Case of Study of a Distributed Machine Learning Pipeline

Supervisor
Prof. Massimo Callisto De Donato

Students
Luca Bianchi
Michele Martini
Angelica Berdini

Aim and Objectives

The aim of the project is to study different ML libraries and how they work in a distributed environment.

Creating a distributed environment with Docker

Study different approaches

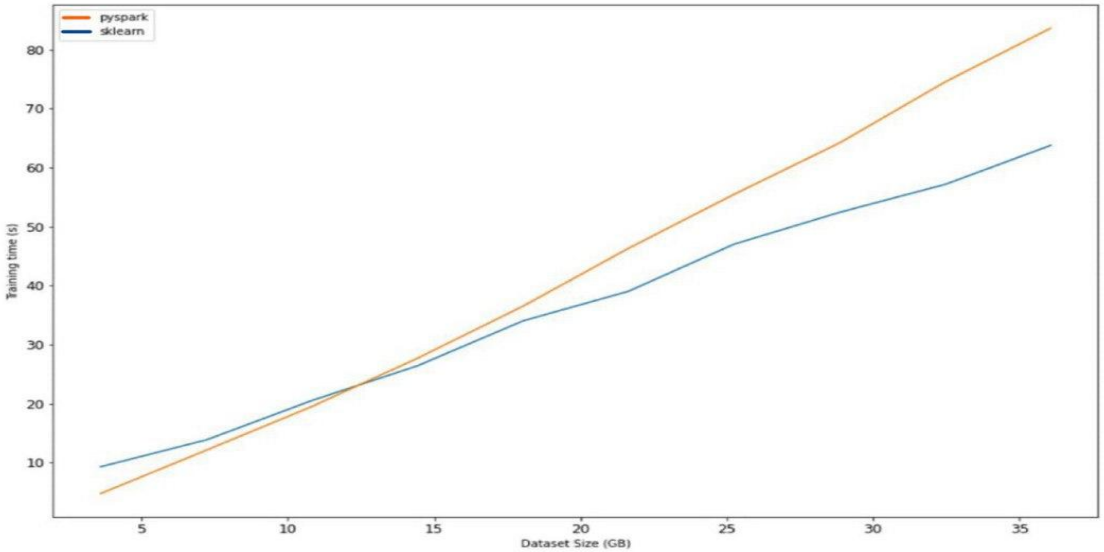
Extract meaningful insight

Non-distributed Pipeline

Distributed Pipeline

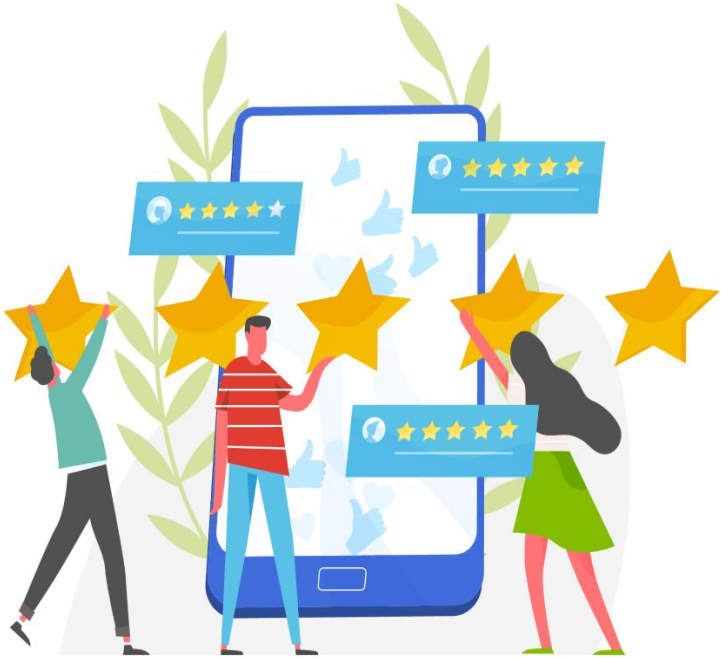
Performance: PySparkML vs Scikit-learn

Pyspark generally works better as the dataset volume increases



https://medium.com/geekculture/when-should-you-use-pyspark-over-scikit-learn-b10b91e41252

