textpressoapi Documentation

Release 0.1

Valerio

Apr 16, 2020

Contents:

Textpresso API provides functions to query the Textpresso database, a collection of annotated scientific articles for biological curation. Articles can be searched by keywords, biological categories, and bibliographical information. Keywords and categories can be searched in the whole text of the documents or in each sentence separately. A score is assigned to each article in Textpresso database according to the match with the provided query, and the articles returned by searches are sorted by their score. Scores for sentence searches is calculated as the sum of the scores for sentences in each document.

The base endpoint of the API is *https://textpressocentral.org:18080/request_name* where *request_name* is the API to call.

CHAPTER 1

Search Documents in Textpresso

Search documents indexed by Textpresso through queries on fulltext or sentences.

These are the APIs to perform document searches:

`POST /v1/textpresso/api/search_documents`

Search for documents indexed by Textpresso. **Requires authentication**

Request JSON Object

- **`token`** (*`string`*) – a valid access token. See *How to obtain an access token* for further information on how to get one.
- **`query`** (*`object`*) – a query object (see *Query Object* for more details)
- **`include_fulltext`** (*`boolean`*) – whether to return the fulltext and abstract of the documents. *Default value* is **false**. Restricted to specific tokens due to copyright.
- **`include_all_sentences`** (*`boolean`*) – whether to return the text of all the sentences in the text. *Default value* is **false**. Restricted to specific tokens due to copyright.
- **`include_match_sentences`** (*`boolean`*) – whether to return the text of each matched sentence. Valid only for sentence searches. *Default value* is **false**
- **`since_num`** (*`int`*) – used for pagination. Skip the first results and return entries from the specified number. Note that the counter starts from 0 - i.e., the first document is number 0.
- **`count`** (*`int`*) – used for pagination. Return up to the specified number of results. *Maximum value* is **200**

Response Datatype Format

The returned data is a json array of objects, each of which contains the following fields:

Response JSON Object

- **`identifier`** (*`string`*) – the document identifier
- **`score`** (*`string`*) – the score of the document - an absolute number that indicates the degree to which the document matches the provided query

- **`title`** (*string*) – the title of the document
- **`author`** (*string*) – the author(s) of the document
- **`accession`** (*string*) – the accession of the document
- **`journal`** (*string*) – the journal of the document
- **`year`** (*string*) – publication year
- **`doc_type`** (*string*) – the type of document (e.g., research article, review)
- **`fulltext`** (*string*) – the fulltext of the document. Only if *include_fulltext* is set to **true** in the request.
- **`abstract`** (*string*) – the abstract of the document. Only if *include_fulltext* is set to **true** in the request.

Response JSON Array of Objects

- **`all_sentences`** (*string*) – the text of each sentence. Only if *include_all_sentences* is set to **true** in the request.
- **`matched_sentences`** (*string*) – the text of each matched sentence. Only if *include_match_sentences* is set to **true** in the request and the query type is set to **sentence**.

Example request:

```
POST /v1/textpresso/api/search_documents HTTP/1.1
Host: textpressocentral.org:18080
Accept: application/json

{
   "token": "123456789",
   "query": {
      "keywords": "DYN-1",
      "type": "document",
      "case_sensitive": false,
      "sort_by_year": false,
      "count": 2,
      "corpora": [
                    "C. elegans",
                    "C. elegans Supplementals"
                 ]
   }
}
```

Example response:

```
HTTP/1.1 200 OK
Vary: Accept
Content-Type: text/javascript

[
   {
      "doc_type": "Journal_article",
      "score": 0.0418161,
      "identifier": "I5m",
      "title": "Factors regulating the abundance and localization of␣
↪synaptobrevin in the plasma membrane.",
      "author": "Dittman JS ; Kaplan JM",
      "accession": " Other:doi:10.1073\\/pnas.0600784103 PMID:16844789 ␣
↪WBPaper00027755",
```

(continues on next page)

(continued from previous page)

```
      "journal": "Proc Natl Acad Sci U S A"
   },
   {
      "doc_type": "Journal_article",
      "score": 0.032331,
      "identifier": "B4r",
      "title": "A dynamin GTPase mutation causes a rapid and reversible␣
↪temperature-inducible locomotion defect in C. elegans.",
      "author": "Clark SG ; Shurland D-L ; Meyerowitz EM ; Bargmann CI ; Van der␣
↪Bliek AM",
      "accession": " Other:cgc2892 doi:10.1073\\/pnas.94.19.10438 PMID:9294229 ␣
↪WBPaper00002892",
      "journal": "Proc Natl Acad Sci U S A"
   }
]
```

Example request using Curl from the shell

```
curl -k -d "{\"token\":\"XXXXX\", \"query\": {\"keywords\": \"yeast AND two AND␣
↪hybrid\", \"year\": \"2017\", \"type\": \"sentence\", \"corpora\": [\"C.␣
↪elegans\"]}, \"include_sentences\": true}" https://textpressocentral.org:18080/
↪v1/textpresso/api/search_documents
```

`POST /v1/textpresso/api/get_documents_count`

Get the number of documents that match a search query. **Requires authentication**

