbetrug Trickbetrüger Wettbetrug Untreue")
  (title adj "Neffentrick-Opfer")
  (subtitle any "Falschgeld Schwindel")
  (text_content adj/stemming "gewerbsmäßigen Betruges")
  // terms related to fraud that need qualification //
  (and
    (text_content any/stemming "Fälschung")
    (text_content any/stemming "Geldschein")
  )
  (or/countunique>1
    ("Betrug")
    ("Betrüger")
    ("Betrügerin")
    (text_content adj "Fake-Internetshop")
    (text_content adj "Fake-Shop")
    (text_content adj "Untreue-Prozess")
    (text_content adj "Wett-Manipulation")
    (text_content adj/stemming "Love-Scam")
    (text_content adj/stemming "Love-Scam-Fälle")
    (text_content any/stemming "Abgabenhinterziehung")
    (text_content any/stemming "Autoschieberbande")
    (text_content any/stemming "betrogen")
    (text_content any/stemming "Betrügerei")
    (text_content any/stemming "betrügerisch")
    (text_content any/stemming "Betrugsfall")
    (text_content any/stemming "Betrugsprozess")
    (text_content any/stemming "Betrugsverdacht")
    (text_content any/stemming "Falschgeldstatistik")
    (text_content any/stemming "Fälschung")
    (text_content any/stemming "Finanzpolizisten")
    (text_content any/stemming "Geldwäschevorwurf")
    (text_content any/stemming "Heiratsschwindler")
    (text_content any/stemming "Identitätsdieb")
    (text_content any/stemming "Identitätsbetrüger")
    (text_content any/stemming "Kreditkartenbetrug")
    (text_content any/stemming "Neffentrick")
    (text_content any/stemming "Neffentrickbetrüger")
    (text_content any/stemming "Pfusch")
    (text_content any/stemming "Scheinidentität")
    (text_content any/stemming "Schwarzgeldzahlungen")
    (text_content any/stemming "Schwindel")
    (text_content any/stemming "Telefonbetrug")
    (text_content any/stemming "Trickbetrüger")
    (text_content any/stemming "unterschlagen")
    (text_content any/stemming "Untreue")
    (text_content any/stemming "Urkundenfälschung")
    (text_content any/stemming "Wettbetrug")
    (text_content any/stemming "Wettmanipulation")
    (text_content any/stemming "Zechprellerei")
    (prox/unit=sentence/distance=1
      (body any/stemming "falsch")
      (body any/stemming "Kreditkarte")
    )
  )
)
```



## Computerkriminalitaet (Computer Crime)

```
(or
  (and
    // specific forms of attacks //
    (or
      (text_content adj/stemming "Cyber-Attacke")
      (text_content adj/stemming "Cyber-Betrüger")
      (text_content adj/stemming "DDOS-Attacke")
      (text_content any "Cyberangriff Cyberattacke Cybermobbing Datendiebstahl")
    )
    // exclusion terms //
    (not
      (text_content = "Börse")
      (title adj "W I S S E N S C H A F T - B I L D U N G - T A G E S P R O G R A M M")
    )
  )
  // terms related to a specific attack on an organization which need to appear in the same sentence //
  (prox/unit=sentence/distance=1
    ("FACC")
    ("Cyberangriff")
  )
  // terms related to a form of attack that appear in close proximity within the same sentence //
  (prox/unit=sentence/distance=1
    ("Missbrauch")
    ("Daten")
  )
)
```

## Datenschutz (Data Protection)

```
(or
  // all quite specific terms that are quite likely only to appear in the context of data protection //
  (prox/unit=sentence/distance=1
    (body = "Abkommen")
    (or
      (body adj "Privacy Shield")
      (body adj "Safe Harbor")
    )
  )
  (text_content adj "Datenschutz-Reform")
  (text_content adj "datenschutzrechtliche Grundlage")
  (text_content adj "personenbezogene Daten")
  (text_content adj "Handy-Standortdaten")
  (text_content adj "Recht auf Privatsphäre")
  (text_content adj "zentrales Kontenregister")
  (text_content any "Informationsfreiheitsgesetz Verkehrsdaten Auskunft Vorratsdatenspeicherung")
  (text_content any/stemming "Datenschutzanliegen Datenschutzbeauftragte Datenschutzbehörde Datenschutzgedanke
  Datenschutzgruppe Datenschutzpaket Datenschutzrecht Datenschutzregelung")
  // also interesting terms, in APA test corpora they had a negative impact on precision though //
  // ( " Datenschutzgesetz " ) //
  // ( body adj " persönlichen Daten " ) //
  // ( " Datenschutz " ) //
)
```

## Drogenhandel (Drug trafficking)

```
(or
  (or/countunique>2
    // term drug trafficking or synonyms / near synonyms //
    (text_content any/stemming "Dealer")
  )
)
```

