API Documentation

hdp-node11:14880/ws/documentation/index

TXT Werk - the Neofonie text mining API - analyzes text according to semantic criteria. Various methods of natural language technology are used. The API takes text as input and classifies them to topics, it extracts keywords which can be used as tags. If a text contains dates or date ranges, they will be annotated. Mentions of names (Named Entities) of places, persons, organizations, and concepts are recognized and annotated. If the entity can be found in the Wikidata ontology, an URI will be provided.

Authentication

The authentication is done by providing the API Key in the header "X-Api-Key".

Example Request

Details of the parameters and a detailed description of the return format can be found in the Request- or Response- documentation.

Request

```
curl "https://hdp-node11/rest/txt/analyzer" \
   -H "X-Api-Key: 1f163d44-de63-da89-bd2a-a310285ea80f" \
   --data-urlencode text='Angela Merkel wurde am 17. Juli 1954 in Hamburg als Angela
Dorothea Kasner geboren.' \
   -d services='categories,entities,tags,dates'
```

Response

- {
    - text: "Angela Merkel wurde am 17. Juli 1954 in Hamburg als Angela Dorothea Kasner geboren.",
    - language: "de",

- entities: [
    - {
        - confidence: 47.150089263916016,
        - end: 13,
        - label: "Angela Merkel",
        - start: 0,
        - surface: "Angela Merkel",
        - type: "PERSON",
        - uri: "https://www.wikidata.org/wiki/Q567"
    - },
    - {
        - confidence: 46.01070785522461,
        - end: 47,
        - label: "Hamburg",
        - start: 40,
        - surface: "Hamburg",
        - type: "PLACE",
        - uri: "https://www.wikidata.org/wiki/Q1055"
    - },
    - {
        - confidence: 75.0,
        - end: 74,
        - label: null,
        - start: 52,
        - surface: "Angela Dorothea Kasner",
        - type: "PERSON",
        - uri: null
    - }
-]

- tags: [
  - {
    - confidence: 0.9967904107197822,
    - term: "Angela Merkel"
  - },
  - {
    - confidence: 0.9927268430144784,
    - term: "Juli"
  - },
  - {
    - confidence: 0.9751561498425574,
    - term: "Hamburg"
  - },
  - {
    - confidence: 0.7406453816287002,
    - term: "Angela Dorothea Kasner"
  - }
-]
- dates: [
  - {
    - dateEnd:
      - {
        - bc: false,
        - day: 17,
        - month: 7,
        - year: 1954
      - }
    - dateStart:
      - {
        - bc: false,
        - day: 17,
        - month: 7,
        - year: 1954
      - }
    - end: 36,
    - start: 23,
    - surface: "17. Juli 1954"
  - }
-]

  - categories: [
    - {
      - confidence: 0.9840945695370302,
      - label: "politik"
    - },
    - {
      - confidence: 0.010815793425103136,
      - label: "wirtschaft"
    - },
    - {
      - confidence: 0.005075348628913112,
      - label: "kultur"
    - },
    - {
      - confidence: 1.097029997677795e-05,
      - label: "sport"
    - },
    - {
      - confidence: 1.8793566199359706e-06,
      - label: "reisen"
    - },
    - {
      - confidence: 8.053138213925 74e-07,
      - label: "wissenschaft"
    - },
    - {
      - confidence: 6.269585510453 14e-07,
      - label: "internet"
    - },
    - {
      - confidence: 6.4799843584039 16e-09,
      - label: "auto+technik"
    - }
  -]
- }

API Documentation - Request

hdp-node11:14880/ws/documentation/showRequest

The request can be send as GET or POST request to the URL
https://hdp-node11/rest/txt/analyzer
The document and the used services are passed as parameters.

Document

The document, which should be annotated, can be passed directly as text.

Alternatively, you can simply specify the URL of a website to be analyzed. In this case, the site of the main text content is crawled, determined and processed. Foreign elements, such as navigation or teaser text will be removed.

Services

The document can be analyzed with different techniques. Choose from the following services:

entities	Named Entities based on the Wikidata ontology.
tags	Keywords which appear in the text and describe and summarize the content.
categories	Assignment of text to categories of politics, business, cars & technology, internet, culture, travel, sports, human interest, science.
dates	Dates and periods.
entities-ml	Alternative named entity service based on a machine learning algorithm.
measures	Measurements that occur in the text.
authors	Authors of the article available as an HTML document.
fingerprints	Fingerprints for the text for near duplicate detection.
lexiconEntities	Named Entites based on a lexicon managed in TXT Werk.