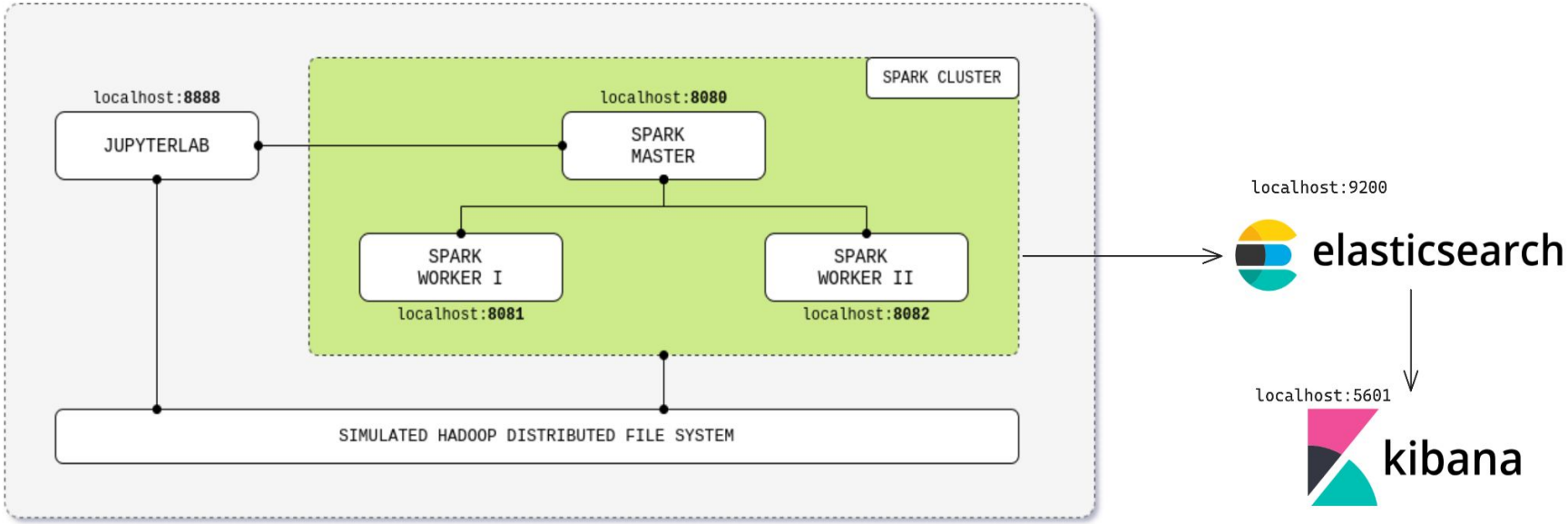
kaggle



Yelp Dataset on Kaggle

The Yelp **dataset** is a rich collection of real-world data encompassing information on businesses, **reviews**, user interactions, pictures, tips, business attributes, and aggregated check-ins from multiple metropolitan areas.

