termite-toolkit

Release 0.2.3

SciBite Data Science Team

Jun 20, 2019

CONTENTS

CHAPTER

ONE

#1 – TERMITE

TERMiteRequestBuilder- make requests to the TERMite API and process results.

class termite_toolkit.termite.**TermiteRequestBuilder**

Class for creating TERMite requests

execute(*display_request=False*)

Once all settings are done, POST the parameters to the TERMite RESTful API

Parameters **display_request** – if True request will be printed out before being submitted

Returns request response

set_basic_auth(*username=''*, *password=''*, *verification=True*)

Pass basic authentication credentials **ONLY change verification if you are calling a known source**

Parameters

- **username** – username to be used for basic authentication
- **password** – password to be used for basic authentication
- **verification** – if set to False requests will ignore verifying the SSL certificate, can also pass the path

to a certificate file

set_binary_content(*input_file_path*)

For annotating file content, send file path string and process file as a binary multiple files of the same type can be scanned at once if placed in a zip archive

Parameters **input_file_path** – file path to the file to be sent to TERMite

set_entities(*string*)

Limit the entities to be annotated

Parameters **string** – a comma separated string of entity types, e.g. 'DRUG,GENE'

set_fuzzy(*bool*)

Use fuzzy matching?

Parameters **bool** – set to True if fuzzy matching is to be enabled

set_input_format(*string*)

Set input format e.g. txt, medline.xml, node.xml, pdf, xlsx

Parameters **string** – string input format

set_max_docs(*integer*)

When tagging a zip file of multiple documents, limit how many to scan also applies where there are multiple document records in a single xml e.g. from a medline XML export

Parameters **integer** – number of documents to limit annotation too

set_no_empty(*bool*)
Reject all documents where there were no hits

Parameters **bool** – if True do not return any docs with no hits

set_options(*options_dict*)
For bulk setting multiple TERMite API options in a single call, send a dictionary object here

Parameters **options_dict** – a dictionary of options to be passed to TERMite

set_output_format(*string*)
Set output format e.g. tsv, json, doc.json

Parameters **string** – provide the output format to be used

set_reject_ambiguous(*bool*)
Automatically reject any hits flagged as ambiguous

Parameters **bool** – set True to reject any ambiguous hits

set_subsume(*bool*)
Take longest hit where an entity is a hit against more than one dictionary

Parameters **bool** – set subsume if True

set_text(*string*)
Use this for tagging raw text e.g. if looping through some file content

Parameters **string** – text to be sent to TERMite

set_url(*url*)
Set the URL of the TERMite instance e.g. for local instance http://localhost:9090/termite

Parameters **url** – the URL of the TERMite instance to be hit

termite_toolkit.termite.**all_entities**(*termite_response*)
Parses TERMite response and returns a list of VOCab modules with hits

Parameters **termite_response** – JSON or doc.JSONx TERMite response

Returns list

termite_toolkit.termite.**all_entities_df**(*termite_response*)
Parses JSON or doc.JSONx TERMite response into summary of hits dataframe

Parameters **termite_response** – JSON or doc.JSONx TERMite response

Returns pandas dataframe

termite_toolkit.termite.**annotate_files**(*url*, *input_file_path*, *options_dict*)
Wrapper function to execute a TERMite request for annotating individual files or a zip archive

Parameters

- **url** – url of TERMite instance
- **input_file_path** – path to file to be annotated
- **options_dict** – dictionary of options to be used during annotation

Returns result of request

termite_toolkit.termite.**annotate_text**(*url*, *text*, *options_dict*)
Wrapper function to execute a TERMite request for annotating strings of text

Parameters

- **url** – url of TERMite instance
- **text** – text to be annotated
- **options_dict** – dictionary of options to be used during annotation

Returns result of request

termite_toolkit.termite.**bool_to_string**(*bool*)
Convert a boolean to a string

Parameters **bool** – provide boolean to be converted

Returns string

termite_toolkit.termite.**docjsonx_payload_records**(*docjsonx_response_payload*, *reject_ambig=True*, *score_cutoff=0*, *remove_subsumed=True*)
Parses TERMite doc.JSONx payload into records, includes rules to filter out ambiguous and low-relevance hits

