Wayne Williams

Professor Wiejie Pang

Data Science Foundation

November 23, 2025

Methodology

My project follows a step-by-step NLP workflow. The goal is to turn raw articles into structured topics that can be compared across English and Chinese.

1. Cleaning the Text

- Remove extra symbols, punctuation, and noise
- Convert English text to lowercase
- Remove stopwords in both languages
- For Chinese, use Jieba to cut sentences into usable word segments (because Chinese has no spaces)

2. Turning Text Into Numbers

To analyze text, it has to be represented numerically:

- Convert all words into TF-IDF vectors using Scikit-learn
- Limit the vocabulary to avoid extremely rare or overly common words
- Keep the features balanced across both languages

3. Topic Modeling

To discover the main themes:

- Run Latent Dirichlet Allocation (LDA) on the English dataset
- Run a separate LDA model on the Chinese dataset
- Extract the top representative words for each topic
- Compare which topics appear in each language and how strong they are
- Identify differences, such as which subjects are emphasized more in Chinese news vs. English news

4. Visualizing Results

My project uses:

- Word clouds for each topic
- Bar charts showing the top words per topic
- Tables listing key topics side-by-side (English vs. Chinese)
- Topic distribution charts to show which themes dominate each dataset

These visuals make it easy to see which ideas appear the most.

5. Tools Used

- Python
- Pandas (data handling)
- Jieba (Chinese segmentation)
- Scikit-learn (TF-IDF and LDA modeling)
- Matplotlib / Seaborn (visualizations)

6. What This Answers

My method directly answers your four project questions:

- What topics appear most often in each language?
- Which words define those topics?
- Where do English and Chinese news overlap?
- How can topic modeling reveal patterns and bias in each news system?