# CMPE255_Team_Project_InsightfulMiners

## Project Proposal: Cross-Language Topic Modeling with TF-IDF

### Objective:

To develop a robust cross-language topic modeling system that utilizes TF-IDF (Term Frequency-Inverse Document Frequency) to identify and align topics across documents written in different languages. The objective is to enhance the understanding of multilingual content by creating a model that can accurately and efficiently map topics from one language to another.

### Project Scope:

#### Inclusion Criteria:

- Selection of multilingual datasets for analysis.
- Development of preprocessing tools for language normalization.
- Application of TF-IDF for term importance weighting.
- Implementation of a topic modeling algorithm compatible with TF-IDF outputs.

### Exclusion Criteria:

- Non-textual data.
- Single-language datasets.
- Pre-existing topic models without TF-IDF integration.

## **Methodology**:

1. **Data Collection**: Gather multilingual datasets from specified sources, ensuring a variety of languages and topics.
2. **Preprocessing**: Normalize the data by cleaning, tokenizing, and stemming to prepare for cross-language analysis.
3. **TF-IDF Application**: Apply the TF-IDF algorithm to highlight key terms in each language, reducing the influence of common words.
4. **Topic Modeling**: Use an advanced topic modeling algorithm (e.g., Latent Dirichlet Allocation) to identify topics within each language corpus.
5. **Cross-Language Alignment**: Implement a strategy to align the topics discovered across different languages, possibly using bilingual dictionaries or machine translation.
6. **Evaluation**: Evaluate the model's effectiveness using metrics such as coherence, perplexity, and alignment accuracy.

## **Deliverables**:

- A cross-language topic modeling software tool.
- Documentation detailing the methodology, usage, and limitations of the tool.
- A final report presenting the findings, including a comparative analysis with monolingual topic models.
- A presentation summarizing the project outcomes and potential applications.