Name	**Data type**	**Description**
retrievedDate	Date	retrieved date of the news article, using SimpleDateFormat("yyyy-MM-dd'T'HH-mm-ss'Z'").
articleID	int	article id of the news article, given by the IJS newsfeed.
mention_NP	String	noun phrase (NP)
startOffset_NP	int	start offset of the NP in the news article
endOffset_NP	int	end offset of the NP in the news article
PosTag	String	Part-of-speech (POS) tag of the NP; often NN, NNP, …
allCaps	boolean	true if term consists only of capitalized chars.
containsDigitAndAlpha	boolean	true if an item contains at least one digit as well as at least one alpha character
containsNonAlpha	boolean	true if an item contains at least one non alpha character;
endsWithPeriod	boolean	true if an item contains a period "." at the end
firstCapital	boolean	true if the first character of an item is a capital letter;
firstLetterCapitalized	boolean	true if the first letter is a capital letter, i.e. [^a-z]*[A-Z]+.*
containsAZChar	boolean	true if the NP contains at least one "real letter", i.e. [A-Za-z]
hasInternalApostrophe	boolean	true if an item contains an apostrophe ("'")
hasInternalPeriod	boolean	true if an item contains a period "." internally;
internalCaps	boolean	true if an item contains at least one capitalized character at any position except the first character;
isHyphenated	boolean	true if an item contains a hyphen "-"
suffix	int	suffix of an item. A suffix is defined as four characters at the end of an item. After that the hash code is computed for the suffix. Finally, the computed hash code is returned for each suffix;
summarizedPattern	int	e.g. the item “iPhone 7” is mapped to the pattern xXxzy because all small letters are mapped to x, all capital letters to X, all digits to y, all spaces to z and all other characters to w. After that the hash code is computed for the pattern. Finally the computed hash code is returned for each summarized pattern;
triggerWordForLOC	boolean	true if an item contains a word that occurs in a list of trigger words for locations. This list is based on observations made by looking at named entities of type LOCATION in OntoNotes 4.0. (E.g. city, village, town etc.);
triggerWordForORG	boolean	true if an item contains a word that occurs in a list of trigger words for organizations. This list is based on observations made by looking at named entities of type ORGANIZATION in OntoNotes 4.0. (E.g. Inc., Co., Ltd. etc.);
triggerWordsForPERSON	boolean	true if an item contains a word that occurs in a list of trigger words for persons. This list is based on observations made by looking at named entities of type PERSON.
npLength	int	number of characters of NP
country	String	country where published
feedTitle	String	title of the RSS feed (correlated with feedURI probably)
feedURI	String	URI of the RSS feed
hostName	String	host_name (e.g., "www.n24.de"; correlates with source_name if one of them is not null).
sourceName	String	name of the source (e.g., "Boston Business Journal"; often empty)
sourceTags	String (word vector, separated by ",")	e.g., "geo_source:whois,usa_state:Kansas,usa_census_region:Midwest,usa_sub_region:W N Cen"
tags	String (word vector, separated by ",")	tags attached to the news article (e.g., "ConsumerAffairs,TAXUD,Consumers", quite often empty)
title	String	title of the news article
URI	String	URL of the news article
mention_Annotation	String	mention of annotation, if annotated
startOffset_Annotation	int	start offset of the mention of the annotation, if annotated
endOffset_Annotation	int	end offset of the mention of the annotation, if annotated
entityURI	String	Wikipedia URL of annotation, if annotated
entityWeight	float	weight of annotation, if annotated
noveltyClass	int	novelty class (1,2,3,4) if annotated (but due to classification, no class 1 and 3 included)
minutesToCreatedDate	int	minutes between news article retrieving date and date where WP entity was inserted; only if novel entity
redLink	boolean	true if the NP occurred as red link in a WP dump before (in March 2015)
pageView24hExist	boolean	returns true if there is at least one pageview value for this NP in the last 24h
pageView24hSum	int	sum of pageview values within the last 24h
pageViewDays7d	int (between 0 and 7)	given the values of the last 7 days, return number of days where pageview value is > 0 (i.e., a value is existing in the db)
pageViewDays14d	int (between 0 and 14)	given the values of the last 14 days, return number of days where pageview value is > 0 (i.e., a value is existing in the db).
pageViewDays7dMin100	int (between 0 and 7)	given the values of the last 7 days, return number of days where pageview value is >= 100.
pageViewDays14dMin100	int (between 0 and 14)	given the values of the last 14 days, return number of days where pageview value is >= 100.