Request JSON Object

- **`token`** (*`string`*) – a valid access token. See *How to obtain an access token* for further information on how to get one.
- **`query`** (*`object`*) – a query object (see *Query Object* for more details)

Response Datatype Format

Response JSON Object

- **`counter`** (*`int`*) – the number of documents matching the query

Example request:

```
POST /v1/textpresso/api/get_documents_count HTTP/1.1
Host: textpressocentral.org:18080
Accept: application/json

{
   "token": "123456789",
   "query": {
      "keywords": "DYN-1",
      "type": "document",
      "case_sensitive": false,
      "sort_by_year": false,
      "count": 2,
      "corpora": [
                    "C. elegans",
                    "C. elegans Supplementals"
                 ]
   }
}
```

Example response:

```
HTTP/1.1 200 OK
Vary: Accept
Content-Type: text/javascript

{
  "counter": 229
}
```

GET /v1/textpresso/api/available_corpora

Get the list of corpora available on the server

Response Data Format

A json array of strings

Example request:

```
GET /v1/textpresso/api/available_corpora HTTP/1.1
Host: textpressocentral.org:18080
```

Example response:

```
HTTP/1.1 200 OK
Vary: Accept
Content-Type: text/javascript

["C. elegans","C. elegans Supplementals","PMCOA C. elegans","PMCOA Animal"]
```

POST /v1/textpresso/api/get_category_matches_document_fulltext

Get the list of words in the fulltext of one or more documents that match a specified category. **Requires authentication**

Request JSON Object

- **token** (*string*) – a valid access token. See *How to obtain an access token* for further information on how to get one.
- **query** (*object*) – a query object used to search for the documents (see *Query Object* for more details)
- **category** (*string*) – a valid category in Textpresso format (e.g., "Gene (C. elegans) (tpgce:0000001)") - see Textpresso central category browser for the complete list of supported categories.

Response Datatype Format

The returned data is a json array of objects, each of which represents a document matched by the provided query, and contains the following fields:

Response JSON Object

- **identifier** (*string*) – the document identifier

Response JSON Array of Objects

- **matches** (*string*) – the list of words in the fulltext of the document that matched the specified category

Example request:

```
POST /v1/textpresso/api/get_category_matches_document_fulltext HTTP/1.1
Host: textpressocentral.org:18080
Accept: application/json

{
   "token": "123456789",
   "query": {
      "accession": "WBPaper00050052",
      "corpora": [
                    "C. elegans",
                    "C. elegans Supplementals"
                 ]
   },
   "category": "Gene (C. elegans) (tpgce:0000001)"
}
```

Example response:

```
HTTP/1.1 200 OK
Vary: Accept
Content-Type: text/javascript

[
   {
      "identifier":"C. elegans/WBPaper00050052/WBPaper00050052.tpcas",
      "matches": ["apl-1","cdc-42","ceh-36","daf-16","glp-1","hsf-1","ins-33",
↪"lin-14","lin-4","mec-4","pmp-3","rab-3","snb-1"]
   }
]
```

CHAPTER 2

Query Object

Specify a query to search documents in textpresso

Calls that require a query object must contain the following fields:

ANY query

Request JSON Object

- **keywords** (*string*) – *(optional)* the keywords to match in the text. Can contain logical operators AND and OR and grouping by round brackets
- **exclude_keywords** (*string*) – *(optional)* the keywords to exclude. Can contain logical operators AND and OR and grouping by round brackets
- **year** (*string*) – *(optional)* year of publication of the paper
- **author** (*string*) – *(optional)* the author(s) of the paper
- **accession** (*string*) – *(optional)* the accession of the paper
- **journal** (*string*) – *(optional)* the journal where the paper has been published
- **paper_type** (*string*) – *(optional)* the type of paper (e.g., research_article, review)
- **exact_match_author** (*bool*) – *(optional)* apply exact match on the author field
- **exact_match_journal** (*bool*) – *(optional)* apply exact match on the journal field
- **categories_and_ed** (*bool*) – *(optional)* use AND logical operator between the provided categories
- **type** (*string*) – the type of search to perform. Accepted values are: **document** to query the fulltext of documents and **sentence** to search in each sentence separately. *Default value* is **document**
- **case_sensitive** (*boolean*) – whether to perform a case sensitive search. *Default value* is **false**
- **sort_by_year** (*boolean*) – whether the results have to be sorted by publication date. *Default value* is **false**

Request JSON Array of Objects

- **categories** (*string*) – *(optional)* a set of categories to match in the text
- **corpora** (*string*) – *(optional)* restrict the search to the specified list of corpora

Example:

```
{
  "query":
  {
    "keywords": "DYN-1",
    "type": "document",
    "case_sensitive": false,
    "sort_by_year": false,
    "corpora": [
                  "C. elegans",
                  "C. elegans Supplementals"
               ]
  }
}
```

CHAPTER 3

How to obtain an access token

Textpresso API requires authentication for most of its endpoints - i.e., a valid *token* string must be supplied with the requests.

To obtain a token, contact valearna@caltech.edu.

CHAPTER 4

Indices and tables

- genindex
- search

HTTP Routing Table