DSC-495 PROJECT REPORT **SPRING 2024**
ANAGH SHRIVASTAVA

English-to-Hindi Language Translation using Apache Spark & NLP

This project aimed to develop an effective language translation system to facilitate communication between English and Hindi speakers, leveraging the power of Apache Spark's distributed computing framework and natural language processing (NLP) techniques.

- **Background:**
  Hindi, a prominent language in South Asia, has a rich history dating back centuries. With its widespread use in India and significant diaspora communities worldwide, there is a growing need for accurate translation tools to bridge the linguistic gap between Hindi and English.

- **Approach:**
  The project utilized Apache Spark's parallel processing capabilities to efficiently handle the translation tasks. NLP methods were employed for text analysis, preprocessing, and the application of a specialized Hindi language model trained on extensive Hindi language data.

- **Dataset:**
  The "Hindi-English Truncated Corpus" dataset from Kaggle was used, consisting of English text paired with confirmed Hindi translations. This dataset served as the foundation for training and evaluating the translation model.

- **Execution:**
  The project was executed on a high-performance computing (HPC) cluster using the LSF scheduler and a custom wrapper script (lsf-spark-submit.sh) for resource allocation. The Apache Spark executable was configured with the necessary dependencies, including the spark-nlp package, and appropriate memory settings.

- **Performance Testing:**
  Performance testing was conducted by varying the number of CPU cores allocated to the translation tasks. The optimal performance was achieved with 8 CPU cores, resulting in a translation time of approximately 2.7 minutes.

- **Results:**
  The translated Hindi text generated by the model was evaluated against the confirmed English translations in the dataset, demonstrating the reliability and accuracy of the Hindi language model. The project successfully facilitated communication and understanding between Hindi and English speakers, fostering cultural exchange and collaboration.

- **Conclusion:**
  The project showcased the power of Apache Spark, NLP, and specialized language models in bridging linguistic barriers and promoting cross-cultural communication. Future directions include exploring GPU acceleration for improved performance, working on multiple applications, and further leveraging Apache Spark's capabilities.

- **Acknowledgments:**

The project was supervised by Prof. Andrew Petersen and guided by Mr. Qasim Adegbite and Mr. Vikram Pande. The resources and support provided by NC State University were instrumental in the successful completion of this project.