This dataset contains headlines, URLs, and categories for 422,937 news stories collected by a web aggregator between March 10th, 2014 and August 10th, 2014.

News categories included in this dataset include business; science and technology; entertainment; and health. Different news articles that refer to the same news item (e.g., several articles about recently released employment statistics) are also categorized together.

## Content

The columns included in this dataset are:

* **ID** : the numeric ID of the article
* **TITLE** : the headline of the article
* **URL** : the URL of the article
* **PUBLISHER** : the publisher of the article
* **CATEGORY** : the category of the news item; one of: -- b : business -- t : science and technology -- e : entertainment -- m : health
* **STORY** : alphanumeric ID of the news story that the article discusses
* **HOSTNAME** : hostname where the article was posted
* **TIMESTAMP** : approximate timestamp of the article's publication, given in Unix time (seconds since midnight on Jan 1, 1970)

## Acknowledgments

This dataset comes from the [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml). Any publications that use this data should cite the repository as follows:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

This specific dataset can be found in the UCI ML Repository at [this URL](http://archive.ics.uci.edu/ml/datasets/News+Aggregator)

## Inspiration

What kinds of questions can we explore using this dataset? Here are a few possibilities:

* can we predict the category (business, entertainment, etc.) of a news article given only its headline?
* can we predict the specific story that a news article refers to, given only its headline?