Information Retrieval: Final Project

1. Khmer Text Summarization

Description:
This project focuses on building an automatic summarization system for Khmer text. Explore both **extractive** and **abstractive** summarization methods and analyze challenges related to Khmer script, lack of spacing, and complex morphology. The system will be tested on Khmer news articles, educational content, or social media posts.

Key Tasks:

- Collect Khmer articles and create a summarization dataset
- Study extractive methods: TextRank, TF-IDF ranking, clustering-based extraction
- Explore abstractive methods: Seq2Seq, Transformers (mT5, KhmerBERT if available)
- Evaluate summaries using ROUGE or human ratings
- Identify Khmer-specific challenges (e.g., segmentation, formal vs. informal language)

2. Developing a Customized Khmer Stop-Word Removal System

Description:
Design a stop-word removal system specifically for Khmer. This includes identifying frequent functional words (e.g., “ជា”, “នៃ”, “ឆន”, “ក៏”, “ហើយ”), analyzing their linguistic roles, and evaluating their impact on IR tasks. The project should measure how stop-word removal affects **precision**, **recall**, **TF-IDF weighting**, and **search ranking** on Khmer documents.

Key Tasks:

- Build a Khmer stop-word list using frequency analysis + linguistic rules
- Implement stop-word removal in a preprocessing pipeline
- Test the impact on IR models or text classification
- Provide recommendations for standardized Khmer stop-word lists

3. Sentiment Analysis of Khmer Text Using ML

Description:
Perform sentiment analysis on Khmer social media posts, reviews, or news comments. Collect real Khmer text data, clean it, and apply preprocessing such as tokenization, normalization, and handling informal Khmer writing. Various ML models (SVM, Naive Bayes, Logistic Regression, LSTM, BERT-based Khmer models if available) can be compared.

Key Tasks:

- Build a labeled dataset of Khmer sentiment (positive/negative/neutral)
- Apply Khmer-specific preprocessing (Unicode normalization, slang handling)
- Train multiple ML models for comparison
- Analyze errors and challenges specific to Khmer sentiment classification

4. Building a Bag-of-Words Model for Khmer Text Classification

Description:
Develop a Bag-of-Words representation for Khmer documents and use it to classify texts into categories such as news, education, entertainment, or literature. Should evaluate preprocessing needs, handling large vocabularies, and apply dimensionality reduction methods like **TF-IDF**, **PCA** feature selection.

Key Tasks:

- Tokenize Khmer documents and create a vocabulary
- Build BoW and TF-IDF matrices
- Train classifiers (SVM, Random Forest, Logistic Regression, etc.)
- Compare performance before and after dimensionality reduction
- Discuss model challenges for Khmer (spacing, compound words)

5. Khmer Text Prediction

Description:
This project aims to develop a system that predicts the next word or suggests possible words as the user types in Khmer. Will investigate both statistical and deep learning approaches and examine the challenges in predicting Khmer words due to spacing issues, compound structures, and orthographic variants.

Key Tasks:

- Collect Khmer text corpus (news, social media, literature)
- Preprocess the data: tokenization, sentence segmentation
- Compare models:
    - N-gram language models
    - RNN/LSTM/GRU models
    - Transformer-based models (GPT-style, mBERT, mT5)
- Implement a simple demo: keyboard auto-suggestion or sentence completion
- Evaluate accuracy and relevance of predictions

Note: *Submit the GitHub link for code, slide, and report.*