Project : From knowledge extraction to knowledge representation

In the previous tutorials, we have seen different tools and techniques that allow :

Represent knowledge and enrich it using RDF, RDFS and OWL
Extract knowledge from text using Naîve Bayes and SVM

In this project, the objective is to build a **simple end to end pipeline** in order to extract knowledge from a set of documents, represent this knowledge using rdf, query the created rdf graph and eventually use reasoning to enrich the extracted knowledge.

Before starting the implementation of your pipeline. You should collect 30 news articles about different topics from different websites.

In order to build the whole pipeline, you will need different subparts :

1. Collect 30 news articles and save them in a forlder, you can name them using this pattern : ”id_title_class”.txt
2. A program that trains a model to classify news document in order to predict the main topic of it
3. A program that takes as input a document (or a list of documents contained in a directory) and output in a csv file the document class (predicted by the model), and eventually the date creation, the title, the authors or any other information that you can extract
4. A program that takes as input the document, the document ID and the label predicted by the learned model. This program should output an RDF file.
5. Optional :
    a. You can do some topic modelling to extract more knowledge about the documents
    b. You can also perform Named Entity Recognition using some pretrained models in NLTK or Spacy
6. A program that contains generic sparql queries that will be executed on the rdf file once loaded in a Jena model.
7. A program that will take an ontology modeling some class hierarchies (example : basket is subclassof sport) about the labels you have, the rdf file and will perform reasoning in order to enrich the extracted knowledge.
8. Optional : You can run some queries over the dbpedia (or any other Linked Open Dataset) to get more knowledge about the topics you extract from document.

Tip : If you want to execute your python and java programs one after another, you can write a bash script that will call them from the terminal.

Important information :

One Project = One student

I am expecting several outputs from you :

- The project report (5 to 10 pages)
- The collected dataset (a directory of several .txt files)
- The project's code source. if you use multiple languages (Python and Java for example), you can write a bash file to execute your pipeline.

Report :
The following is a suggested structure for the report
Author's name
Project's name
Abstract: It should not be more than 400 words.
Introduction: This section introduces your overall approach to the problem.
Approach: This section details your approach to the problem. For example, this is the section where you would describe the architecture of your programs, UML's diagrams... You should be specific.
Experiment and pipeline: In this section, you have to describe:

The dataset(s) you used
The machine learning you used and the different evaluation metrics you obtained (with explanation)
The sparql queries you wrote and eventually used ontology.
Your results you obtained and the difficulties you faced (show example results, and explain things you found hard)

Conclusion: What have you learned? What would you have done if you had more time ? Any suggestions ?