```

(text_content any/stemming "Drogen")
(text_content any/stemming "Drogendeal")
(text_content any/stemming "Drogenhandel")
(text_content any/stemming "Drogenhändler")
(text_content any/stemming "Drogenmarkt")
(text_content any/stemming "Drogenmärkte")
(text_content any/stemming "Drogenschmuggler")
(text_content any/stemming "Suchtgift")
(text_content any/stemming "Suchtgifthandel")
(text_content any/stemming "Suchtgiftverkäufen")
// drugs //
("Cannabis")
("Haschisch")
("Heroin")
("Kokain")
("Marihuana")
("Phenetylamin")
// compound nouns related to drugs //
(text_content any/stemming "Cannabisblüte")
(text_content any/stemming "Cannabisharz")
(text_content any/stemming "Cannabiskraut")
(text_content any/stemming "Cannabispäckchen")
(text_content any/stemming "Cannabispflanze")
(text_content any/stemming "Cannabisstaude")
(text_content any/stemming "Hanfplantage")
// terms related to police / judge activities //
(text_content adj/stemming "Drogen-Notfälle")
(text_content adj "Kokain-Fund")
(text_content any/stemming "Drogenfund")
(text_content any/stemming "Haftbefehl")
(text_content any/stemming "Haftstrafe")
(text_content any/stemming "Suchtgiftermittler")
(text_content any/stemming "Suchtmittelspürhund")
("EMCDDA")
(text_content any/stemming "Drogenbeobachtungsstelle")
)
)

```

## Einwanderung (Immigration)

```

(or
(text_content any "Immigration Einwanderung Schengen Flüchtling Flüchtlinge")
)

```

## Essen und Trinken (Food)

```

(or
// specific types of food //
(or
(text_content adj "American-Burger")
(text_content adj "Bio-Forelle")
(text_content adj "Coca Cola-Life")
(text_content adj "Diät-Shakes")
(text_content adj "Hauswürstel nach Kärntner Art")
(text_content any/stemming "Crepes Cupcakes Heringssalat Heringsschmaus Krapfen Olivenöl Taubengericht")
)
// food fairs //
(or
(text_content adj "Street Food Festival")
("Genussmesse")
)
// other food-related term ( s ) //
(or

```

```
(text_content adj/stemming "Food-Blog")
)
(or
(title adj "Cafe Willendorf")
)
)
```

## Feuer (Fire)

```
(or
  // non-specific terms //
  (or/countunique>1
    // individual any-stemming statements to be able to trigger countunique more than once //
    ("Berufsfeuerwehr")
    ("Brand")
    ("Brandstiftung")
    ("Brandursache")
    ("Einsatzfahrzeug")
    ("Feuer")
    ("Feuerwehr")
    ("Verbrennungen")
  )
  (or
    // fire in title //
    (title = "Brand")
    (title = "Feuerwehr")
    // specific types of fire //
    (text_content any "Kleinbrand Wohnhausbrand Wohnungsbrand")
    // other specific terms //
    (text_content adj "Waldbrand-Alarm")
    (text_content any "Brandstifter Einsatzjacke Rauchgas Rauchgasvergiftung Serienbrandstifter Zigarettenglut")
  )
)
```

## Film (Cinema)

```
(or
  (title = "Oscars")
  // cinema awards //
  (text_content adj "Academy Award")
  (text_content adj "Oscar-Verleihung")
  (text_content adj "Tiger Award")
  (text_content adj "Teddy Award")
  ("Berlinale")
  // action movie franchise //
  (text_content adj "Action-Reihe")
  (and
    (or
      // relation to terms including German word for cinema //
      (text_content any "Kino Kinofilm Kinopremiere")
    )
    // cinema-related terms //
    (or
      (text_content any/stemming "Dokumentarfilm Drehbuch Drehbücher Filmadaption Filmfestival Filmfestspiel Filmkomponist Filmmusik Filmreihe filmschaffend Oscar Regie Regisseur")
    )
  )
)
```

## Fußball (Soccer)

```
(and
  (or
    (text_content any "Fußball Torwette")
  )
)
```

```

// exclusion //
(not
  (title adj "APA - S P O R T N A C H R I C H T E N")
)
)

```

## Gewerkschaft (Unions)

```

(and
  (or
    // German word for union in title or subtitle //
    (title any/stemming "Gewerkschaft")
    (subtitle any/stemming "Gewerkschaft")
    // specific unions //
    (title any "ÖGB")
    (subtitle any "ÖGB")
    (title adj "Spielergewerkschaft VfF")
    (text_content adj "Gewerkschaft Bau-Holz")
    (text_content adj "Gewerkschaft der Privatangestellten")
    (text_content adj "GPA-djp")
    (prox/unit=sentence/distance=1
      ("Ärztegewerkschaft")
      ("Asklepios")
    )
  )
)
// exclusion terms, mainly related to general news //
(not
  (title adj "APA Medien News")
  (title adj "Sieben Tage Wirtschaft")
)
)

```

## Golf

```

(or
  (and
    (or
      // golfers //
      (or
        (text_content adj "Rory McIlroy")
        (text_content adj "Bernd Wiesberger")
        (text_content adj "Rickie Fowler")
        (text_content adj "Martin Kaymer")
        (text_content adj "Gary Stal")
        (text_content adj "Lukas Nemezc")
      )
      // golf tournaments //
      (or
        (text_content adj "Eurasia Cup")
        (text_content adj "US-PGA-Tour")
        (text_content adj "FedEx-Cup")
      )
    )
  )
  // exclusion terms //
  (not
    (title adj "APA - S P O R T N A C H R I C H T E N")
  )
)
)

```

## Homosexualität (Gays and lesbians)

