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

Anagh Shrivastava
DSC-495 (002)
Final Project

Background : HINDI Origins

Hindi, a prominent language in South Asia, has a rich history dating back centuries. Evidence suggests its significance in ancient Indian kingdoms. It's spoken primarily in India, with sizable communities in Nepal, Fiji, and other diaspora hubs. Through migration and cultural exchange, Hindi has spread to countries like Mauritius, Suriname, Trinidad and Tobago, Guyana, and South Africa, shaping local cultures and identities.

Some Indian/Hindi Origin Leaders in USA

Sundar Pichai
CEO, Alphabet INC.

Satya Nadella
CEO, Microsoft

Neal Mohan
CEO, YouTube

Overview of Apache Spark and Natural Language Processing (NLP)

- **Apache Spark**: A robust distributed computing framework adept at processing extensive datasets efficiently.
- **Parallel Processing**: Apache Spark enables parallel execution of translation tasks, ensuring optimal utilization of computational resources.
- **NLP**: Natural Language Processing (NLP) is a branch of artificial intelligence focused on computer-human language interactions.
- **Language Translation**: NLP techniques are applied for language translation, enhancing the efficiency and accuracy of translation pipelines.
- **Text Analysis and Processing**: NLP methods facilitate text analysis and processing tasks, enabling deeper insights into language data.

Language Model Used

- The Hindi language model is trained on extensive Hindi language data, enabling it to understand the linguistic nuances of the language.
- The model accurately translates Hindi text into English, providing a valuable tool for cross-lingual communication.
- The translation pipeline consists of several steps, including data preprocessing, translation using the Hindi language model, and output generation.
- Input data is processed through the pipeline, resulting in translated output that facilitates communication between Hindi and English speakers.

Thought Process

Machine Translation Model:

```
marian = MarianTransformer.load("/home/ashriva3/cache_pretrained/opus_mt_en_hi_xx_3.1.0_2.4_1622560150184") \
    .setInputCols(["sentence"]) \
    .setOutputCol("translation") \
    .setMaxInputLength(30) \
    .setMaxOutputLength(30)

# Define translation UDF
pipeline = Pipeline() \
    .setStages([
      documentAssembler,
      sentence,
      marian
    ])
```

Dataset used: https://www.kaggle.com/datasets/umasrikakollu72/hindi-english-truncated-corpus

```
anagh_df = spark.read.csv("/share/dsc495s24/commonData/Hindi_English_Truncated_Corpus.csv")
```

Dastaset Evaluation

- **Input Data:**

A Dataset consisting of English text paired with confirmed Hindi translations.

A	B	C	D
source	english_sentence	hindi_sentence	
ted	politicians do not have permission to do what needs to be done.	राजनीतिज्ञों के पास जो कार्य करना चाहिए, वह करने कि अनुमति नहीं है .	
ted	I'd like to tell you about one such child,	मई आपको ऐसे ही एक बच्चे के बारे में बताना चाहूंगी,	

- **Output Data:**

Translated output includes the ID and the translated Hindi text generated by the model.

A	B	C	D
id	result		
0	सूचना (_m)		
1	नेताओं को क्या करने की अनुमति नहीं है ।		
2	मैं तुम्हें एक ऐसे बच्चे के बारे में बताना चाहता हूँ,		

Execution

Conda Activation:

conda activate ../commonInputs/nlp521j11_env

Spark Executable:

$SPARK_HOME/bin/spark-submit --packages com.johnsnowlabs.nlp:spark-nlp_2.12:5.2.1 --conf spark.driver.memory=10g --conf spark.executor.memory=10g ../commonInputs/translation3.2.py

LSF Wrapper:

bsub -Is -n 8 -R "span[ptile=4]" -W 100 -env "all,SPARK_PID_DIR=/scratch,SPARK_EXECUTOR_MEMORY=40G,SPARK_LOCAL_DIRS=/scratch,SPARK_DRIVER_MEMORY=40G" ./lsf-spark-submit.sh --conf "spark.eventLog.enabled=true" --conf "spark.eventLog.dir=/share/dsc495s24/$USER/apache-spark/spark-events" --jars $(echo /home/ashriva3/.ivy2/jars/*.jar | tr ' ' ',') ../commonInputs/translation3.2.py