Service Control

More parameters are available for individual services affecting the analysis or the result.

Example Request

Example of a POST request where the document is passed directly as text:

```
curl "https://hdp-node11/rest/txt/analyzer" \
   -H "X-Api-Key: 1f163d44-de63-da89-bd2a-a310285ea80f" \
   -d text='Angela Merkel wurde am 17. Juli 1954 in Hamburg als Angela Dorothea Kasner geboren.' \
   -d services='entities'
```

Example of a POST request where a HTML file is passed directly as input parameter:

```
curl "https://hdp-node11/rest/txt/analyzer" \
   -H "X-Api-Key: 1f163d44-de63-da89-bd2a-a310285ea80f" \
   -F htmlFile='@' \
   -F services='entities'
```

Overview of Parameters

Parameter	Area	Description
text	Document	Contains the annotated to document as text. If you have longer texts, please send the request as POST request and pass the text in the request body. mandatory: either text or htmlFile values: text
htmlFile	Document	Contains the annotated to document as html text. mandatory: either htmlFile or text values: html file text
title	Document	Title of the document. By additionally specifying a title, the result can be improved and will only be applied to the following services: tags. mandatory: no values: text
teaser	Document	Teaser of the document. By adding a teaser, the result can be improved and will only be applied to the following services: tags. mandatory: no values: text

Parameter	Area	Description
services	Services	List of requested services. mandatory: yes values: comma-separated list that contains at least one of the supported services: [entities, tags, categories, dates, entities-ml, measures, authors, fingerprints, lexiconEntities]
language	Service control	Language of the document. Language-dependent components can be specifically activated by setting this parameter. mandatory: no, will then be auto-detected values: 'en' or 'de'
ntags	Service control	Maximum number of keywords (tags) which are requested. Service: tags. mandatory: no, default: 10 values: non-negative integer
ncategories	Service control	Number of returned categories. Service: categories. mandatory: no values: non-negative integer
nentities	Service control	Number of returned entities. Service: entities. mandatory: no values: non-negative integer
nerMinConfidence	Service control	Threshold for the entity confidence. Service: entities. mandatory: no values: non-negative integer

Parameter	Area	Description
nerMinRelevance	Service control	Schwellwert für die Relevanz bei den Entitäten. Service: entities. mandatory: no values: non-negative integer
nerFormat	Service control	Response format for the entities. Service: entities. mandatory: no values: 'list' or 'aggregate' (aggregated list of entities, sorted by relevance)

API Documentation - Response

hdp-node11:14880/ws/documentation/showResponse

The response is always in json format. It contains the analyzed text and the language of the text, and for every requested service the response contains a block. The content of the response block is service-specific and contains the actual analysis result for this service. For clear documentation, the response block will be omitted here, but will be described later in detail for each service.

For an example of a complete response, see section Overview.

Response Format

- {
    - text: "Angela Merkel wurde am 17. Juli 1954 in Hamburg als Angela Dorothea Kasner geboren.",
    - timestamp: 1400247994051,
    - language: "de",
    - entities: [
    -]
    - lexiconEntities: [
    -]
    - tags: [
    -]
    - dates: [
    -]
    - categories: [
    -]
    - measures: [
    -]
- }

An empty result list will be returned if a service has successfully analyzed the text, but found no results. In case of an error of a single service, the returned HTTP status will be 200 and the response content will contain the results of all the services, except for the failed service block.

Description of each field:

text	The analyzed text. If you passed an URL, the extracted plain text (with boiler plate removal) will be displayed. If you passed an plain text within the parameter 'text', the unchanged text will be shown.
language	The language for the text, eg "de" , "en", or "ru" or others.

timestamp	The timestamp of the response (in milliseconds since January 1, 1970).