```
(or
  (title = "Homosexuelle")
  (subtitle = "Homosexuelle")
  (title adj "Homo-Rechte")
  (subtitle adj "Homo-Rechte")
  (title adj/stemming "Homo-Familie")
  (subtitle adj/stemming "Homo-Familie")
  (title adj/stemming "Homo-Partnerschaft")
  (subtitle adj/stemming "Homo-Partnerschaft")
  (title adj/stemming "Homo-Politiker")
  (subtitle adj/stemming "Homo-Politiker")
  (title adj/stemming "Homosexuellenverbände")
  (title = "Ampelpärchen")
  ("Homosexualität")
  (text_content adj "Schwulen- und Lesbenhaus")
  (title any/stemming "Lebenspartnerschaft")
  (prox/unit=sentence/distance=1
    ("homosexuelle")
    ("Paare")
  )
)
```

## Immobilien (Real Estate)

```
(and
  (or
    // terms in real estate business //
    (or/countunique>1
      (title adj "Immo-Kursziel")
      (text_content adj/stemming "privater Wohnbau")
      (text_content adj "Immo-Branche")
      (text_content adj "Immo-Deal")
      (text_content adj "Immo-Entwickler")
      (text_content adj "Immo-Treuhänder")
      (text_content adj "Immobilien-Lagen")
      (text_content adj "Immobilien-Verkauf")
      (text_content adj "Immoinvest-Volumen")
    )
    // individual any-stemming statements to be able to trigger countunique more than once //
    (text_content any/stemming "Immobiliendeal")
    (text_content any/stemming "Immobilienentwicklung")
    (text_content any/stemming "ImmoFirma")
    (text_content any/stemming "Immobilien dienstleister")
    (text_content any/stemming "Wohnbaupaket")
    (text_content any/stemming "Immobilien gesellschaft")
    (text_content any/stemming "Immobilienmakler")
    (text_content any/stemming "Immobilienfirma")
    (text_content any/stemming "Immobilienpreise")
    (text_content any/stemming "Immobilienmarktplatz")
    (text_content any/stemming "Immobilienentwickler")
    (text_content any/stemming "Immobilienentwicklers")
    (text_content any/stemming "Immobilientreuhänder")
    (text_content any/stemming "Immobilienbesitz")
    (text_content any/stemming "Immobilienfond")
    (text_content any/stemming "Immobilienfond")
    (text_content any/stemming "Einzelhandelsimmobilien")
    (text_content any/stemming "Immoinvestmentgruppe")
    (text_content any/stemming "Immobilientochter")
    (text_content any/stemming "Wohnimmogesellschaft")
  )
)
```

```

(text_content any/stemming "Einzelhandelsimmobilien")
(text_content any/stemming "Hauspreise")
(text_content any/stemming "Immobilienmarkt")
(text_content any/stemming "Immobilienmanagement")
(text_content any/stemming "Immobilienbereich")
(text_content any/stemming "Kaufhausprojekt")
(text_content any/stemming "Immobilienleasing")
(text_content any/stemming "Immobilieninvestition")
(text_content any/stemming "Immobilienentwickler")
(text_content any/stemming "Luxusimmobilien")
(text_content any/stemming "Immobilientreuhänder")
(text_content any/stemming "Wohnbau")
(text_content any/stemming "Fertighausbranche")
(text_content any/stemming "Wohnimmobilien")
(text_content any/stemming "Bauträgerverträge")
(text_content any/stemming "Wohnungskäufer")
(text_content any/stemming "Wohnungskauf")
(text_content any/stemming "Wohnbauoffensive")
(text_content any/stemming "Wohnbaustadtrat")
(text_content any/stemming "Gemeindewohnung")
)
// real estate companies //
(or
  (prox/unit=sentence/distance=1
    (or
      (body any "Baufirma Immobilie Immobilienentwicklers Immobilienentwickler Porr")
    )
    (body = "UBM")
  )
  (text_content adj "6B47-Gruppe")
  (text_content adj "Atrium European Real Estate Limited")
  (text_content adj "Adler Real Estate")
  (text_content adj "CA Immo")
  (text_content adj "CA Immobilien Anlagen AG")
  (text_content adj "EHL Immobilien")
  // typo in an article //
  (text_content adj "EHL Immobilien")
  (text_content adj "EHL-Immobilien-Gruppe")
  (text_content adj "Finanzgruppe Imperial")
  (text_content adj "s-Immo-Gruppe")
  (text_content adj "s-Immo")
  (text_content adj "Wiener Immobilienrendite AG")
  (text_content adj "UBM Development AG")
  (text_content any "6B47 Atrium Buwog CBRE conwert Gesiba Immofinanz")
)
// exclusion terms, mostly related to stock market //
(not
  (title adj "Wiener Rentenmarkt")
  (title adj "Wiener Börse")
  (title adj "Devisen Wien")
  (title adj "Devisen USA")
  (title adj "US-Börsen")
  ("Aktienmarkt")
  (title = "Kursgewinner")
  (title adj "Die zehn größten Kursgewinner und -verlierer des Tages")
  (title adj "Die zehn größten Kursgewinner und -verlierer der Woche")
  (title adj "W I S E N S C H A F T - B I L D U N G - TAGESPROGRAMM")
  (title adj "Sieben Tage Wirtschaft")
  (title adj "Sieben Tage Innenpolitik")
)
)

```

## Insolvenz (Bankruptcy)