Parameters

- **docjsonx_response_payload** – doc.JSONx TERMite response.
- **reject_ambig** – boolean
- **score_cutoff** – a numerical value between 1-5
- **remove_subsumed** – boolean

Returns TERMite response in records format

termite_toolkit.termite.**entity_freq**(*termite_response*)
Parses TERMite JSON or doc.JSONx response and returns dataframe of entity type frequencies

Parameters **termite_response** – JSON or doc.JSONx TERMite response

Returns pandas dataframe

termite_toolkit.termite.**get_entity_hits_from_docjsonx**(*termite_response*, *filter_entity_types*)
Parses doc.JSONx TERMite response and returns a summary of the hits

Parameters

- **termite_response** – doc.JSONx TERMite response
- **filter_entity_types** – comma separated list

Returns dictionary of filtered hits

termite_toolkit.termite.**get_entity_hits_from_json**(*termite_json_response*, *filter_entity_types*, *reject_ambig=True*, *score_cutoff=0*)
Extract entity hits from TERMite JSON

Parameters

- **termite_json_response** – JSON returned from TERMite
- **filter_entity_types** – string of entity types separated by commas
- **reject_ambig** – boolean
- **score_cutoff** – a numeric value between 1-5

Returns dictionary of filtered hits

termite_toolkit.termite.**get_termite_dataframe**(*termiteResponse*, *cols_to_add=''*, *reject_ambig=True*, *score_cutoff=0*, *remove_subsumed=True*)

Parses TERMite JSON or doc.JSONx into a dataframe of hits, filtering out ambiguous and low-relevance hits By default returns docID, entityType, hitID, name, score, realSynList, totnosyns, nonambigsyns, frag_vector_array Additional hit information not included in the default output can be included by use of a comma separated list

Parameters

- **termiteResponse** – JSON or doc.JSONx response from TERMite
- **cols_to_add** – comma separated list of additional fields to include
- **reject_ambig** – boolean
- **score_cutoff** – a numerical value between 1-5
- **remove_subsumed** – boolean

Returns dataframe of TERMite hits

termite_toolkit.termite.**json_payload_records**(*response_payload*, *reject_ambig=True*, *score_cutoff=0*, *remove_subsumed=True*)

Parses TERMite json payload into records, includes rules to filter out ambiguous and low-relevance hits

Parameters

- **response_payload** – REP_PAYLOAD of JSON TERMite response
- **reject_ambig** – boolean
- **score_cutoff** – a numerical value between 1-5
- **remove_subsumed** – boolean

Returns TERMite response in records format

termite_toolkit.termite.**payload_records**(*termiteResponse*, *reject_ambig=True*, *score_cutoff=0*, *remove_subsumed=True*)

Parses TERMite JSON or doc.JSONx output into records format

Parameters

- **termiteResponse** – JSON or doc.JSONx TERMite response
- **reject_ambig** – boolean
- **score_cutoff** – a numerical value between 1-5
- **remove_subsumed** – boolean

Returns TERMite response in records format

termite_toolkit.termite.**process_payload**(*filtered_hits*, *response_payload*, *filter_entity_types*, *doc_id=''*, *reject_ambig=True*, *score_cutoff=0*, *remove_subsumed=True*)

Parses the termite json output to filter out only entity types of interest and their major metadata includes rules for rejecting ambiguous or low-relevance hits

Parameters

- **filtered_hits** – input
- **response_payload** – json response
- **filter_entity_types** – entity types to filter
- **doc_id** – doc id to filter

- **reject_ambig** – boolean reject ambiguous hits
- **score_cutoff** – int score cit-off
- **remove_subsumed** – boolean remove subsumed

Returns dictionary of filtered hits

termite_toolkit.termite.**top_hits_df**(*termite_response*, *selection=10*, *entity_subset=None*, *include_docs=False*)

Parses JSON or doc.JSONx TERMite response and returns a pandas dataframe of the most frequent hits. By default the top 10 most frequent hits are returned. The entity types to include can be set by a comma separated list For multidoc results the documents in which hits occur can be included

