Tutorial 3 : Text Mining for knowledge extraction

We have seen in the previous tutorial how to model knowledge in ontology in order to automate its processing. Based on RDF, RDFS and OWL standards, we have used the Jena framework to represent various type of knowledge. We have also tested some of SPARQL's capacities in order to query this knowledge and enrich it.

In this tutorial, we will continue exploring the different processes leading to an **end to end Knowledge Discovery System**. But this time, we will see some of the techniques that can be used in order to extract useful information from text.

The main objectives of this tutorial are :

- Build a classification model over a text dataset
- Optimize the built model in order to get better accuracy
- Use the output of the model and transform it into RDF triples

We will use the 20 newsgroups dataset to illustrate the different steps of knowledge extraction from text. The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon a messages posted before and after a specific date.

Part 1 : Text Classification

1.1 Environment setting :

We first begin by setting up the environment.

- Check to see if your Python installation has pip. Enter the following in your terminal:
  `pip -h`
  If pip is not installed : https://pip.pypa.io/en/latest/installing/

- Install The virtualenv with pip:
  `pip install virtualenv`

- Create a virtual environment by specifying its path.
  - For example to create one in the local directory called 'mypython', type the following: `virtualenv -p python3 mypython`

- Activate your virtual env
    - source mypython/bin/activate
- Download the requirements file from moodle and Install the packages using this command
    - pip install -r path_to_tutorial_directory/requirements.txt
- Create a new kernel for jupyter notebook :
    - ipython kernel install --user --name=pyml
- Launch your notebook using jupyter notebook or jupyter lab :
    - jupyter notebook / jupyter lab
- Once started, select the 'pyml' kernel you just created

1.2 Jupyter Notebook basics :

http://nbviewer.jupyter.org/github/jupyter/notebook/blob/master/docs/source/examples/Notebook/Notebook%20Basics.ipynb#Notebook-Basics

- **Exercise :**Before you start processing the dataset and learning the models, be sure to follow the above tutorial !

For the rest of this first part, open the notebook available on moodle and follow the written instructions.

Part 3 : Representation of the extracted knowledge

After building and optimizing the model. You can now search for some news articles of your choice and try to classify them using your model.
In preparation for the final project, try to build a program (using Jena) that takes as input the predictions and the document IDs or titles and construct rdf statements that give those documents URIs and add to them the predicted class
Example :

myns:Docuemt1 ---- myns:is_about ---> myns:Sport