Project architecture



Sentiment analysis

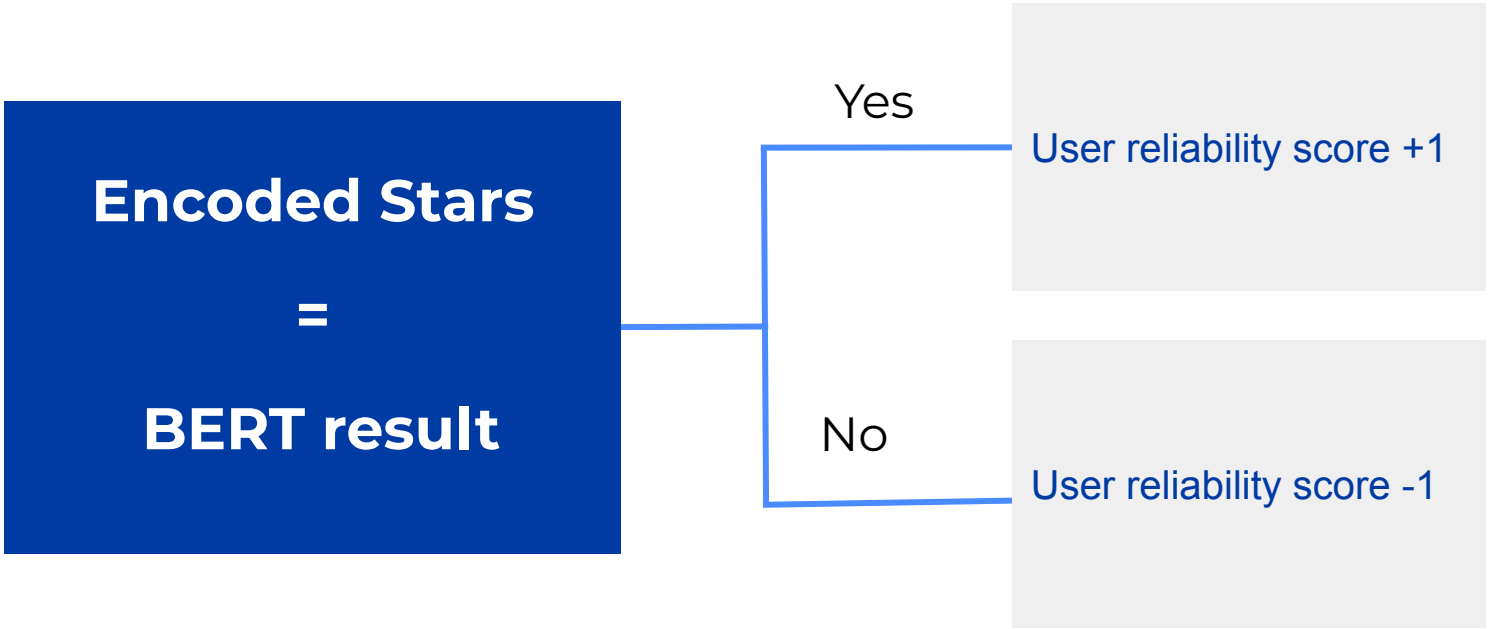
PySpark ML

Bert-sentiment

To train	Pre-trained
Results depends on stars and text	Results depends only on text
Low accuracy depending on low reliability	High accuracy

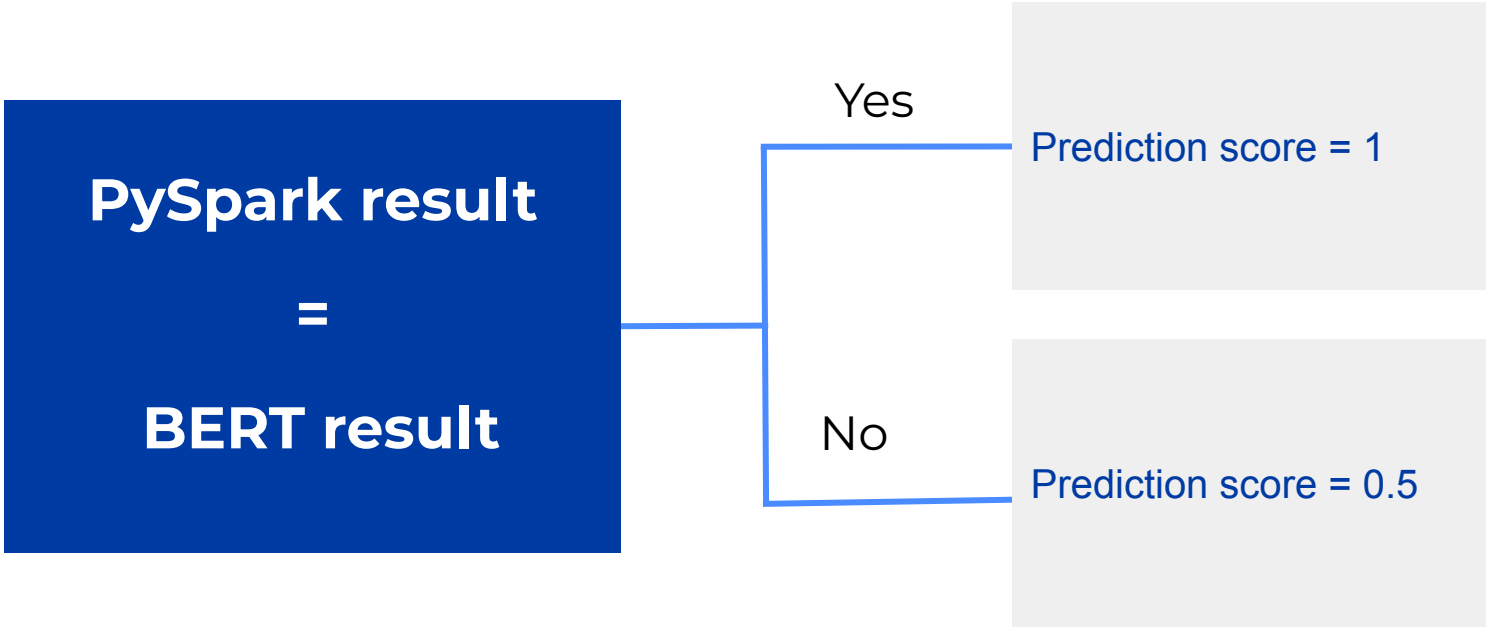
User reliability

User reliability is calculated for each user by comparing the sentiment and the stars.



