ECON6087: Textual Analysis for Economists

Assignment 2: Topic Modeling

Spring 2026

General Instructions

- **Submission:** Please submit a short report and your source code.
  - **Report Format:** Accepted formats include `.md`, `.doc`, `.docx`, `.pdf`, or embedded directly within a Jupyter Notebook (`.ipynb`).
  - **Code:** You may use any programming language.
- **File Format:** This assignment uses **.csv** files. Ensure your environment has the necessary dependencies installed to read this format (e.g., `pandas` or `polars`).
- **Reproducibility:** Ensure your code is well-commented. If your algorithms involve randomization (e.g., classifier initialization), set a fixed **random seed** to ensure your results are reproducible.

Topic Modeling

In this exercise, you will work with the People's Daily dataset.

1.1 Data Preparation

Download the following files from Moodle or Kaggle:

- `RenMin_Daily_*.csv`: News articles from People's Daily.

  All files contain two primary columns: `date` and `title`. You may use the optional `content` column for additional text data if your computational resources allow, but it is not required for this assignment.

1.2 Task Details

(a) Use any topic modeling algorithm of your choice (e.g., Latent Dirichlet Allocation) to identify topics within the news articles. You may use any vectorization method (e.g., Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), pre-trained word embeddings, etc.) to represent the text data before applying the topic modeling algorithm.

(b) Briefly describe the topic modeling algorithm you chose. You should choose the number of topics (K) appropriately. After applying the algorithm, analyze the results to identify the main topics present in the news articles. For each topic, list the top 10 words that are most representative of that topic.

(c) If you have limited computational resources, you may choose to use only a subset of the data (e.g. a random sample of articles). However, ensure that your sample is sufficiently large to capture the diversity of topics in the dataset.

1.3 Reporting Requirements

Report the following:

1. **Topics:** List the top 10 words for each of the topics identified by your topic modeling algorithm.
2. **Topic Distribution:** Provide a summary of the distribution of topics across the dataset (year). You can use tables, charts, or any other visualizations to illustrate this distribution effectively.
3. **Interpretation:** Provide a brief interpretation of the identified topics and their distribution. Discuss any trends or patterns you observe in the context of the news articles.