SCHOOL OF SCIENCE & TECHNOLOGY

COURSE: **DSA 4020A Natural Language Processing**

Document Title: **Semester Project**

Semester Project: Language Technologies for African Indigenous Languages

Project Description:

Each of these projects allows students to engage with real-world NLP challenges, enrich their understanding of African language processing, and contribute to language technology for the underrepresented languages.

Project Duration: 4 Weeks

1. **Sentiment Analysis in African Languages**
   - **Objective**: Develop a sentiment analysis model for a selected African language.
   - **Tasks**: Collect or use an existing dataset of social media posts or product reviews in an African language. Preprocess the data, develop a sentiment analysis model, and evaluate its accuracy.
   - **Expected Outcome**: A sentiment classifier capable of determining the positive, negative, or neutral sentiment in texts of the chosen language.
2. **Named Entity Recognition (NER) in African Languages**
   - **Objective**: Build a Named Entity Recognition (NER) model to identify entities (e.g., names of people, places, organizations) in an African language text.
   - **Tasks**: Annotate text data in the chosen language for NER, train an NER model, and test the model's performance.
   - **Expected Outcome**: An NER tool that can identify and classify entities specific to the cultural and linguistic context of the language.
3. **Machine Translation for an African Language Pair**
   - **Objective**: Implement a basic machine translation system for translating between an African language and English or another prominent African language.
   - **Tasks**: Use existing parallel datasets, train a statistical or neural translation model, and evaluate it with metrics like BLEU score.
   - **Expected Outcome**: A working translation system that can translate basic sentences between the selected languages.
4. **Text Classification in African Languages**

- **Objective**: Develop a model to classify text documents in an African language (e.g., news articles, social media posts) by topic (e.g., sports, politics, entertainment).
- **Tasks**: Use or collect a labeled dataset, preprocess it, and train a classification model.
- **Expected Outcome**: A classifier that can categorize texts into predefined topics with a reasonable accuracy.

5. **Automatic Speech Recognition (ASR) for African Languages**
   - **Objective**: Develop a speech-to-text model for an African language with available audio data.
   - **Tasks**: Preprocess audio data, train an ASR model using a toolkit like Mozilla DeepSpeech, and evaluate its transcription accuracy.
   - **Expected Outcome**: An ASR tool capable of transcribing speech in the selected language.
6. **POS Tagging for African Indigenous Languages**
   - **Objective**: Build a Part-of-Speech (POS) tagger for an African language with existing annotated data.
   - **Tasks**: Use annotated text data to train a POS tagger using Hidden Markov Models (HMM) or Conditional Random Fields (CRF).
   - **Expected Outcome**: A POS tagging tool that accurately labels parts of speech in the text of the selected language.
7. **Word Sense Disambiguation (WSD) in African Languages**
   - **Objective**: Implement a Word Sense Disambiguation model for a specific African language.
   - **Tasks**: Compile a dataset with ambiguous words in context, build a WSD model, and evaluate its accuracy in disambiguating meanings.
   - **Expected Outcome**: A system that can identify the correct sense of words in context for better comprehension of African language text.
8. **Chatbot Development in an African Language**
   - **Objective**: Develop a simple rule-based or machine learning-based chatbot that can converse in an African language.
   - **Tasks**: Design and build a conversation dataset, train the chatbot, and conduct interaction tests.
   - **Expected Outcome**: A chatbot capable of handling basic conversational tasks in the selected language, with cultural nuances embedded in responses.

9. **Topic Modeling in African Languages**
    - **Objective**: Extract and identify topics from a corpus in an African language using topic modeling techniques.
    - **Tasks**: Preprocess the corpus, apply Latent Dirichlet Allocation (LDA) or other algorithms, and analyze the most prevalent topics.
    - **Expected Outcome**: An understanding of key topics in the dataset, with identified words or phrases common to each topic.
10. **Multilingual Sentiment Analysis for African Language Code-Switching**

- **Objective**: Develop a sentiment analysis model for text with code-switching between an African language and English.
- **Tasks**: Prepare or collect a dataset with mixed-language text, train a classifier to recognize sentiment, and evaluate its performance.
- **Expected Outcome**: A sentiment classifier adapted to the language-switching context common in many African texts.