```
(or
  (and
    // entities that can go bankrupt //
    (or
      (text_content adj "Fußball-Zweitligist")
      (text_content any/stemming "Academy Akadmie Baufirma Baufirmen Handelskette Handelsunternehmen
Lebensmittelhandelskette Modehändler Supermarktkette Tochterunternehmen Unternehmen")
    )
    // general terms related to bankruptcy //
    (or
      (text_content any/stemming "Ausfallwahrscheinlichkeit Firmenpleite insolvent Insolvenz Konkurs pleite pleitegegangen
Sanierungsplan")
    )
  )
  // specific terms related to bankruptcy //
  (or
    (text_content any "Insolvenzrahmen Konkursverfahren")
    (text_content any/stemming "Insolvenzantrag Insolvenzursache Insolvenzverwalter Insolvenzverwalterin Masseverwalter
Privatinsolvenz Privatkonkurs")
  )
)
```

## Islam

```
(and
  (or
    (title any/stemming "Burka")
    (title any/stemming "Islam")
    (title any/stemming "islamisch")
    (title any/stemming "Jihadist")
    (or/countunique>1
      (text_content adj "Al-Azhar-Universität")
      (text_content adj "IS-Kämpfer")
      (text_content adj "Islamic Banking")
      (text_content adj/stemming "Islamischer Staat")
      (text_content any/stemming "Beschneidung")
      (text_content any/stemming "Islamgesetz")
      (text_content any/stemming "islamisch")
      (text_content any/stemming "Islamisierung")
      (text_content any/stemming "Jihadist")
      (text_content any/stemming "Moschee")
      (text_content any/stemming "Muslim")
      (text_content any/stemming "muslimisch")
      (text_content any/stemming "Schiiten")
      (text_content any/stemming "schiitisch")
      (text_content any/stemming "Sunnit")
      (text_content any/stemming "sunnitisch")
      (text_content any/stemming "Wahhabit")
      ("IGGiÖ")
    )
  )
  // exclusion terms //
  (not
    (title adj "Historisches Kalenderblatt")
    (title adj "Sieben Tage")
  )
)
```



## Kartellfragen (Antitrust Issue)

```
(or
  (title any/stemming "Bundeswettbewerbsbehörde Bundeswettbewerbbehörde BWB Kartellstrafe Wettbewerbsbehörde
Wettbewerbshüter")
  (and
    (text_content any "Konsortium Übernahme Wettbewerbsbedenken")
    (text_content any "Bundeswettbewerbsbehörde BWB")
  )
)
```

## Leichtathletik (Athletics)

```
(or
  // topic name in title //
  (title = "Leichtathletik")
  // below roughly equates to above //
  (and
    (or/countunique>1
      // topic name needs to be in text so other terms are more likely to be relevant //
      ("Leichtathletik")
      // athletes //
      (or
        (text_content adj "Antonio Jimenez")
        (text_content adj "Jennifer Wenth")
        (text_content adj "Lukas Weißhaidinger")
        (text_content adj "Sarah Lagger")
        (text_content adj "Dominik Hufnagl")
        (text_content adj "Andreas Vojta")
        (text_content adj "Victor Salazar")
        (text_content adj "Ivan Hierro")
        (text_content adj "Beate Schrott")
      )
      // competitions //
      (or
        (text_content adj "Austrian Athletics Award")
        (text_content adj "Golden Fly Series")
        (text_content adj "Gugl-Indoor-Meeting")
      )
      // athletic unions international / national //
      (or
        (text_content adj "Europäischen Leichtathletikverbandes")
        (text_content adj "Österreichischen Leichtathletik-Verbandes")
        (text_content adj "Internationalen Leichtathletik-Verbandes")
        (text_content adj "Weltverband der Leichtathleten")
      )
      // union officials //
      (or
        (text_content adj "Rune Andersen")
        (text_content adj "Sebastian Coe")
        (text_content adj "Svein Arne Hansen")
        (text_content adj "Ralph Vallon")
      )
    )
  )
  // exclusion terms - do not consider sports news //
  (not
    (title adj "APA - S P O R T N A C H R I C H T E N")
    (title adj "SPORT - TAGESPROGRAMM")
  )
)
```

## Medizinische Forschung (Medical Research)

```
(and
  (or
    (or
      // research- and hospital-related //
      (prox/unit=sentence/distance=1
        ("Resistenzen")
        ("Antibiotika")
      )
      (text_content adj "Arachidonyl-Glyzerol-Spiegel")
      (text_content adj "Austrian Stroke Prevention Study")
      (text_content adj "Biotech-Medikament")
      (text_content adj "Deutschen Krebspreis")
      (text_content adj "Medizinische Wissenschaft")
      (text_content adj "Monoglyzerid-Lipase")
      (text_content adj "Phase-I-Untersuchungen")
      (text_content adj "Smart Drugs")
      (text_content adj "Symposiums zur Suchterkrankung")
      (text_content all/stemming "Erforschung Erkrankung")
      (text_content all/stemming "Wissenschaft Medizin")
      (text_content any/stemming "Elementarteilchendetektor Impfplan Molekularbiologie Organerhalt Radiologenkongress")
      (text_content any "Bioethikkommission Doppelorgantransplantation Herzschrittmacher Krankenhaushygiene Krebstherapie
Mammografieprogramm Medikationsmanagement Neurowissenschaftler Onkologie Tropenmedizin Weltkrebstag")
    )
    (or/countunique>1
      // general terms //
      (text_content any/stemming "Forschung Studie")
      // disease-related //
      (text_content adj "Zika-Virus-Infektion")
      (text_content any/stemming "Leukämieerkrankung Lungenkarzinom Suchterkrankung")
      (text_content any "Cholera Diabetes Herzinfarkt Masern Mikrozephalie Prostatakarzinom Psoriasis Schlaganfallgefahr")
      ("Darmkrebsrisiko")
      ("Krebskranke")
      // related to immune system //
      (text_content adj "T-Zell-Rezeptoren")
      (text_content any "Acinetobacter Antikörper Immunsystem Immunabwehr Krankenhausinfektion")
      // bacteria- or virus-related //
      (text_content adj "Acinetobacter-Keimes")
      (text_content adj "Campylobacter-Keime")
      (text_content adj "multiresistenten Keim")
      // hospitals or medical institutes //
      (text_content adj "Europäischen Zentrums für Krankheitskontrolle")
      (text_content adj "Med-Uni")
      (text_content adj "Medizin-Fakultät")
      (text_content adj "MedUni Wien")
      (text_content adj "Wiener Wilhelminenspitale")
      (text_content any/stemming "Universitätsklinik")
    )
  )
  (not
    (title adj "W I S S E N S C H A F T - B I L D U N G - T A G E S P R O G R A M M")
    (text_content any/stemming "Aufnahmeverfahren")
  )
)
```

## Militaerische Ausruestung (Military Equipment)

