Apache Spark

In this directory I have two projects that use `Apache Spark` framework for distributed data processing for . In detail:

1. `RDD.ipynb` : This Python notebook leverages the PySpark `RDD API` to analyze a parallel corpus of transcriptions from the European Parliament, containing texts in Greek and English. It involves tasks such as:

   a. loading & preprocessing the text
   b computing the 10 most frequent words in each language, and using the parallel corpus to mine translation pairs by pairing words found in lines of equal length in both languages.

2. `DataFrames_SQL.ipynb` : this Python notebook leverages the `PySpark DataFrames/SQL API` to analyze a dataset of Los Angeles Parking Citations, which has been preloaded into an HDFS cluster. Tasks performed include:

   1. loading, exploring and preprocessing the dataset
   2. finding the maximum fine amount
   3. finding the top 20 most frequent vehicle makes
   4. finding what's the most frequent colour value for Toyotas