Parameters

- **termite_response** – JSON or doc.JSONx TERMite response
- **selection** – number of most frequent hits to return
- **entity_subset** – comma separated list
- **include_docs** – boolean

Returns pandas dataframe

CHAPTER

TWO

#2 – TEXPRESS

TExpressRequestBuilder- make requests to the TExpress API and process results.

class termite_toolkit.texpress.**TexpressRequestBuilder**

Class for creating TEXpress requests

execute(*display_request=False*)

Once all settings are done, POST the parameters to the TERMite RESTful API

Parameters **display_request** – if True request will be printed out before being submitted

Returns request response

set_allow_ambiguous(*bool*)

Allow matches containing ambiguous entity hits to be returned

Parameters **bool** – string boolean

set_alwaysadd(*bool*)

Always return an annotated sentence, even if no hit. Note, use the pattern !ANNOTATE to obtain this without any pattern search

Parameters **string** – string boolean

set_basic_auth(*username=''*, *password=''*, *verification=True*)

Pass basic authentication credentials **ONLY change verification if you are calling a known source**

Parameters

- **username** – username to be used for basic authentication
- **password** – password to be used for basic authentication
- **verification** – if set to False requests will ignore verifying the SSL certificate, can also pass the path to a certfile

set_binary_content(*input_file_path*)

For annotating file content, send file path string and process file as a binary multiple files of the same type can be scanned at once if placed in a zip archive

Parameters **input_file_path** – file path to the file to be sent to TERMite

set_bundle(*bundle_name*)

Provide a bundle to be used during TExpress search. Please ensure that this bundle is loaded on the server which you are calling

Parameters **bundle_name** – name of the bundle you wish to call

set_entities(*string*)

Limit the entities to be annotated

Parameters **string** – a comma separated string of entity types, e.g. 'DRUG,GENE'

set_fuzzy(*bool*)
Use fuzzy matching?

Parameters **bool** – set to True if fuzzy matching is to be enabled

set_input_format(*string*)
Set input format e.g. txt, medline.xml, node.xml, pdf, xlsx

Parameters **string** –

set_max_docs(*integer*)
When tagging a zip file of multiple documents, limit how many to scan also applies where there are multiple document records in a single xml e.g. from a medline XML export

Parameters **integer** – number of documents to limit annotation too

set_no_empty(*bool*)
Reject all documents where there were no hits

Parameters **bool** – if True do not return any docs with no hits

set_options(*options_dict*)
For bulk setting multiple TERMite API options in a single call, send a dictionary object here

Parameters **options_dict** – a dictionary of options to be passed to TERMite

set_output_format(*string*)
Set output format e.g. tsv, json, doc.json

Parameters **string** –

set_pattern(*pattern*)
Provide a pattern to be used during TExpress search.

Parameters **pattern** – pattern string

set_pivot(*bool*)
List TExpress hits by entity rather than document. Will result in redundant data and only works for some output formats

Parameters **string** – string boolean

set_reverse(*bool*)
Should we look for this reverse version of this pattern?

Parameters **bool** – boolean look for reverse

set_subsume(*bool*)
If another TExpress hit full overlaps this hit, hits to this pattern are removed

Parameters **bool** – set subsume if True

set_text(*string*)
Use this for tagging raw text e.g. if looping through some file content

Parameters **string** – text to be sent to TERMite

set_tx_group(*bool*)
Capture entities matching non-spacer groups into a *group* parameter

Parameters **string** – string boolean

set_url(*url*)
Set the URL of the TERMite instance e.g. for local instance http://localhost:9090/termite

Parameters **url** – the URL of the TERMite instance to be hit

termite_toolkit.texpress.**annotate_files**(*url*, *input_file_path*, *options_dict*)
Wrapper function to execute a TExpress request for annotating individual files or a zip archive

Parameters

- **url** – url of TERMite instance
- **input_file_path** – path to file to be annotated
- **options_dict** – dictionary of options to be used during annotation

termite_toolkit.texpress.**annotate_text**(*url*, *text*, *options_dict*)
Wrapper function to execute a TExpress request for annotating strings of text