```
(or
  // quite specific types military equipment, here mostly related to grenades //
  (body adj "Flak-Granate")
  (body adj "Steyr-AUG-Sturmgewehren")
  (text_content any/stemming "Eihandgranate Fliegerbombe Gewehrgranate Granatenexport Luftabwehrsystem Splittergranate")
  // tests of nuclear or h bombs //
  ("Atombombentest")
  (prox/unit=sentence/distance=1
    ("Wasserstoffbombe")
    ("Test")
  )
  (prox/unit=sentence/distance=1
    ("Wasserstoffbombe")
    ("getestet")
  )
  // specific treaties / laws related to military equipment //
  (body adj "Comprehensive Nuclear-Test-Ban Treaty")
  ("Kriegsmaterialengesetz")
  // export of military equipment //
  ("Waffenexporte")
)
```

## Mode (Fashion)

```
(or
  (and
    (or
      (text_content any/stemming "Luxuskonzern Mode Modedesigner Modegruppe Modehändler Modehaus Modelabel
Modesatire Modeschau Modeschöpfer")
    )
    // fashion brands //
    (or
      (text_content adj "Dolce&Gabbana")
      (text_content adj "Dolce & Gabbana")
      (text_content adj "Louis Vuitton")
      (text_content adj "Roger Vivier")
      (text_content any "Luxottica Prada Safilo Swarovski Swatch Texhages")
    )
  )
  // fashion awards //
  (or
    (text_content adj "Austrian Fashion Awards")
    (text_content adj "Berliner Modesalon")
    (text_content adj "Fashion Week Berlin")
    (text_content adj "Mailänder Modewoche")
    (text_content adj "Mido-Messe")
    (text_content adj "Pitti Uomo")
  )
  // season collections //
  (or
    (text_content adj "Frühjahr/Sommer-Kollektion")
    (text_content adj "Frühjahrs-Kollektion")
    (text_content adj "Sommer-Kollektion")
    (text_content adj "Herbst/Winter-Kollektion")
    (text_content adj "Herbst-Kollektion")
    (text_content adj "Winter-Kollektion")
  )
  // designs //
  (or
    (text_content adj "Glitter-Optik")
    (text_content adj "Bohemien-Style")
  )
)
```

```

(text_content adj "Seventies-Style")
(text_content any/stemming "Ethnomuster Zuckerwattefarben")
)
(and
  // star designers ( real or imagined ) //
  (or
    (text_content adj "Alexander McQueen")
    (text_content adj "Lena Hoschek")
    (text_content adj "Marina Hoermanseder")
    (text_content adj "Modezar Mugatu")
    (text_content adj "Nhut La Hong")
  )
  // only trigger rule if star designers not a guest in a tv show or similar //
  (not
    (text_content any/stemming "Gast Stargast")
  )
)
)
// specific accessories //
(or
  (text_content adj "Titan Minimal Art")
)
)

```

## Parlamentswahlen (National Elections)

```

(or
  (title adj "Irland-Wahl")
  (title adj "Slowakei-Wahl")
  (title adj "US-Wahlen")
  (and
    (or
      (title any/stemming "kroatisch mazedonisch")
      (title any/stemming "Afghanistan Iran Kroatien Mazedonien Montenegro Serbien Slowakei Uganda Vanuatu")
    )
    (or
      (title any/stemming "Neuwahl Parlamentswahl Regierungsbildung Regierungsteilnahme Übergangsregierung Wahl
Wahldatum Wahlergebnis Wahlkommission Wahltermin Wahlvoraussetzung")
      (subtitle any/stemming "Neuwahl Parlamentswahl Regierungsbildung Regierungsteilnahme Übergangsregierung Wahl
Wahldatum Wahlergebnis Wahlkommission Wahltermin Wahlvoraussetzung")
    )
  )
)
)

```

## Pharmaindustrie (Pharmaceutical)

```

(and
  (or
    // names of pharmaceutical industry companies //
    (or
      ("Sanochemia")
      ("Boehringer")
      ("Pfizer")
      ("Valneva")
      ("Gilead")
    )
    // terms related to term pharmaceutical industry //
    (or/countunique>1
      ("Pharmaindustrie")
      ("Pharmawirtschaft")
      ("Biotechunternehmen")
    )
  )
)

```

