Lab Distributed Big Data Analytics

Worksheet-4: **Spark ML and SANSA**

Dr. Hajira Jabeen, **Gezim Sejdiu**, **Prof. Dr. Jens Lehmann**

May 9, 2017

In this lab we are going to perform basic Spark ML and SANSA operations (described on "Spark Fundamentals II (Spark MLlib)" and "SANSA").
In this lab, you will use MLlib and SANSA to find out the subject distribution over nt file. The purpose is to demonstrate how to use the Spark MLlib and SANSA using Spark.

IN CLASS

1. Spark ML
    a. After a file (page_links_simple.nt.bz2) have been downloaded, unzipped, and uploaded on HDFS under /yourname folder you may need to create an RDD out of this file.
    b. First create a Scala class `Triple` containing information about a triple read from a file, which will be used as schema. Since the data is going to be type of .nt file which inside contains rows of triples in format `<subject> <predicate> <object>` we may need to transform this data into a different format of representation. Hint: Use `map` function.
    c. Create an RDD of a `Triple` object
    d. Use the filter transformation to return a new RDD with a subset of the triples on the file by checking if the first row contains "#", which on .nt file represent a comment.
    e. Implement **TF-IDF** algorithm for finding the most used classes on the given dataset.
        i. TF-IDF in Spark uses
            1. TF: HashingTF is a Transformer which takes sets of terms and converts those into fixed-length feature vectors.
            2. IDF:IDF is a Estimator which fits on a dataset and produces an IDFModel which takes feature vectors and scales each column.
        ii. List classes into a separate RDD.
        iii. Split each label into words using Tokonizer. For each sentence we use Hashing TF to hash the sentence into a feature vector. And then use IDF to rescale the feature vectors.
        iv. Pass this feature vector into a learning algorithm.

   f. Collect and print the results.

2. SANSA
   a. After a file (page_links_simple.nt.bz2) have been downloaded, unzipped, and uploaded on HDFS under /yourname folder you may need to create an RDD out of this file.
   b. Read a RDF file by using SANSA and retrieve a Spark RDD representation of it.
   c. Read an OWL file by using SANSA and retrieve a Spark RDD/DataSet representation of it.
   d. Use SANSA-Inference layer in order to apply some inference/reasoning over RDF file by applying RDFS profile reasoner.

AT HOME

---

1. Read and explore
   a. Spark Machine Learning Library (MLlib) Guide
   b. SANSA Overview and SANSA FAQ.
2. Further readings
   a. MLlib: Machine Learning in Apache Spark