TikaAndNLTK

Natural Language Toolkit (NLTK) with Tika

NLTK is a python library and in order to integrate and use its capabilities with Tika one must install the server endpoint created for NTLK using Flask to extract Named Entities from text passed to it.

- Natural Language Toolkit (NLTK) with Tika
  - Installation
  - Start NLTK Server
  - Preparing resources for NLTK in Tika-App
  - Running NLTK with Tika

Installation

1. **Simple with pip**

```
$ pip install --process-dependency-links nltkrest
```

2. **Setuptools and/or Distribute**: The module can be downloaded from github and then installed with the following commands:

```
$ cd NLTKRest/nltkrest
$ python setup.py install nltkrest
```

After installation you will have a command called *nltk-server*. By default it starts a REST server on port 8881. You can change the port by typing *--port* or *-p*. You can also turn on verbose mode by typing *-v*.

Start NLTK Server

```
$ nltk-server -v --port 8888
```

You should see a message:
Running on http://127.0.0.1:8888 The server is up and ready!

Preparing resources for NLTK in Tika-App

You can either perform steps 1 & 2 together or just 3.

1. **Activate Named Entity Parser** In order to use any of the NamedEntityParser implementations in Tika , the parser responsible for handling the name recognition task needs to be enabled. This can be done with Tika Config XML file, as follows

```
<?xml version="1.0" encoding="UTF-8"?>
<properties>
    <parsers>
        <parser class="org.apache.tika.parser.ner.NamedEntityParser">
            <mime>text/plain</mime>
            <mime>text/html</mime>
            <mime>application/xhtml+xml</mime>
        </parser>
    </parsers>
</properties>
```

This configuration has to be supplied in the later phases, so store it as 'tika-config.xml'.

2. **Supply NLTKServer.properties file**
It is imperative that Tika should know on what host you are running the *nltk-server*. By default Tika will assume your server runs on port 8881. In order to specify any other port, you must supply a NLTKServer.properties file. Sample NLTKServer.properties file. My file looks like the following:

```
nltk.server.url=http://localhost:8881
```

In an nutshell

```
#Create a directory for keeping the config and properties file.
export NLTK_RES=$HOME/NLTKRest-resources
mkdir -p $NLTK_RES
cd $NLTK_RES
#config file must be stored in this directory
pwd

export PATH_PREFIX="$NLTK_RES/org/apache/tika/parser/ner/nltk"
mkdir -p $PATH_PREFIX
#create and edit the properties file
vim $PATH_PREFIX/NLTKServer.properties
```

3. **Download NLTKRest-resources**
Better yet, you could skip the previous two steps completely and save the hassle. Simply download the project NLTKRest-resources and edit the properties file

```
cd $HOME && git clone https://github.com/manalishah/NLTKRest-resources
export NLTK_RES=$HOME/NLTKRest-resources
vim $NLTK_RES/org/apache/tika/parser/ner/nltk/NLTKServer.properties
```

Running NLTK with Tika

Finally, we've reached this point where we can smile and let Tika do the working!

```
export TIKA_APP={your/path/to/tika-app}/target/tika-app-1.13-SNAPSHOT.jar

#set the system property to use NLTKNERecogniser class
java -Dner.impl.class=org.apache.tika.parser.ner.nltk.NLTKNERecogniser -classpath $NLTK_RES:$TIKA_APP org.
apache.tika.cli.TikaCLI --config=$NLTK_RES/tika-config.xml -m  http://www.hawking.org.uk/
```

This will output metadata along with named entities extracted using nltk:

NER_NAMES: Gonville
NER_NAMES: Einstein
NER_NAMES: Briefer History
NER_NAMES: ALS
NER_NAMES: Unbreakable Code
NER_NAMES: Click
NER_NAMES: Cambridge
NER_NAMES: George
NER_NAMES: Lucasian Professor
NER_NAMES: Stephen Hawking
NER_NAMES: Latest
NER_NAMES: Hubble
NER_NAMES: 1979
NER_NAMES: Brief History
NER_NAMES: California
NER_NAMES: University
NER_NAMES: 1663
NER_NAMES: 1982
NER_NAMES: London
NER_NAMES: US National Academy
NER_NAMES: Baby Universe
NER_NAMES: Home About Stephen The Computer Stephen
NER_NAMES: Applied Mathematics
NER_NAMES: Santa Cruz
NER_NAMES: Leiden University
NER_NAMES: CBE
NER_NAMES: Science
NER_NAMES: Caius College
NER_NAMES: HUDF09 Team
NER_NAMES: Dennis Stanton Avery
NER_NAMES: 2009
NER_NAMES: ESA
NER_NAMES: Annie
NER_NAMES: NASA
NER_NAMES: Black Holes
NER_NAMES: Universe
NER_NAMES: Sally Tsui
NER_NAMES: eXtreme Deep Field
NER_NAMES: Centre
NER_NAMES: Royal Society
NER_NAMES: Weebly
NER_NAMES: Cambridge Lectures Publications Books Images Films Videos Stephen
NER_NAMES: 1963
NER_NAMES: Unbreakable Code Hot
NER_NAMES: Isaac Newton
NER_NAMES: Lucy
NER_NAMES: Stephen