```

        ("Biotechkonzern")
    )
)
// exclusion terms, mainly related to stock market //
(not
    (title = "Börse")
    (title adj "US-Börsen")
    ("Aktienmarkt")
    (title = "Kursgewinner")
    (title = "Markt")
    (title adj "Die zehn größten Kursgewinner und -verlierer des Tages")
    (title adj "Die zehn größten Kursgewinner und -verlierer der Woche")
    (title adj "W I S S E N S C H A F T - B I L D U N G - T A G E S P R O G R A M M")
    (title adj "Sieben Tage Wirtschaft")
)
)

```

## Radsport (Cycling)

```

(or
    (title = "Radsport")
    (or
        (and
            (not
                // not consider news items that have narrower topic //
                (text_content any "Eintagesrennen Querfeldein Massensprint Straßenrennen Zeitfahren")
                // not consider general sports news //
                (title adj "A P A - S P O R T N A C H R I C H T E N")
            )
            (or
                // cyclists //
                (or
                    (text_content adj "Alberto Contador")
                    (text_content adj "Alejandro Valverde")
                    (text_content adj "Christoph Strasser")
                    (text_content adj "Felix Großschartner")
                    (text_content adj "Geraint Thomas")
                    (text_content adj "Jonathan Castroviejo")
                    (text_content adj "Matthias Krizek")
                    (text_content adj "Matthias Brändle")
                    (text_content adj "Tony Martin")
                    (text_content adj "Simon Gerrans")
                    (text_content adj "Tejay Van Garderen")
                )
                // cycling championships //
                (or
                    (text_content adj "Andalusien-Rundfahrt")
                    (text_content adj "Algarve-Rundfahrt")
                    (text_content adj "GP Laguna Porec")
                    (text_content adj "Race Across America")
                    (text_content adj "Tour-de-France")
                )
            )
            // cycling teams //
            (or
                (text_content adj "Felbermayr Simplon Wels")
                (text_content adj "Orica-GreenEdge")
                (text_content = "Movistar")
            )
        )
    )
)
)
)
)

```

## Rente (Pension)

```
(and
  (or
    // terms specifically related to pension //
    (title any/stemming "Beamtenpension Frauenpension Invaliditätspension Pension Pensionsdebatte Pensionistenpreisindex
Pensionsdeal Pensionsantrittsalter Pensionskassen Pensionsreform Pensionsreserve Pensionssystem Pensionstransfer
Vorsorgekasse")
    (subtitle any/stemming "Beamtenpension Frauenpension Invaliditätspension Pension Pensionsdebatte Pensionistenpreisindex
Pensionsdeal Pensionsantrittsalter Pensionskassen Pensionsreform Pensionsreserve Pensionssystem Pensionstransfer
Vorsorgekasse")
  )
  (not
    (text_content any/stemming "Hotel Tourismus")
  )
)
```

## Sexuelle Gewalt (Sexual Assault)

```
(or
  (or/countunique>3
    (text_content adj/stemming "geschlechtlicher Nötigung")
    ("Anklage")
    ("Anzeige")
    ("Attacke")
    ("attackiert")
    ("bedrängt")
    ("erpresst")
    ("Kinderporno")
    ("minderjährig")
    ("Missbrauch")
    ("missbraucht")
    ("Obszönität")
    ("Opfer")
    ("Opferschutz")
    ("Pornofilme")
    ("Sex")
    ("Strafverfahren")
    ("Verdächtiger")
    ("Vergewaltiger")
    ("Vergewaltigung")
    (prox/unit=sentence/distance=1
      ("sexueller")
      ("Belästigung")
    )
  )
  (prox/unit=sentence/distance=1
    ("sexuelle")
    ("Belästigung")
  )
  (prox/unit=sentence/distance=1
    ("sexuell")
    ("belästigt")
  )
  (prox/unit=sentence/distance=1
    ("Frau")
    ("bedrängt")
  )
)
```

```

(prox/unit=sentence/distance=1
  ("unsittlichen")
  ("Berührungen")
)
(prox/unit=sentence/distance=1
  ("Gesäß")
  ("betastet")
)
)
(or
  (text_content adj "unsittlich berührt")
  ("Pädophilie")
  ("Silvesterattacken")
  ("vergewaltigt")
  ("vergewaltigen")
  (prox/unit=sentence/distance=1
    ("Übergriffe")
    ("Frauen")
  )
  (prox/unit=sentence/distance=1
    ("belästigt")
    ("Frauen")
  )
  (prox/unit=sentence/distance=1
    ("sexuellen")
    ("Missbrauchs")
  )
  (prox/unit=sentence/distance=1
    ("sexuellen")
    ("Übergriff")
  )
  (prox/unit=sentence/distance=1
    ("sexueller")
    ("Übergriffe")
  )
  (prox/unit=sentence/distance=1
    ("sexueller")
    ("Übergriff")
  )
  (prox/unit=sentence/distance=1
    ("Rock")
    ("fassen")
  )
)
)
)

```

## Staatshaushalt (Government Budget)

```

(or
  (or
    (text_content adj "Bund der Steuerzahler")
    (title all "Bundesheer Budget")
    (title = "Budgetdisziplin")
  )
  (or/countunique>1
    (title = "Steuerreform")
    (prox/unit=sentence/distance=1
      ("Einnahmen")
      ("Staates")
    )
    (title any "Budget")
    (text_content adj "Einkommenssteuer-System")
    (text_content adj "Kalte Progression")
    ("Budgetsprecher")
  )
)
)

```