Response Format: Entities

- {
    - entities: [
        - {
            - confidence: 47.72833251953125,
            - relevance: 15.534404754638672,
            - surface: "Angela Merkel",
            - label: "Angela Merkel",
            - uri: "https://www.wikidata.org/wiki/Q567",
            - type: "PERSON",
            - start: 0,
            - end: 13
        - },
        - {
            - confidence: 39.60715866088867,
            - relevance: 14.97057819366455,
            - surface: "Hamburg",
            - label: "Hamburg",
            - uri: "https://www.wikidata.org/wiki/Q1055",
            - type: "PLACE",
            - start: 40,
            - end: 47
        - },
        - {
            - confidence: 100.0,
            - relevance: 17.836894989013672,
            - surface: "Angela Dorothea Kasner",
            - label: null,
            - uri: null,
            - type: "PERSON",
            - start: 52,
            - end: 74
        - }
    -]
- }

Description of each field:

label	The unique label of the entity.
surface	The surface form of the entity in the text.

type	Type of entity. Possible values are "PERSON", "PLACE", "ORGANISATION", "JOB TITLE", "WORK", "EVENT", "CONCEPT". This is determined heuristically and may vary in some cases from the expected value. Example: A city can act as an employer and can be therefore classified as an organization.
uri	The Wikidata URI of the named entity. Set to 'null' if there is no entity URI in the Wikidata knowledge base.
confidence	Confidence value about the discovered entity. A higher value represents a more secure detection. The upper value of the confidence is unlimited.
relevance	Relevance value for the discovered entity. A higher value represents a more important entity. The upper value of the relevance is unlimited.
start	The start position of the entity in the text.
end	The end position of the entity in the text.

Response Format: Top Entities

- {
    - topEntities: [
        - {
            - confidence: 717.3840942382812,
            - relevance: 40.1431999206543,
            - label: "Angela Merkel",
            - uri: "https://www.wikidata.org/wiki/Q567",
            - type: "PERSON",
            - matches: [
                - {
                    - surface: "Angela Merkel",
                    - start: 0,
                    - end: 13
                - },
                - {
                    - surface: "Merkel",
                    - start: 89,
                    - end: 95
                - },
                - {
                    - surface: "Bundeskanzlerin",
                    - start: 104,
                    - end: 119
                - }
            -]
        - },

        - {
            - confidence: 100.0,
            - relevance: 17.836894989013672,
            - label: "Angela Dorothea Kasner",
            - uri: null,
            - type: "PERSON",
            - matches: [
                - {
                    - surface: "Angela Dorothea Kasner",
                    - start: 52,
                    - end: 74
                - }
            -]
        - },
        - {
            - confidence: 39.51301193237305,
            - relevance: 14.95887279510498,
            - label: "Hamburg",
            - uri: "https://www.wikidata.org/wiki/Q1055",
            - type: "PLACE",
            - matches: [
                - {
                    - surface: "Hamburg",
                    - start: 40,
                    - end: 47
                - }
            -]
        - }
    -]
- }

Description of each field:

label	The unique label of the entity.
type	Type of entity. Possible values are "PERSON", "PLACE", "ORGANISATION", "JOB TITLE", "WORK", "EVENT", "CONCEPT". This is determined heuristically and may vary in some cases from the expected value. Example: A city can act as an employer and can be therefore classified as an organization.
uri	The Wikidata URI of the named entity. Set to 'null' if there is no entity URI in the Wikidata knowledge base.

confidence	Confidence value about the discovered entity. A higher value represents a more secure detection. The upper value of the confidence is unlimited.
relevance	Relevance value for the discovered entity. A higher value represents a more important entity. The upper value of the relevance is unlimited.
matches	The matches of the entity in the text.
matches.surface	The surface form of the entity in the text.
matches.start	The start position of the entity in the text.
matches.end	Die Endposition der Fundstelle im Text.

Response Format: Entities ML

The machine learning (ML) Service uses a machine learning model to detect the correct entities. Generally speaking, it is more accurate for lesser known entities as they may occur in blog posts. For more well-known entities - as they often occur in news articles - the regular entity service will generally perform better. If you are unsure which service to pick, use the regular Entity service. The response format of the Entity ML service is identical to that of the regular Entity service.

Response Format: Lexicon Entities

These Named Entities are based on a lexicon managed in TXT Werk. Different to the Wikidata entities, they are determined without any disambiguation. The response format is the same as for 'entities', except the different response block name 'lexiconEntities'.