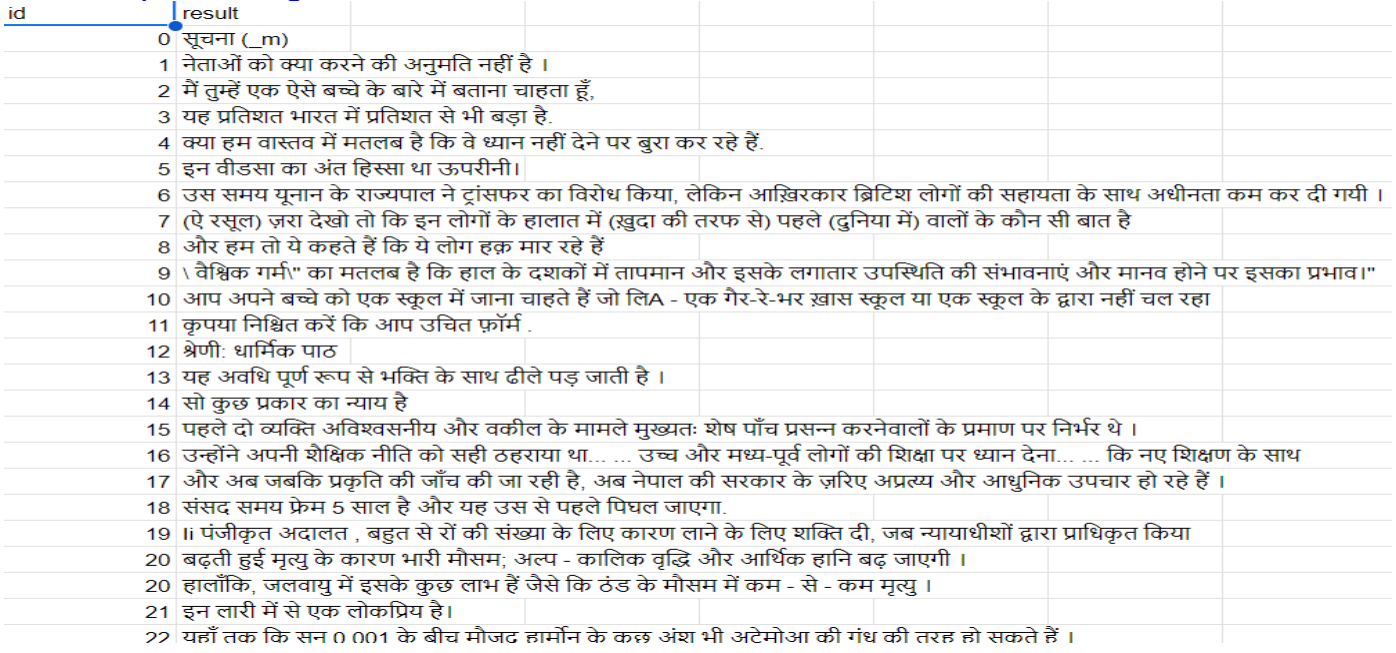
```
import sparknlp
from sparknlp.base import *
from sparknlp.annotator import *
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType
from pyspark.ml import Pipeline
from pyspark.sql.functions import monotonically_increasing_id

# Start Spark NLP session

spark = sparknlp.start()

# Load necessary components

documentAssembler = DocumentAssembler() \
    .setInputCol("text") \
    .setOutputCol("document")

sentence = SentenceDetectorDLModel.load("/share/dsc495s24/commonInputs/cache_pretrained/sentence_detector_dl_xx_2.7.0_2.4_1609610616998/") \
    .setInputCols(["document"]) \
    .setOutputCol("sentence")
```

```
marian = MarianTransformer.load("/home/ashriva3/cache_pretrained/opus_mt_en_hi_xx_3.1.0_2.4_1622560150184") \
    .setInputCols(["sentence"]) \
    .setOutputCol("translation") \
    .setMaxInputLength(30) \
    .setMaxOutputLength(30)

# Define translation UDF
pipeline = Pipeline() \
    .setStages([
      documentAssembler,
      sentence,
      marian
    ])
```

```
# Remove unnecessary column _c0
anagh_df = anagh_df.drop("_c0")

anagh_df_final = anagh_df.withColumn("id", monotonically_increasing_id())
anagh_df_final = anagh_df_final.select("id", "_c1")
anagh_df_final = anagh_df_final.withColumnRenamed("id", "id").withColumnRenamed("_c1", "text")

anagh_df_final.show()
anagh_df_final = anagh_df_final.limit(100)

result = pipeline.fit(anagh_df_final).transform(anagh_df_final)
exploded_result = result.selectExpr("id", "explode(translation.result) as result")
exploded_result.show(truncate=False)

exploded_result.write.csv("hindi_result.csv", header=True, mode="overwrite")


#data = spark.read.text("/share/dsc495s24/commonData/Test10.txt")
#data.show()

# Apply translation UDF to create a new column

#translated_data = data.withColumn("translation_hindi", translate_udf(data["Yoruba"]))

# Select top 10 rows

#top_10_rows = translated_data.limit(10)
top_10_rows_hindi = exploded_result.limit(10)
```

