

RCV1 dataset

Reuters Corpus Volume I (RCV1) is an archive of over 800,000 manually categorized newswire stories made available by Reuters, Ltd. for research purposes. The dataset is extensively described in [1].

Data Set Characteristics:

Classes	103
Samples total	804414
Dimensionality	47236
Features	real, between 0 and 1

`:func:`sklearn.datasets.fetch_rcv1`` will load the following version: RCV1-v2, vectors, full sets, topics multilabels:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\sklearn\datasets\descr\ (scikit-learn-main) (sklearn) (datasets) (descr) rcv1.rst, line 19); [backlink](#)

Unknown interpreted text role "func".

```
>>> from sklearn.datasets import fetch_rcv1
>>> rcv1 = fetch_rcv1()
```

It returns a dictionary-like object, with the following attributes:

data: The feature matrix is a scipy CSR sparse matrix, with 804414 samples and 47236 features. Non-zero values contains cosine-normalized, log TF-IDF vectors. A nearly chronological split is proposed in [1]: The first 23149 samples are the training set. The last 781265 samples are the testing set. This follows the official LYRL2004 chronological split. The array has 0.16% of non zero values:

```
>>> rcv1.data.shape
(804414, 47236)
```

target: The target values are stored in a scipy CSR sparse matrix, with 804414 samples and 103 categories. Each sample has a value of 1 in its categories, and 0 in others. The array has 3.15% of non zero values:

```
>>> rcv1.target.shape
(804414, 103)
```

sample_id: Each sample can be identified by its ID, ranging (with gaps) from 2286 to 810596:

```
>>> rcv1.sample_id[:3]
array([2286, 2287, 2288], dtype=uint32)
```

target_names: The target values are the topics of each sample. Each sample belongs to at least one topic, and to up to 17 topics. There are 103 topics, each represented by a string. Their corpus frequencies span five orders of magnitude, from 5 occurrences for 'GMIL', to 381327 for 'CCAT':

```
>>> rcv1.target_names[:3].tolist() # doctest: +SKIP
['E11', 'ECAT', 'M11']
```

The dataset will be downloaded from the [rcv1 homepage](#) if necessary. The compressed size is about 656 MB.

References

- [1] ([1,2](#)) Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research, 5, 361-397.