pageViewSum7d	long	the sum of all pageview values of the last 7 days
pageViewSlope24hSlope	double	slope after applying simple linear regression on the pageview values of the last 24 h
pageViewSlope24hIntercept	double	intercept after applying simple linear regression on pageview values of the last 24 h
pageViewSlope24hRSquare	double	R^2 value after applying simple linear regression on pageview values of the last 24 h
pageViewSlope24hSlopeStdErr	double	slope standard error after applying linear regression on pageview values of the last 24 h
pageViewSlope24hInterceptStdErr	double	intercept standard error after applying simple linear regression on pageview values of the last 24 h

pageViewSlope14dSlope	double	slope after applying simple linear regression on pageview values of the last 14d
pageViewSlope14dIntercept	double	intercept after applying simple linear regression on pageview values of the last 14d
pageViewSlope14dRSquare	double	R^2 value after applying simple linear regression on pageview values of the last 14d
pageViewSlope14dSlopeStdErr	double	slope standard error after applying simple linear regression on pageview values of the last 14d
pageViewSlope14dInterceptStdErr	double	intercept standard error after applying simple linear regression on pageview values of the last 14d
npAsTitleInDEWP	boolean	true if the NP occurred as WP article in the German Wikipedia
npAsTitleInFRWP	boolean	true if the NP occurred as WP article in the French Wikipedia
npAsTitleInESWP	boolean	true if the NP occurred as WP article in the Spanish Wikipedia
npOccurrenceNo1h	int	how often the NP ocurred in the news in the last 1h (exactly); all NPs per article considered
npOccurrenceNo24h	int	how often the NP ocurred in the news in the last 24h (exactly), all NPs per article considered
npOccurrenceSlope24hSlope	double	slope after applying simple linear regression on NP frequency of the last 24h,all NPs per article considered
npOccurrenceSlope24hIntercept	double	intercept after applying simple linear regression on NP frequency of the last 24h,all NPs per article considered
npOccurrenceSlope24hRSquare	double	R^2 value after applying simple linear regression on NP frequency of the last 24h, all NPs per article considered
npOccurrenceSlope24hSlopeStdErr	double	slope standard error after applying simple linear regression on NP frequency of the last 24h, all NPs per article considered
npOccurrenceSlope24hInterceptStdErr	double	intercept standard error after applying simple linear regression on NP frequency of the last 24h, all NPs per article considered
namedEntity	String	if it is tagged as named entity, then the named entity type is written; otherwise empty.
npIsOnlyNE	boolean	true if whole NP is only a named entity (in order to detect noise)
target_label	boolean	target value for classification.
filteredTitleTokens	String	set of words in title of news article; made lowercase, delete non-printable chars.
annotationsInParagraph	String[]	(not-unique) list of annotations in paragraph of NP; separated by ; (used mainly to get entity types). Caution: xlisa buggy.
classesInParagraph	String	set of classes given by rdf:type of the annotations in the paragraph of the NP
containsSignalWord	boolean	true if article contains at least one of the signal words (e.g., "new", "emerging", …) collected manually.
npCountPerArticle	int	Occurrence number of current NP in article.
tokenContextFiltered5	String[]	set of words (tokens) before and after the NP (currently plus minus 5 words at most); stopwords are removed, made lowercase, deleted non-printable chars.
tokenContextFiltered10	String[]	set of words (tokens) before and after the NP (currently plus minus 10 words at most); stopwords are removed, made lowercase, deleted non-printable chars.
googleNgramFrequency	long	occurrence number of current NP in the Google ngram index
googleNgramSlope	float	getSlope() of slope of ngram frequency of current NP in the Google ngram index
googleNgramR2	float	R^2 value reg. ngram of current NP in the Google ngram index
googleNgramUsageSinceYear	int	year since the NP is used according to Google ngram index
googleNgramsPercentProperCaps	float [0,1]	the percentage of case-insensitive matches for a NP where all words began with a capital letter
googleNgramsMatchCountUntil1899	long	occurrence number of current NP in the Google ngram index in books from 1 (actually 1500) until 1899.
paragrapLengthWithNP	int	length of the paragraph which contains the NP
dayOfWeek	int	weakday, encoded in int
lastFullHour	int	hour of the retrieved date of the article
twitter24hSum	int	number of tweets in the last 24h containing NP as exact string
twitterSlope24hSlope	double	slope after applying simple linear regression on the number of tweets in the last 24h containing the NP as exact string, values taken by hour
twitterSlope24hIntercept	double	intercept after applying simple linear regression on the number of tweets in the last 24h containing the NP as exact string, values taken by hour
twitterSlope24hRSquare	double	R^2 value after applying simple linear regression on the number of tweets in the last 24h containing the NP as exact string, values taken by hour
twitterSlope24hSlopeStdErr	double	slope standard error after applying simple linear regression on the number of tweets in the last 24h containing the NP as exact string, values taken by hour
twitterSlope24hInterceptStdErr	double	intercept standard error after applying simple linear regression on the number of tweets in the last 24h containing the NP as exact string, values taken by hour
npDiffArtsOccurrenceNo1h	int	how often the NP ocurred in the news in the last 1h (exactly); NP only once per article considered
npDiffArtsOccurrenceNo24h	int	how often the NP ocurred in the news in the last 24h (exactly), NP only once per article considered
npDiffArtsOccurrenceSlope24hSlope	double	slope after applying simple linear regression on the of NP frequency values of the last 24h, NP only once per article considered
npOccurrenceSlope24hIntercept	double	intercept after applying simple linear regression on the NP frequency values of the last 24h, NP only once per article considered
npDiffArtsOccurrenceSlope24hRSquare	double	R^2 value after applying simple linear regression on the NP frequency values of the last 24h, NP only once per article considered
npDiffArtsOccurrenceSlope24hSlopeStdErr	double	slope standard error after applying simple linear regression on the NP frequency values of the last 24h, NP only once per article considered
npDiffArtsOccurrenceSlope24hInterceptStdErr	double	intercept standard error after applying simple linear regression on NP frequency values of the last 24h, all NPs per article considered
npPositionInArticle	double	position (beginning of startOffSet) of NP in current article; normalized, in [0,1].