Description of each field:

label	See entities.
surface	See entities.
type	Type of entity. Possible values are managed in the lexicon and depend on its state.
uri	A URI associated with this named entity, typically an identifier in an external system.
confidence	See entities. Although in this case the return value is always 1 - it means, it's found.
start	See entities.
end	See entities.

Response Format: Tags

- {
  - tags: [
    - {
      - confidence: 0.9967904107197822,
      - term: "Angela Merkel"
    - },
    - {
      - confidence: 0.9927268430144784,
      - term: "Juli"
    - },
    - {
      - confidence: 0.9751561498425574,
      - term: "Hamburg"
    - },
    - {
      - confidence: 0.7406453816287002,
      - term: "Angela Dorothea Kasner"
    - }
  -]
- }

Description of each field:

term	The found Keyword.
confidence	The confidence value of the phrase. The value is always between 0 to 1.

Response Format: Dates

- {
    - dates: [
        - {
            - dateEnd:
                - {
                    - bc: false,
                    - day: 17,
                    - month: 7,
                    - year: 1954
                - }
            - dateStart:
                - {
                    - bc: false,
                    - day: 17,
                    - month: 7,
                    - year: 1954
                - }
            - end: 36,
            - start: 23,
            - surface: "17. Juli 1954"
        - }
    -]
- }

Description of each field:

surface	The surface form of the date in the text.
start	The start position of the date in the text.
end	The final position of the date in the text.
dateStart	The start date. A date is always represented as time periods, e.g. start and end date may have the same value.
dateEnd	The end date.
day	The day of the start or end date. Possible values are 1-31.
month	The month of the start or end date. Possible values are 1-12.
year	The year of the start or end date.
bc	Describes whether the date refers to the time before Christ. Possible values are true and false.

Response Format: Categories

- {
    - categories: [
        - {
            - confidence: 0.9840945695370302,
            - label: "politik"
        - },
        - {
            - confidence: 0.010815793425103136,
            - label: "wirtschaft"
        - },
        - {
            - confidence: 0.005075348628913112,
            - label: "kultur"
        - },
        - {
            - confidence: 1.09702999767795e-05,
            - label: "sport"
        - },
        - {
            - confidence: 1.8793566199359706e-06,
            - label: "reisen"
        - },
        - {
            - confidence: 8.05313821392574e-07,
            - label: "wissenschaft"
        - },
        - {
            - confidence: 6.26958551045314e-07,
            - label: "internet"
        - },
        - {
            - confidence: 6.479984358403916e-09,
            - label: "auto+technik"
        - }
    -]
- }

Description of each field:

label	The name of the category. Possible values are "politik", "wirtschaft", "auto+technik", "internet", "kultur", "reisen", "sport", "vermischtes", "wissenschaft" (e.g. "politics", "economics", "auto + technology", "internet", "culture", "travel", "sport", "mixed", "economy", "science")
confidence	The confidence value for the category is always between 0 to 1.

Response Format: Measures

- {
    - measures: [
        - {
            - start: 8,
            - end: 15,
            - text: "2 Meter",
            - valueString: "2",
            - unitString: "Meter",
            - type: "LENGTH"
        - }
    -]
- }

Description of each field:

start	The start position of the measurement in the text.
end	The end position of the measurement in the text.
text	The measurement string, exactly as it occurs in the text.
valueString	The value as a string, exactly as it occurs in the text.
unitString	The unit as a string, exactly as it occurs in the text.
type	The type of the measurement. Possible values are "LENGTH", "AREA", "MASS", "TEMPERATURE", "VOLTAGE", "AMPERAGE", "RESISTANCE", "CHARGE", "CAPACITY", "CONDUCTANCE", "INDUCTANCE", "MAGNETIC_STRENGTH", "POWER", "ENERGY", "FORCE", "PRESSURE", "FREQUENCY", "VOLUME", "LUMINOSITY", "ILLUMINANCE", "SPIN", "SUBSTANCE", "RADIOACTIVITY", "CURRENCY", "TIME", "UNKNOWN"

API Documentation - Failure

hdp-node11:14880/ws/documentation/showErrors

Response Format

In case of failure, error details will be displayed in json format: the HTTP status, a TXT Werk-internal error code, a short error message and -if available- more error details. Please find here an example for exceeding the daily limit of API calls:

