Wayne Williams

Professor Wiejie Pang

Data Science Foundation

November 23, 2025

Datasets

My project uses two main sources of news articles: one for Chinese and one for English. Both datasets are large, well-organized, and widely used in NLP research.

Chinese Datasets

1. Yet Another Chinese News Dataset (Kaggle)
   Link: https://www.kaggle.com/datasets/ceshine/yet-another-chinese-news-dataset
   This dataset includes thousands of Chinese news articles across different categories like sports, finance, entertainment, politics, and more. It's clean, well structured, and perfect for topic modeling.

2. THUCNews (GitHub)
   Link: https://github.com/gaussic/text-classification-cnn-rnn/tree/master/thucnews
   A large and popular Chinese news dataset with millions of articles. It is commonly used for text classification and topic modeling tasks. It gives strong coverage of modern Chinese news language.

English Dataset

1. Open Humanities English News Dataset
   Link: https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.62
   This dataset contains English news articles organized across topics like technology, culture, business, and politics. It's formatted clearly and works well for NLP preprocessing and topic extraction.

Why These Datasets Work

- They are publicly available and designed for research.
- They contain thousands of articles, which makes topic modeling effective.
- They cover multiple categories, giving a balanced view of what each language emphasizes.
- They can be cleaned, tokenized, and transformed into TF-IDF vectors without issues.

All three datasets will be cleaned, segmented (for Chinese), and prepared before any modeling is applied.