Machine learning with Spark using PySpark

Objectives

- Review
  - Spark
- Introduction to Spark MLlib (Spark ML) documentation
  - Individual
  - Pair
- Your time today (capstones/resumes/other)

Spark Review

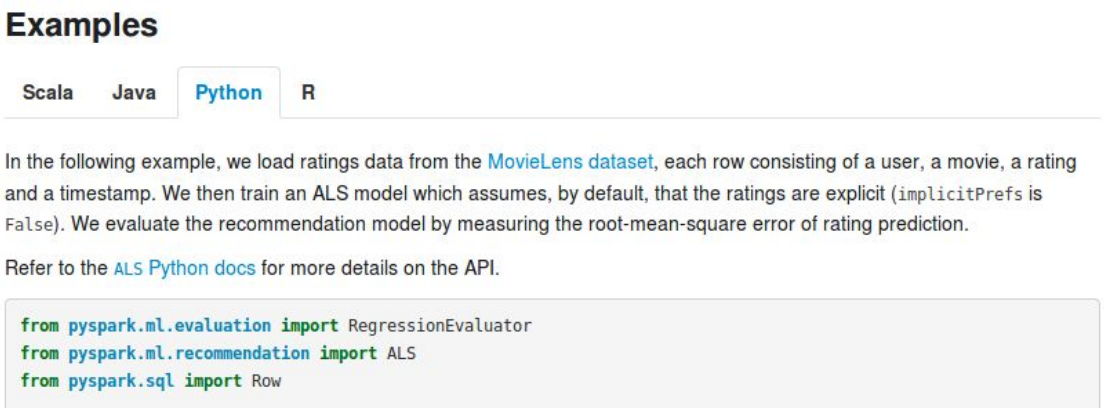
- What is the point of Spark? (What's it for?)
- What part of the Hadoop ecosystem does Spark typically replace?
  - Why is it better?
- What are the four main parts of a Spark application?
- What are the two APIs through which we interact with data in Spark?
  - How are they different?
- What are the types of operations required to get results out of data in Spark?
  - What's lazy evaluation? What's a DAG?
- If you were to whiteboard making a Spark application, what steps would be in it?

Spark MLlib (Spark ML)

https://spark.apache.org/docs/latest/ml-guide.html

- When you are doing machine learning, DataFrames (or DataSets in case of Java or Scala) provide better performance.
- For this reason, the RDD-based version of the machine learning libraries in no longer being developed (MLlib)
- Nice examples in guide (in 4 languages)

Recommenders (collaborative filtering) example:



Examples

Scala | Java | Python | R

In the following example, we load ratings data from the MovieLens dataset, each row consisting of a user, a movie, a rating and a timestamp. We then train an ALS model which assumes, by default, that the ratings are explicit (`implicitPrefs` is `False`). We evaluate the recommendation model by measuring the root-mean-square error of rating prediction.

Refer to the `ALS` Python docs for more details on the API.

```
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.sql import Row
```

Assignments today

- Individual
  - Text processing and sentiment analysis
  - Provided solution uses nltk -> **DO NOT download all of nltk. Just do this ---------->**
- Pair
  - Spark Cluster on EC2
  - K-means clustering of Wikipedia articles
- Optional: your time today (capstones/resumes/other)

```
Python 2.7.13 |Anaconda custom (64-bit)| (default, Dec 20 2016, 23:09
Type "copyright", "credits" or "license" for more information.

IPython 5.1.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details

In [1]: import nltk

In [2]: nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/frank/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Out[2]: True

In [3]: nltk.download('averaged_perceptron_tagger')
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/frank/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
Out[3]: True

In [4]: nltk.download('stopwords')
[nltk_data] Downloading package stopwords to /home/frank/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
Out[4]: True
```