

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)
☒ Repository ☐ Web [View ALL Data Sets](#)

News Aggregator Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: References to news pages collected from an web aggregator in the period from 10-March-2014 to 10-August-2014. The resources are grouped into clusters that represent pages discussing the same story.

Data Set Characteristics:	Multivariate	Number of Instances:	422937	Area:	N/A
Attribute Characteristics:	N/A	Number of Attributes:	5	Date Donated	2016-02-28
Associated Tasks:	Classification, Clustering	Missing Values?	N/A	Number of Web Hits:	56049

Source:

Provided by Artificial Intelligence Lab @ Faculty of Engineering, Roma Tre University - Italy

Contact: Fabio Gasparetti, Faculty of Engineering, Roma Tre University - Italy (gaspare'@'dia.uniroma3.it)

Data Set Information:

News are grouped into clusters that represent pages discussing the same news story.

The dataset includes also references to web pages that, at the access time, pointed (has a link to) one of the news page in the collection.

422937 news pages and divided up into:

152746 news of business category

108465 news of science and technology category

115920 news of business category

45615 news of health category

2076 clusters of similar news for entertainment category
1789 clusters of similar news for science and technology category
2019 clusters of similar news for business category
1347 clusters of similar news for health category

References to web pages containing a link to one news included in the collection are also included. They are represented as pairs of urls corresponding to 2-page browsing sessions. The collection includes 15516 2-page browsing sessions covering 946 distinct clusters divided up into:

6091 2-page sessions for business category
9425 2-page sessions for entertainment category

Attribute Information:

FILENAME #1: newsCorpora.csv (102.297.000 bytes)
DESCRIPTION: News pages
FORMAT: ID TITLE URL PUBLISHER CATEGORY STORY HOSTNAME TIMESTAMP

where:

ID Numeric ID
TITLE News title
URL Url
PUBLISHER Publisher name
CATEGORY News category (b = business, t = science and technology, e = entertainment, m = health)
STORY Alphanumeric ID of the cluster that includes news about the same story
HOSTNAME Url hostname
TIMESTAMP Approximate time the news was published, as the number of milliseconds since the epoch 00:00:00 GMT, January 1, 1970

FILENAME #2: 2pageSessions.csv (3.049.986 bytes)
DESCRIPTION: 2-page sessions
FORMAT: STORY HOSTNAME CATEGORY URL

where:

STORY Alphanumeric ID of the cluster that includes news about the same story
HOSTNAME Url hostname
CATEGORY News category (b = business, t = science and technology, e = entertainment, m = health)
URL Two space-delimited urls representing a browsing session

Relevant Papers:

Fabio Gasparetti. 2017. Modeling user interests from web browsing activities. Data Min. Knowl. Discov. 31, 2 (March 2017), 502-547. DOI: [\[Web Link\]](#)

Citation Request:

Please refer to the Machine Learning Repository's [citation policy](#).



In Collaboration With:



[About](#) || [Citation Policy](#) || [Donation Policy](#) || [Contact](#) || [CML](#)