Analytics of big data - Targil 1
הגשה עד 26.3.19

The goal of this exercise is to learn how to use Spark for processing and preparing data .
Choose one dataset from the following site: https://archive.ics.uci.edu/ml/index.php
The dataset should contains:

a. at least 10000 rows
b. more than 10 columns
c. a timestamp column.
d. A text column.

1. Load your dataset.
2. Transform the data set into RDD pair, where the key is the unique id and the value is a Python list which contains the rest of the columns.
3. Choose 5 important columns and for each column
   a. Count the distinct values in each one of them.
   b. Create histogram to analyze the distribution of the above columns (normalize?).
   c. Explain your results.
4. Fill in bad or missing value (zero,mean or median). Explain your solution.
5. Transform 2 columns which contains categorical features to numerical values. Explain your method.
6. Transform the timestamp into categorical features (hour, day, month)
7. Choose 2 columns and normalize them.
8. Transform one of your text features.
   a. Do tokenization.
   b. Stop word removal - use the function StopWordsRemover from http://spark.apache.org/docs/latest/ml-features.html#stopwordsremover
   c. Binary vectorization.

הוראות הגשה:

1. יש לשלוח לבודק את קבצי ה-JUPYTER עם ההסברים והתוצאות בתוך הקובץ.
2. יש להעביר הרצאה של 5-10 דקות בכיתה בשיעור של ה3.4.19