Review reliability

Review reliability is calculated for each review with a simple formula. First we calculate prediction score



Review reliability

```
[ ] reliability = reliability.withColumn('reliability_score', reliability['prediction_score'] * (reliability['stars'] * reliability['words']))
```

```
reliability.show()
```

```
+--------------------+--------------------+--------------------+-----+-----+--------------+--------------+---------------+-----------------+
|         business_id|           review_id|             user_id|stars|words|prediction_text|bert-sentiment|prediction_score|reliability_score|
+--------------------+--------------------+--------------------+-----+-----+--------------+--------------+---------------+-----------------+
|--ZVrH2X2QXBFdCil...|5IJ8bbgtuaYY44jyl...|QD0DRN_ZY5czjHjzo...|  3.0|   50|      negative|      positive|            0.5|             75.0|
|--ZVrH2X2QXBFdCil...|1OO2oZ33wxPFH66e6...|Tf-TSPR3nqA_7b7_4...|  5.0|   47|      positive|       neutral|            0.5|            117.5|
|--ZVrH2X2QXBFdCil...|kE-NdGwUA1zbqNp9M...|Yy8JcvtMoNajJJW7k...|  5.0|   60|      positive|      positive|            1.0|            300.0|
|--_9CAxgfXZmoFdNI...|hCjfr9owNP4NfiDtX...|4Fjq-rolzbjKwzcdb...|  4.0|  106|       neutral|       neutral|            1.0|            424.0|
|-02xFuruu85XmDn2x...|pSE4t801nC2dX8dEQ...|5hhGQEj5K2urQ1Bcs...|  5.0|  122|      negative|      positive|            0.5|            305.0|
|-02xFuruu85XmDn2x...|UmXZrVok3IQAkkhq8...|lTF5IE8pl0wRMM2WE...|  5.0|  110|      positive|       neutral|            0.5|            275.0|
|-02xFuruu85XmDn2x...|0_-8nKLlT25m0tPed...|qe9cM4t63vKLhFaqd...|  5.0|   46|      positive|      positive|            1.0|            230.0|
|-0Ym1Wg3bXd_TDz8J...|LlvCXMS0Am_zSzDBk...|Wa-DgCDkaB3O0xP3c...|  5.0|  407|      negative|       neutral|            0.5|           1017.5|
|-0fvhILrC9UsQ6gLN...|tdgbQ7ZhWVIl_5uV3...|U-dNFjVZ907wxEFiO...|  5.0|   95|      positive|       neutral|            0.5|            237.5|
|-0fvhILrC9UsQ6gLN...|FUoItAyjds8jVyNWg...|KGnPTPP-i2l3__OTz...|  5.0|   78|      positive|       neutral|            0.5|            195.0|
|-0fvhILrC9UsQ6gLN...|gf9Cdnqe0K_ZcLTA4...|LsvUxdydAazds6ZV6...|  3.0|   48|      positive|      negative|            0.5|             72.0|
|-1MhPXk1FglglUAmu...|sGFHsKZcK7Ldw8T_V...|Fv3v5qxkb5CA9kMdY...|  3.0|   61|      positive|      positive|            1.0|            183.0|
|-1MhPXk1FglglUAmu...|282KkoS_qCeHX4twB...|bDwBtC0jk3is-qLFl...|  4.0|  270|       neutral|       neutral|            1.0|           1080.0|
|-1MhPXk1FglglUAmu...|WUuUVkVwio_0ED8Qr...|RkG5JsU070EZi-ptk...|  5.0|  240|      negative|       neutral|            0.5|            600.0|
|-1MhPXk1FglglUAmu...|HHpsfEqFewJuf3x75...|Yy-_hY62XH2XTCdHu...|  2.0|  120|       neutral|      negative|            0.5|            120.0|
|-1MhPXk1FglglUAmu...|pHxXa8SnGSM9WZg-N...|YXayxgxuR-CBEWvNg...|  3.0|   60|      positive|      positive|            1.0|            180.0|
|-1MhPXk1FglglUAmu...|iiRj2BGfC_pdjSxUO...|S-wipfsarZla1sby1...|  5.0|   54|      positive|       neutral|            0.5|            135.0|
|-1ueCbvIpUPi8KT95...|J-U6n26u1FxfoQu14...|ZRu1ybTPbTox4BU1J...|  4.0|   36|      positive|       neutral|            0.5|             72.0|
|-1ueCbvIpUPi8KT95...|8Nhh8OckuZantthme...|gkIdNDxm_V-tZayX7...|  2.0|  142|       neutral|      negative|            0.5|            142.0|
|-1ueCbvIpUPi8KT95...|zuWjhvGK0FRmXygzb...|867-opKCfFRsqVOg7...|  5.0|   78|      positive|      positive|            1.0|            390.0|
+--------------------+--------------------+--------------------+-----+-----+--------------+--------------+---------------+-----------------+
only showing top 20 rows
```

Conclusions

Research

- Distributed environments as a future development for data science
- PySpark can already outperform standard Data Science technologies on certain scenarios

Conclusions

Study Case

- Reliability played a fundamental role in the trained model
- Complete distributed machine learning pipeline

Conclusions

CONS

- Still new technology and needs many updates
- Cluster execution needs heavy performance

Further Works

- Improve reliability system considering “sentence-distance” approaches
- Leverage on cloud services to have a lighter execution
- Improve Recommendation System using more sophisticated algorithms and approaches

Thanks

For the attention