DATA 6150 Data Science Foundations
Fall 2025

Topic of Individual Project
Due at 11:59 pm on Sunday, Nov 2

Based on the introductions of individual projects, please decide the topic of your individual project. Fill in the following questions and submit the filled PDF version before the deadline.

1. Project Topic and Description

What is the topic or problem you plan to explore?
Provide a brief description explaining why it interests you and why it is worth investigating.
(2–4 sentences are sufficient.)

I want to explore what are some of the main topics that are in English and Chinese texts. I will use different articles and texts that I can find like the BBC News articles and different Chinese texts. The point of this is to see what are the topics that show up the most in both languages and analysis how they are different. I plan to test it by using the frequency model to classify words into a certain order.

2. Dataset Information

For the existing dataset(s), provide the link(s) and a short description of what the data contain.

Chinese: https://www.kaggle.com/datasets/ceshine/yet-another-chinese-news-dataset?utm_source=chatgpt.com

Chinese: https://github.com/gaussic/text-classification-cnn-rnn/tree/master/thucnews

Chinese: https://www.kaggle.com/datasets/ceshine/yet-another-chinese-news-dataset?utm_source=chatgpt.com

English:
https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.62?utm_source=chatgpt.com

3. Research or Analytical Questions

List at least four specific questions that you plan to investigate within this topic.
These questions should guide your analysis and connect directly to your problem statement.

1. What are the most common topics in English news compared to Chinese news?
2. Which words appear most often in each language's top topics?
3. Do both languages share similar main themes like business, sports, or politics?
4. How can topic modeling help us see the focus or bias in news coverage?

4. Quantitative Analysis (Modeling Component)

Among your questions**, identify at least one** that will require quantitative modeling or data-driven prediction (e.g., classification, regression, clustering, dimensionality reduction, NLP). Specify the question that need quantitative modeling. And specify the model or method you plan to use and briefly explain why it fits your question.

What are the main themes in English and Chinese news, and how do they compare?

Steps:

- Load and clean both datasets.
- For Chinese text, use Jieba to cut words into segments.
- Turn text into TF-IDF vectors.
- Apply LDA (Latent Dirichlet Allocation) to find top words for each topic.
- Show results with tables and word clouds.
- Compare topics between English and Chinese articles.
-

Tools: Python, Pandas, Scikit-learn, Matplotlib.

5. Expected Outcome or Insight (optional but recommended)

What do you hope to learn, discover, or demonstrate through this project?
How will answering your questions contribute to understanding the topic?

The project should show about 5 to10 key topics for each language. English news might talk more about tech and entertainment, while Chinese news might focus more on business and politics. The will give a simple visual way to understand what each language's media focuses on.

6. Suggestions or Comments for the Instructor

Do you have any questions, concerns, or suggestions for feedback on your project idea at this stage?

No, not at the moment.