- {
    - status: 403,
    - code: "403-002",
    - reason: "exceeded request quota",
    - details: "number of allowed requests per day (1000) reached"
- }

In case of a validation error, the rejected value and the validation error message will be displayed :

- {
    - status: 422,
    - code: "422-001",
    - reason: "validation failed",
    - fieldErrors: [
        - {
            - field: "ntags",
            - rejectedValue: -2,
            - details: "Must have a nonnegative value."
        - },
        - {
            - field: "htmlURL",
            - rejectedValue: "neofonie.de",
            - details: "Must be a valid HTTP URL."
        - }
    -]
- }

List of error codes

Code	HTTP Status	Error message	Description
400-001	400	request binding error	The request was not identified as a valid request.

Code	HTTP Status	Error message	Description
400-002	400	missing document source parameter	The request must include a parameter either 'text' or 'htmlFile'.
400-003	400	duplicate document source parameter	The request must contain either 'text' or 'htmlFile' parameter, not both.
400-004	400	document source file unknown	The document has been specified via the 'htmlFile' parameter, but is not reachable.
400-005	400	missing service parameter	The services must be passed as a comma-separated list in the parameter 'services'. The allowed values are listed in the API documentation.
400-006	400	illegal service parameter value	In your 'services' parameter list is an unsupported service. Please find the allowed values in the API documentation.
400-010	400	uri already exists	Given uri already exists.
400-011	400	uri does not exist	Given uri does not exists.
400-015	400	missing service parameter	Necessary request parameter missing or wrong.
400-020	400	Field to patch an entry is missing	No field for a patch was found..
400-021	400	Unknown field to patch entry	Given field for patch was wrong.
401-001	401	missing api key header	The request must contain a valid API Key in the header "X-Api-Key" and match the requesting user.
401-002	401	unknown api key	An unkown API Key was passed in the header "X-Api-Key".

Code	HTTP Status	Error message	Description
401-003	401	missing request signature	For the given user, signed requests are mandatory: Please sign the request and pass the signature in the header "X-Signature".
401-004	401	invalid request signature	The header "X-Signature" in the request signature does not match the request and the API Secret of the requesting user.
401-005	401	missing admin role	documentation.error.description.MISSING_ADMIN_ROLE
403-001	403	locked api key	You are using the "X-Api-Key" header for a locked API Key. One cause could be an expired plan or a manual blocking.
403-002	403	exceeded request quota	The number of requests per day according to your chosen plan was exceeded
404-404-000	404	Page was not found on server.	Requested page was not found on server.
422-001	422	validation failed	At least one of the passed parameters is not valid.
500-001	500	unknown server error	An text mining API error has occurred, the request could not be answered.
500-002	500	watt server error	An error has occured in the called nerd services, the request could not be answered.
500-003	500	lexicon server error	An error has occured in the called TXT lexicon services, the request could not be answered.

API Documentation - Experimental Services

hdp-node11:14880/ws/documentation/showExperimental

You have access to other services, which are still under development and therefore not freely accessible to the public. Please note that the request and response format may not be stable.

Services

The services are passed in the services parameter passed as a comma-separated list. The choices are:

quotes	Quotes that are included in the text.
subjectivity	Measure of the subjectivity of the text.

Response Format: Quotes

- {
    - text: "\"Angela Merkel laufen die Kurfürsten in Scharen davon.\", sagte Jürgen Trittin.",
    - language: "de",
    - quotes: [
        - {
            - text: "\"Angela Merkel laufen die Kurfürsten in Scharen davon.\"",
            - source: null,
            - start: 0,
            - end: 55
        - }
    -]
- }

Description of each field :

text	The text of the quote.
source	PLANNED: The Free Base URI of the author, if the author of the quote is recognized from the text and is known as a person in Wikidata.
start	The start position of the quotation in the text.
end	The final position of the quotation in the text.

Response Format: Subjectivity

- {
    - text: "Angela Merkel wurde am 17. Juli 1954 in Hamburg als Angela Dorothea Kasner geboren.",
    - language: "de",
    - subjectivity: 0
- }

Description of each field :

subjectivity A value between 0 and 1 which describes the subjectivity of the text. A higher value indicates a more subjective text.