PROJECT PROPOSAL
CONDUCTING SENTIMENT ANALYSIS USING PySpark

INTRODUCTION

Sentiment analysis, a natural language processing technique, serves to determine the emotional connotation within a body of text, classifying it as positive or negative. This invaluable tool allows us to discern the emotional nuances within textual content, categorizing them as positive or negative.

The primary objective of our project is to carry out Sentiment Analysis utilizing PySpark. PySpark, a Python library, excels in the efficient processing of extensive datasets. Equipped with a distributed computing framework and integration with Apache Spark, PySpark empowers us to construct a scalable and effective sentiment analysis system capable of handling substantial quantities of textual data.

DATASET

The primary data source for this research will be social media platform X.com (formerly known as twitter). The data will be tweets from ordinary as well as government accounts collected through the X.com API. The tweets will be classified as follows -

Positive Tweet:

"Just heard about the humanitarian aid efforts in Ukraine. It's heartwarming to see people coming together to support those in need. 🙏❤️ #UkraineStrong #Hope"

Negative Tweet:

"This ongoing conflict in Ukraine is devastating. It's disheartening to see the suffering and loss of life. When will this madness end? 💔 #UkraineCrisis #EndTheWar"

Neutral Tweet:

"Keeping an eye on the developments in the Russia-Ukraine conflict. It's a complex situation with global implications. #RussiaUkraine #CurrentEvents

The link to the API is *https://developer.twitter.com/en/products/twitter-api.*

Data Preprocessing

1. **Data extraction** : For the initial milestone, our focus will be on retrieving data from the provided dataset. Within the PySpark framework, we have the flexibility to employ different APIs, including Spark SQL, PySpark DataFrame API, or PySpark RDD API, for the purpose of data extraction.

2. **Data Cleaning and Preprocessing:** In this initial phase, our primary objective is to ensure the dataset's cleanliness and tidiness. This includes addressing the following aspects:
    a) **Handling Missing Data & Duplicate removal:** We'll identify and manage any missing values within the dataset to prevent potential issues during analysis. Duplicates will be identified and removed, promoting data integrity.
    b) **Data Cleansing:** In the data cleansing process, we will ensure the removal of stop words, punctuation, special characters, and handling of numerical data. Normalization, such as lowercasing, will be applied to maintain uniformity in text. These steps collectively aim to refine the data for sentiment analysis of the Russia-Ukraine conflict, eliminating irrelevant elements and enhancing the quality of the text.
    c) **Text Data Preprocessing & Splitting:** This involves text operations such as tokenization, stemming and lemmatization (reducing words to their base form) to prepare textual data for analysis. Once the data is prepared, the entire dataset is split into a Training, Validation & Test dataset.

3. Model Development

We plan to develop a lightweight softmax logistic regression classification model

4. Evaluation

- ☐ Classification task: Accuracy score, precision, recall
- ☐ Performance metric: Latency. Calculation of the time taken to analyze and classify a single tweet. Lower latency is crucial for real-time applications.

5. Milestone Goals

- ☐ **Using a Pre-labeled dataset to train our model**
    - ☐ Select a pre-labeled twitter sentiment dataset
    - ☐ Preprocess data, see 2.
    - ☐ Train models using pyspark ML : TF-IDF Logistic Regression (softmax)
- ☐ **Real-time Data pipeline**
    - ☐ Using twitter API, select a topic, generate data stream with text and timestamps, and store them in spark DF
    - ☐ Process data
    - ☐ Predict with model

6. Final Project Goals

a. A final graph showing the latency of inference, analyzing how it evolves over time as it processes a continuous stream of tweets.
b. A real-time data pipeline storing and processing tweets relevant to interested search term
c. A time series counting the frequency of each positive, negative, neutral tweets in regards to the topics selected (if time permits)