Preprocessing Justification

Preprocessing Steps for Each Method

This study applies **Document-Term Matrix (DTM) exploration, cosine similarity analysis, and category classification using a Large Language Model (LLM)**. To ensure consistency and comparability across methods, preprocessing is standardized:

- **Stopword Removal (Chinese & English)**
  - **Chinese**: Based on the **Harbin Institute of Technology** stopword set, supplemented with additional words (e.g., "有个," "情况").
  - **English**: Based on **spaCy's** stopword list, with additional terms removed after content review.
- **Text Cleaning**
  - **Lowercasing** (English only), **removing punctuation** (`re.sub(r'[^\w\s]', '', text)`) and **digits** (`re.sub(r'\d+', '', text)`).
- **Language Separation**
  - **Using Unicode detection (`ord(char) > 127`)** to separate **Chinese** and **English** text.
  - **Processing them independently** to prevent segmentation errors.
- **Tokenization**
  - **Chinese**: `jieba.lcut()`
  - **English**: `spaCy`, including **lemmatization** and **stopword removal**.
- **Text Extraction**
  - Extracting **desc, comments_CN, comments_NonCN** from JSON files and storing them in a **DataFrame**.
- **TF-IDF Computation**
  - Using `TfidfVectorizer` to construct a **DTM matrix** and extract **high-frequency keywords per category**.
  - Adjusting **`min_df`, `max_df`, and `ngram_range`** to fine-tune feature selection.
- **Visualization**
  - **Word clouds** (`WordCloud`) and **bar charts** (`matplotlib`) for keyword representation.

Rationale Behind Preprocessing Choices

- **Stopword Removal**: Enhances keyword relevance by eliminating frequent but non-informative words.
- **Tokenization & Lemmatization**: Prevents fragmentation and ensures consistent term representation.
- **Punctuation & Digit Removal**: Eliminates noise, improving analysis accuracy.

- **TF-IDF Weighting**: Prioritizes informative words for topic extraction.

Impact on Results

- **Improved segmentation for mixed-language text**
- **Higher-quality text analysis by reducing noise**
- **Enhanced keyword extraction for meaningful themes**
- **More reliable LLM-based classification by filtering irrelevant terms**
- **Better visualization for intuitive interpretation**

Method Selection

Why These Methods?

A **multi-perspective approach** combining:

1. **DTM Exploration**
    - Identifies **key topics** within each category.
    - Serves as input for **similarity analysis and classification**.
2. **Cosine Similarity + Visualization**
    - Measures **comment clustering patterns**.
    - **Network graphs and PCA** uncover linguistic and thematic structures.
3. **LLM-Based Classification**
    - Automates categorization using **Zhipu AI**.
    - Useful when **annotated training data is limited**.

Research Questions Addressed

1. **Structural Differences Between Chinese & English Comments**
    - How do **topic diversity and clustering** vary across languages?
    - PCA scatter plots highlight **linguistic discussion trends**.
2. **Content Similarity & User Behavior**
    - Do highly similar comments suggest **automated interactions**?
    - Identifies **scripted engagement strategies**.
3. **Cross-Cultural Shared Themes**
    - Are there **common topics across linguistic groups**?
    - Network graphs highlight **content convergence**.
4. **Content Categorization**

- Can we distinguish **social interaction, brand marketing, fan culture, and cross-cultural exchange**?
- LLM classification provides **structured categorization**.

Results Analysis

Interpretation of Results

Cross-Border Marketing

- Both **Chinese & English** comments discuss **U.S. social trends**.
- Example: **"egg"** (English) and **"鸡蛋"** (Chinese) appear frequently, reflecting shared concerns about **inflation and shortages**.
- **Implication**: Rednote serves as a **cross-cultural information hub**, bridging global discussions.

Social Interaction & Cross-Cultural Communication

- **Positive sentiment dominates**, with frequent words like *"love," "friend," "support."*
- **Minimal negativity** suggests these discussions focus on **experience-sharing rather than controversy**.
- **Future Improvement**: Compare **synonymous sentiment words across languages** (e.g., *"love"* vs. *"喜欢"*).

Political Discourse

- **Race & social justice** are key themes, with:
  - **English**: *"black," "white"*
  - **Chinese**: *"歧视" (discrimination), "警察" (police)*
- **Implication**: Different linguistic approaches to discussing **race and policy**.
- **Future Improvement**: Use **LDA topic modeling** to break down discussions into subcategories like:
  - **Identity**
  - **Systemic Discrimination**
  - **Social Media Discourse on Race**

Comparative Reflection

Complementary & Contradictory Findings

- **DTM identifies topic structure, cosine similarity reveals interaction patterns, and LLM performs categorization**.
- **Potential contradictions**: LLM classification sometimes assigns **political topics** where DTM suggests **news-related discussions**, highlighting **pre-trained bias vs. actual keyword distribution**.

Most Valuable Methods

- **DTM for topic identification**

- **Cosine similarity for interaction pattern detection**
- **Cross-language sentiment comparison (future improvement)**
- **LLM classification + topic modeling for political discussions**