```

("Budgetziele")
("Bundesbudget")
("Bundeshaushalt")
("Staatsbudget")
("Steuergerechtigkeit")
("Steuerzahler")
)
)

```

## Staatsverschuldung (Government Debt)

```

(or/countunique>1
  (title adj "10-Jahresrenditen")
  (title adj "10-Jahres-Renditen")
  (title any/stemming "Bundesanleihen")
  (subtitle any/stemming "Bundesanleihen")
  (text_content adj "Euro-Stabilitätspakt")
  ("Bundesanleihen")
  ("Jahresanleihen")
  ("Pensionsreform")
  ("Schuldenabbau")
  ("Schuldenstand")
  ("Staatsschulden")
  ("Staatsverschuldung")
  ("Verschuldung")
  (and
    (text_content any/stemming "Defizit")
    (or
      (text_content any/stemming "Staat")
      (text_content adj/stemming "EU-Staat")
    )
  )
)
)
  (and
    (or
      ("Eurogruppe")
      ("Eurozone")
    )
    ("Griechenland")
  )
)
  (prox/unit=sentence/distance=1
    (or
      (body = "Bundesländer")
      (body = "Regierungen")
      (body adj "Internationalen Währungsfonds")
    )
    (or
      (body = "Finanzierung")
      (body = "Kapitalaufnahme")
      (body = "Kreditabkommen")
    )
  )
)
)
)

```

## Steuer (Taxation)

```

(or
  (or
    (title = "Bankenabgabe")
    (title = "Finanzstrafbehörde")
    (subtitle = "Finanzstrafbehörde")
    (title adj "LKW-Maut")
    (title = "Registrierkassenpflicht")
    (subtitle = "Registrierkassenpflicht")
  )
)

```



```

(title any/stemming "Steuerbehörde")
(title = "Steuerhinterziehung")
("Steuerquote")
("Steuereinnahmen")
)
(or/countunique>1
(text_content adj/stemming "Anti-Steuerbetrugsmaßnahmen")
(text_content adj/stemming "Einkommenssteuer-System")
(text_content adj/stemming "Einkommenssteuer-Stufen")
(text_content adj "EU-Mautharmonisierung")
(text_content adj/stemming "EU-Steuerkommissar")
(text_content adj/stemming "kalten Progression")
(text_content adj "Lkw-Mautpflicht")
(text_content adj/stemming "Maut-Harmonisierung")
(text_content adj/stemming "Mehrwertsteuer-Regel")
(text_content adj/stemming "Mehrwertsteuer-Betrug")
(text_content adj/stemming "MwSt-Regel")
(text_content adj/stemming "Ökostrom-Förderung")
(text_content adj/stemming "Steuervermeidungs-Richtlinie")
(text_content adj "Tobin-Steuer")
("Bankenabgabe")
("Dienstwagenbesteuerung")
("Energiesteuerfreibetrag")
("Erbschaftssteuer")
("Finanzbehörden")
(text_content any/stemming "Finanzpolizist")
("Finanzpolizei")
("Finanzsteuer")
("Finanzstrafbehörde")
("Finanztransaktionssteuer")
("Lohnsteuer")
("Lohnsteuerentlastung")
("Mehrwertsteuerbetrug")
("Maut")
("Mineralölsteuer")
("Negativsteuer")
("Nichtbesteuerung")
("Ökosteur")
(text_content any/stemming "Ökostromgesetz")
("Reverse Charge")
("Schenkungssteuer")
("Schwarzarbeit")
("Sozialdumping")
("Steuerbaukasten")
("Steuerbegünstigung")
("Steuerbehandlung")
("Steuerbehörden")
("Steuerberater")
("Steuerbetrug")
("Steuerdeal")
("Steuerdokumente")
("Steuerentlastung")
("Steuerflucht")
("Steuergeld")
("Steuergerechtigkeit")
("Steuergesetz")
("Steuerkommissar")
("Steuermissbrauch")
("Steuermittel")
("Steueroasen")
("Steuerpflicht")
("Steuerpläne")
("Steuerplanung")
("Steuerrechner")
("Steuerreform")
("Steuerschuldnerschaft")

```

```

("Steuerswellen")
("Steuersystem")
("Steuervermeidung")
("Steuerverschwendung")
("Steuerverwaltung")
("Steuervorteile")
("Steuerwettbewerb")
("Steuerzahler")
("Vorsteuerabzug")
(prox/unit=sentence/distance=1
  (body = "Besteuerung")
  (body = "Dieselkraftstoff")
)
(prox/unit=sentence/distance=1
  (body = "steuerlich")
  (body = "absetzbar")
)
(prox/unit=sentence/distance=1
  (body = "steuerlichen")
  (body = "Begünstigung")
)
(prox/unit=sentence/distance=1
  (body = "angemeldete")
  (body = "Erwerbstätigkeit")
)
(prox/unit=sentence/distance=1
  (body = "Steuer")
  (or
    (body any "Alkohol Strom Zigaretten")
  )
)
)
)
)

```

## Terroranschlag (Act of Terror)

```

(or
  (title any/stemming "Attentat Attentäter Anschlag Anschlagversuch Doppelanschlag Sprengstoffgürtel Terroranschlag")
  (and
    ("Anschlag")
    (or
      (text_content adj "Charlie Hebdo")
      ("Hurghada")
    )
  )
)
)

```

## Umwelt (Environment)