Output

- https://docs.google.com/spreadsheets/d/1pC01M0hnsdLxvTzkqmrnV2fQ6YyBxsP7zMNTc6tPo-Y/edit?usp=sharing



id	result
0	सूचना (_m)
1	नेताओं को क्या करने की अनुमति नहीं है ।
2	मैं तुम्हें एक ऐसे बच्चे के बारे में बताना चाहता हूँ,
3	यह प्रतिशत भारत में प्रतिशत से भी बड़ा है.
4	क्या हम वास्तव में मतलब है कि वे ध्यान नहीं देने पर बुरा कर रहे हैं.
5	इन वीडसा का अंत हिस्सा था ऊपरीनी।
6	उस समय यूनान के राज्यपाल ने ट्रांसफर का विरोध किया, लेकिन आख़िरकार ब्रिटिश लोगों की सहायता के साथ अधीनता कम कर दी गयी ।
7	(ऐ रसूल) ज़रा देखो तो कि इन लोगों के हालात में (ख़ुदा की तरफ से) पहले (दुनिया में) वालों के कौन सी बात है
8	और हम तो ये कहते हैं कि ये लोग हक़ मार रहे हैं
9	\ वैश्विक गर्म\" का मतलब है कि हाल के दशकों में तापमान और इसके लगातार उपस्थिति की संभावनाएं और मानव होने पर इसका प्रभाव।"
10	आप अपने बच्चे को एक स्कूल में जाना चाहते हैं जो लिA - एक गैर-रे-भर ख़ास स्कूल या एक स्कूल के द्वारा नहीं चल रहा
11	कृपया निश्चित करें कि आप उचित फ़ॉर्म .
12	श्रेणी: धार्मिक पाठ
13	यह अवधि पूर्ण रूप से भक्ति के साथ ढीले पड़ जाती है ।
14	सो कुछ प्रकार का न्याय है
15	पहले दो व्यक्ति अविश्वसनीय और वकील के मामले मुख्यतः शेष पाँच प्रसन्न करनेवालों के प्रमाण पर निर्भर थे ।
16	उन्होंने अपनी शैक्षिक नीति को सही ठहराया था... ... उच्च और मध्य-पूर्व लोगों की शिक्षा पर ध्यान देना... ... कि नए शिक्षण के साथ
17	और अब जबकि प्रकृति की जाँच की जा रही है, अब नेपाल की सरकार के ज़रिए अप्रत्य और आधुनिक उपचार हो रहे हैं ।
18	संसद समय फ्रेम 5 साल है और यह उस से पहले पिघल जाएगा.
19	Ii पंजीकृत अदालत , बहुत से रों की संख्या के लिए कारण लाने के लिए शक्ति दी, जब न्यायाधीशों द्वारा प्राधिकृत किया
20	बढ़ती हुई मृत्यु के कारण भारी मौसम; अल्प - कालिक वृद्धि और आर्थिक हानि बढ़ जाएगी ।
20	हालाँकि, जलवायु में इसके कुछ लाभ हैं जैसे कि ठंड के मौसम में कम - से - कम मृत्यु ।
21	इन लारी में से एक लोकप्रिय है।
22	यहाँ तक कि सन 0.001 के बीच मौजूद हार्मोन के कुछ अंश भी अटेमोआ की गंध की तरह हो सकते हैं ।

Performance Testing

No of CPU cores	Time Taken
8	2.7 mins
4	3.4 mins
6	3.2 mins
2	4.8 mins

Results

Spark 3.4.1 **History Server**

Event log directory: file:/share/dsc495s24/ashriva3/apache-spark/spark-events

Last updated: 2024-04-22 23:43:31

Client local time zone: America/New_York

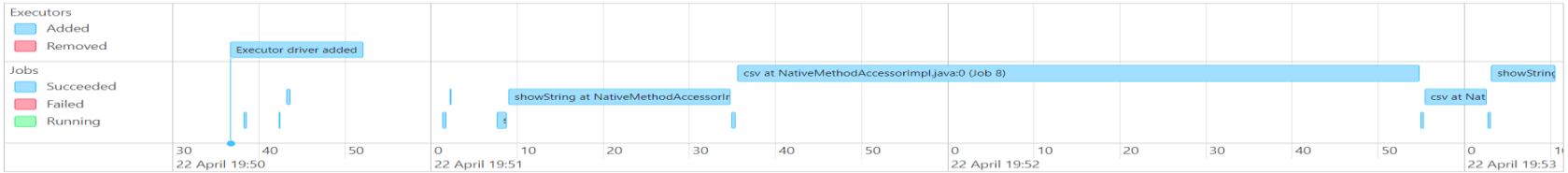
Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.4.1	local-1713843245483	Spark NLP	2024-04-22 23:34:03	2024-04-22 23:38:49	4.8 min	ashriva3	2024-04-22 23:38:49	Download
3.4.1	local-1713842534101	Spark NLP	2024-04-22 23:22:11	2024-04-22 23:25:37	3.4 min	ashriva3	2024-04-22 23:25:37	Download
3.4.1	local-1713842173657	Spark NLP	2024-04-22 23:16:12	2024-04-22 23:19:27	3.2 min	ashriva3	2024-04-22 23:19:27	Download
3.4.1	local-1713841642226	Spark NLP	2024-04-22 23:07:21	2024-04-22 23:10:01	2.7 min	ashriva3	2024-04-22 23:10:01	Download

Spark Jobs (?)

