Spark - Exercises

Exercise MLlib

- Classification problem
- Input:
  - A training data set containing a set of sentences
    - One sentence per line
    - Schema
      - label: 1 (Spark related sentence) or 0 (Non-spark related sentence)
      - text: a sentence about something
  - A set of unlabeled sentences

Exercise MLlib

- Output:
  - For each unlabeled sentence the predicted class label value by using a logistic regression algorithm
- You must train the model by using as input two predictive features:
  - The number of words in each sentence
  - A Boolean value associated with the presence/absence of the word "Spark" in the sentences

Exercise MLlib

- Training data

```
label,text
1,The Spark system is based on scala
1,Spark is a new distributed system
0,Turin is a beautiful city
...
```

- Unlabeled data

```
label,text
,Spark performs better than Hadoop
,Turin is in Piedmont
...
```