Parameters

- **url** – url of TERMite instance
- **text** – text to be annotated
- **options_dict** – dictionary of options to be used during annotation

termite_toolkit.texpress.**bool_to_string**(*bool*)
Convert a boolean to a string

Parameters **bool** – provide boolean to be converted

termite_toolkit.texpress.**docjsonx_records**(*docjsonx_response*, *remove_subsumed=True*)
Parses doc.JSONx TExpress into records, includes filter to remove subsumed hits

Parameters

- **docjsonx_response** – TExpress doc.JSONx response
- **remove_subsumed** – boolean

Returns TExpress hits in records format

termite_toolkit.texpress.**get_entity_hits_from_json**(*termite_json_response*, *score_cutoff=0*)
Remove the entity hits from returned TExpress JSON and return a dictionary in the format (pattern_id : (orig_sentence, [entities]))

Parameters

- **termite_json_response** – JSON returned from TExpress
- **score_cutoff** – a numeric value between 1-5

termite_toolkit.texpress.**get_texpress_dataframe**(*texpress_response*, *cols_to_add=''*, *remove_subsumed=True*)
Get a dataframe from TEXpress response

Parameters

- **texpress_response** – texpress JSON response
- **cols_to_add** – additional column names to be included
- **remove_subsumed** – remove subsumed pattern hits

Returns

termite_toolkit.texpress.**json_resp_records**(*json_resp_texpress*, *remove_subsumed=True*)
parses JSON RESP_TEXPRESS into records, includes filter to remove subsumed hits.

Parameters

- **remove_subsumed** – remove the subsumed hits
- **json_resp_texpress** – RESP_TEXPRESS of TExpress JSON response

Returns TExpress hits in records format

termite_toolkit.texpress.**process_payload**(*texpress_hits*, *response_payload*, *doc_id=''*, *score_cutoff=0*, *remove_subsumed=True*)

Parses the termite json output to filter out only entity types of interest and their major metadata includes rules for rejecting ambiguous or low-relevance hits

Parameters

- **texpress_hits** – TExpress hits to be processed
- **response_payload** – total payload
- **doc_id** – document id
- **score_cutoff** – a numeric value between 1-5
- **remove_subsumed** – boolean

termite_toolkit.texpress.**texpress_records**(*texpress_response*, *remove_subsumed=True*)

Parses TExpress JSON or doc.JSONx response into records, with filtering to remove subsumed hits

Parameters

- **texpress_response** – TExpress JSON of doc.JSONx response
- **remove_subsumed** – boolean

Returns records of TExpress hits

CHAPTER

THREE

#3 – UTILITIES

Utility functions- including autocomplete

class termite_toolkit.utilities.**UtilitiesRequestBuilder**

Class for creating utility requests

call_autocomplete(*input*, *vocab*, *taxon=''*)

Complete a call to the auto complete API

Parameters

- **input** – input string
- **vocab** – vocabs to limit ac too
- **taxon** – taxon to limit ac too

get_entity(*entity_id*, *entity_type*)

Entity lookup function, given and entity type (e.g. GENE, INDICATION) and entity ID (e.g. CSF1, D010024) creates and runs GET call of the format: http://localhost:9090/termite/toolkit/tool.api?t=describe&id=INDICATION:D001249

Parameters

- **entity_id** – id of entity of interest
- **entity_type** – type of entity of interest

Returns request response

get_entity_details(*entity_id*, *entity_type*)

Returns a subset of metadata from the get_entity result: ID, name, mappings to external IDs

Parameters

- **entity_id** – id of entity of interest
- **entity_type** – type of entity of interest

Returns entity details

set_basic_auth(*username=''*, *password=''*, *verification=True*)

Pass basic authentication credentials. ** ONLY change verification if you are calling a known source **

Parameters

- **username** – username to be used for basic authentication
- **password** – password to be used for basic authentication

set_url(*url*)
Set the URL of the TERMite instance e.g. for local instance http://localhost:9090/termite/toolkit/autocomplete.api

Parameters **url** – the URL of the TERMite instance to be hit

PYTHON MODULE INDEX

t

INDEX

T

U