CSC423: Special Topics

Title: NLP Term Project (40%)

OBJECTIVE

Overall, the objective of the Project is to build an accurate and efficient NLP application **IN NATIVE LANGUAGE(S) or Swahili**, which can be applied to various types of text data, such as product reviews, social media posts, or customer feedback.

You are required to develop an NLP (Natural Language Processing) application from the following categories i.e.

1. Spelling correction (Auto correction): Text suggestions in documents
2. Sentiment Analyzer: Analyze sentiments behind a phrase
3. Machine Translator: Given a source text, generate an equivalent text in a target language.
4. Handwriting recognition: Tell whose handwriting
5. Chatbot or Question Answering Machine. : Reading a sequence of text and then answering questions about that text

General Guidelines to be followed (Note: these are dependent on the type of App)
1. Data collection: Gathering a dataset of text reviews, customer comments or tweets
2. Data labelling : use human annotators to label a sample of the text e.g. with positive, negative or neutral sentiment labels in the case of a sentiment analyzer
3. Data preprocessing: Cleaning and preparing the text data for analysis by removing stop words, punctuation, and other noise.
4. Feature extraction: Creating features from the text data that can be used to train a machine learning model. For example, features could be the frequency of specific words, the length of the text, or the sentiment of the words.
5. Training the model: Using a machine learning algorithm such as a Naive Bayes or Support Vector Machine (SVM) to train a model to classify the text into one of the three sentiment categories.
6. Evaluating the model: Assessing the accuracy and performance of the model using a separate test dataset.
7. Deploying the model: Using the trained model to classify new text data in real-time.

NOTE: **THE NATURAL LANGUAGES INVOLVED SHOULD TARGET NATIVE LANGUAGES AND REGIONAL LANGUAGES with minimal inclusion of English or Indo European languages (for translation or documentation purposes)**