User: ashriva3
Total Uptime: 2.7 min
Scheduling Mode: FIFO
Completed Jobs: 13

Event Timeline
Enable zooming



Effectiveness

- The accuracy of our translation is evaluated against confirmed English translations, demonstrating the reliability of our Hindi language model.

Impact

- My project has a significant impact on facilitating communication and understanding between Hindi and English speakers, fostering cultural exchange and collaboration.

Conclusion

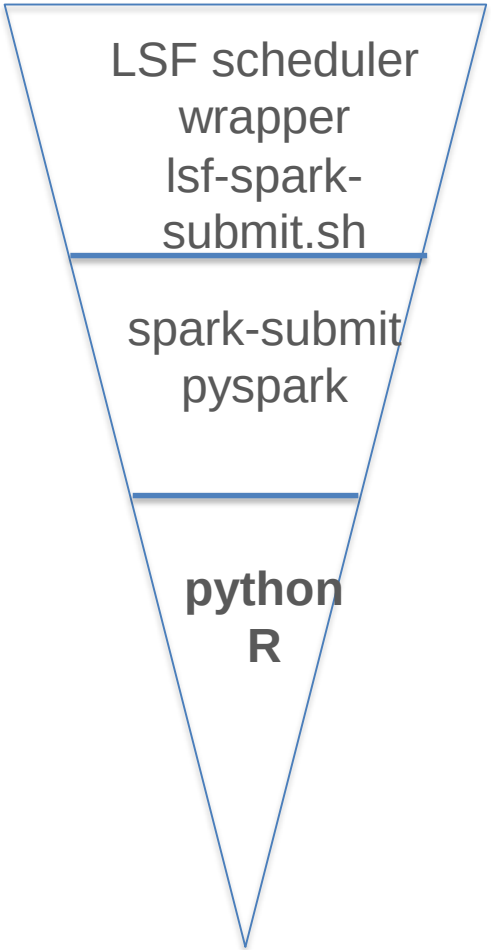
- **Summary**

  My project showcases the power of technology, specifically Apache Spark, NLP, and specialized language models, in bridging linguistic barriers and promoting cross-cultural communication.

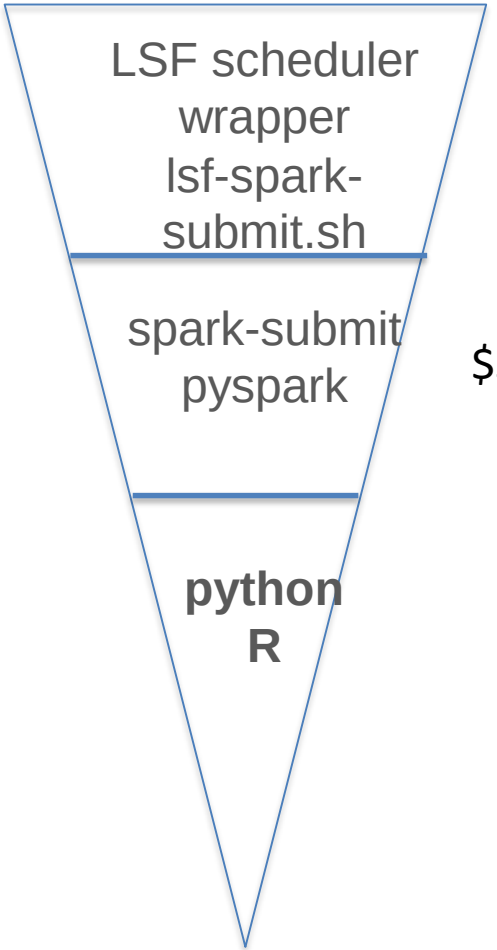
- **Future Directions**

  My project was only able to use the local computer for it's execution. I aim to figure out the issue and work on multiple apps in the future with an interest in learning in the world of Apache Spark.

  Also, to test on GPU for better performance



•/lsf-spark-submit.sh --jars $(echo /home/ashriva3/.ivy2/jars



$SPARK/bin/spark-submit…
on HPC-VCL

Acknowledgement

- Proff. Andrew Petersen
- Mr. Qasim Adegbite
- NC State
- Mr. Vikram Pande

THANK YOU!