<https://ir4u2workshop.wixsite.com/ir4u2>

Website to advocate for research about Information Retrieval for unconventional users (understudied or disadvantaged groups)

<https://www.ecir2024.org/>

Home website for the European Conference on Information Retrieval

<https://commoncrawl.org/>

A place you can download web crawl data (raw webpage data or other data from all the public websites you could ever want)

These two links lead to the same subset of the crawl data that specifically has to do with news:

<https://paperswithcode.com/dataset/cc-news>

<https://commoncrawl.org/blog/news-dataset-available>

**In theory, this should contain news articles for the time covered by the MIND dataset**, but I don’t know how well I’ll be able to access, organize, and match them up.

\*\*Papers with code is a website that has a ton of datasets and related papers on it. Sometimes you have to follow links to their github or read through the paper to find out where to get the dataset:

<https://paperswithcode.com/dataset/inshorts-news>

- Github: https://github.com/sandeep16064/ats/tree/main

- Presents a dataset with just the titles and a short summary

- Also uses the following dataset: <https://paperswithcode.com/dataset/duc-2004>

* **500 news articles**, each **paired with four human written summaries**. News articles from 1998 – 2000.

<https://paperswithcode.com/dataset/cnn-daily-mail-1>

A dataset with around 310,000 news articles (**contains article content** but not too much else, dates unknown, no click data)

* Go to <https://cs.nyu.edu/~kcho/DMQA/> to download the dataset
* Unzip file:
  + tar -xvzf file\_path.tgz
* <https://github.com/abisee/cnn-dailymail> for instructions on how to use the data

<https://paperswithcode.com/dataset/otto-recommender-systems-dataset>

Data from OTTO, a German online shopping website. It does not include news articles or anything else related to news, unfortunately.

* <https://github.com/otto-de/recsys-dataset>

<https://paperswithcode.com/dataset/ag-news>

Dataset about news articles**, does not contain article content**. Contains source, url, title, image, category, description, rank, pubdate, video.

* <http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html>
* Download AG news Article’s corpus
* Unzip file:
  + bzip2 -d filename.bz2

<https://arxiv.org/abs/2209.11429>

They present a news dataset that includes the link to the news articles if we wanted to scrape them. All articles are from huffpost.

Dataset can be found here:

* <https://www.kaggle.com/datasets/rmisra/news-category-dataset/data>

<https://arxiv.org/abs/1906.01749>

A lot of news article data including article content can be found at:

* <https://github.com/Alex-Fabbri/Multi-News/tree/master/data>