```

(or
  // non-specific terms //
  (or/countunique>1
    (title adj "VW-Abgasskandal")
    (text_content adj "VW-Skandal")
    (text_content adj "CO2-Ausstoß")
    (prox/unit=sentence/distance=1
      ("Umwelt")
      ("Konsumentenschutz")
    )
  )
  // individual any-stemming statements to be able to trigger countunique more than once //
  (text_content any/stemming "Umweltminister")
  (text_content any/stemming "Umweltmesse")
)

```

```

(text_content any/stemming "Umweltschutz")
(text_content any/stemming "Umweltschutzmaßnahmen")
(text_content any/stemming "Abgaswert")
(text_content any/stemming "Umweltsprecher")
(text_content any/stemming "Umweltgesetzgebung")
(text_content any/stemming "Umweltverträglichkeitsprüfung")
(text_content any/stemming "Umweltausschuss")
(text_content any/stemming "Umweltschutzorganisationen")
(text_content any/stemming "Abgasaffäre")
(text_content any/stemming "umweltschädlich")
(text_content any/stemming "Emissionshandelssystem")
(text_content any/stemming "Umwelteffekt")
(text_content any/stemming "Umweltverbrauch")
(text_content any/stemming "Abgasskandal")
(text_content any/stemming "Naturkatastrophe")
(text_content any/stemming "Weltrisikobericht")
(text_content any/stemming "Naturgewalt")
(text_content any/stemming "Umweltkommissar")
(text_content any/stemming "Umweltprogramm")
(text_content any/stemming "Umweltanwalt")
(text_content any/stemming "Stromkennzeichnung")
(text_content any/stemming "Umweltaktivist")
(text_content any/stemming "Umwelthilfe")
(text_content any/stemming "Abgasthematik")
(text_content any/stemming "Risikoeinschätzung")
(text_content any/stemming "Umweltinstitut")
(text_content any/stemming "Grenzwert")
(text_content any/stemming "umweltbelastend")
)

(or
  // terms in title or specific terms //
  (title any/stemming "Abgastest")
  (text_content adj "Benefit Corporation")
  (text_content adj "B Impact Assessment")
  (text_content adj "Jane Goodall")
  (text_content adj "ökologische Kriterien")
  ("Anthropozän")
  ("Umweltsünder")
  ("Donaudeklaration")
  (prox/unit=sentence/distance=1
    ("klimafreundlich")
    ("Verkehr")
  )
)
(and
  (text_content any/stemming "Ölbohrung")
  ("Referendum")
)
)
)

```

## Universität (University)

```

(or
  // words related to university in title and subtitle //
  (title any "Architekturfakultät Doktorat Ehrendoktorat Ehrendoktorat Erasmus Privatuni Professur Uni Unis")
  (title any/stemming "Hochschulstudien Universität Universitätsbibliothek Uniklinikum Universitätsstudium Wissenschaftsrat")
  (title adj/stemming "universitär")
  (subtitle adj/stemming "universitär")
  (subtitle any/stemming "Universität Universitätsstudium")
  (title adj "Uni-Zugang")
  (title adj "Uni-Budget")
  // types of universities and research institutes //
  (title adj/stemming "Technischen Unis")
)

```

```

(subtitle adj/stemming "Technischen Unis")
(title adj "Uni-Kliniken")
(subtitle adj "Uni-Kliniken")
// specific universities and institutes //
(text_content adj/stemming "Internationalen Bodenseehochschule")
(text_content adj "Johannes Kepler Universität")
(title adj "Kunstuni Graz")
(subtitle adj "Kunstuni Graz")
(text_content adj "Kunstuniversität Graz")
(title adj/stemming "Max Planck-Institut")
(subtitle adj/stemming "Max Planck-Institut")
(title adj "Med-Uni Graz")
(subtitle adj "Med-Uni Graz")
(title adj "TU Graz")
(subtitle adj "TU Graz")
(title adj "Universität Graz")
(subtitle adj "Universität Graz")
// university grant fund //
(text_content adj "Wiener Wissenschafts-, Forschungs- und Technologiefonds")
)

```

## Zusatzversicherung (Health Insurance)

```

(and
  (or/countunique>1
    // relevant types of insurance //
    (or
      ("Gesundheitsversicherung")
      ("Pflichtversicherung")
    )
    // specific insurance benefits //
    (or
      (body adj "Hepatitis-Präparat")
      (body adj "Magnetresonanz-Untersuchungen")
      (body adj "Rheumatologen-Termin")
      (text_content any/stemming "Vorsorgeuntersuchung")
      (prox/unit=sentence/distance=1
        ("Versorgung")
        ("Bevölkerung")
      )
    )
  )
  // financial terms related to health insurance //
  (or
    (text_content any/stemming "Ambulanzgebühr Medikamentenkosten Umlageverfahren")
  )
  // contracts related to health insurance //
  (or
    (text_content adj/stemming "Rahmen-Pharmavertrag")
    (text_content adj/stemming "Pharma-Rahmenvertrag")
  )
  (or
    (body adj "Elektronische Gesundheitsakte")
  )
  // entities //
  (or
    (text_content adj "Österreichischen Schmerzgesellschaft")
    (text_content any/stemming "Versicherungsanstalt")
  )
  // other terms //
  (or
    ("Gesundheitswoche")
  )
)
)

```