Conversion from UTF-8 Text file to tokenized JSON for CATMA import (CATMA-X)

CATMA-X is an advanced annotation tool which can convert plain utf-8 text files to annotated JSON files for importing to CATMA application for further analysis. The tool is responsible for creating tokens and start-end indices. Differentiation between direct and indirect speech is possible through this tool. Color coding and automated text annotation can be done through this tool with state of the art NLP technologies. Through trained models in german and english, the tool can analyse texts for annotation.

Some key features of this tool are:

- Direct and Indirect speech annotation
- Tokenize sentence start and end indices
- Import output JSON files to CATMA

All relevant code files are available in GitHub Public repository Page [Link:][df1]

Technology

CATMA-X uses the following technology stack to work properly:

- Python - Interpreted, high-level and general-purpose programming language!
- JSON - JSON (JavaScript Object Notation) is a lightweight data-interchange format.
- spaCy - An Open-source software library for advanced natural language processing (NLP), written in the programming languages Python and Cython.

Installation (Windows)

CATMA-X requires Python virtual environment to run. Install the dependencies to start the workflow. Spacy can be installed using pip (python package manager). You can use a virtual environment to avoid depending on system-wide packages.

```
create a new folder Scriptfolder and navigate to that directory from your desktop
using CMD window
$ cd C:\Users\XYZ\Desktop\Scriptfolder
if python virtualenv in not already installed, execute the following code
$ pip install virtualenv
create a new virtual env, activate it and install spacy
$ python3 -m venv venv
$ cd venv
$ .\Scripts\activate
$ pip install spacy
```

Download Spacy models and data

```
For English trained models:
$ python -m spacy download en_core_web_sm
For Deutsch trained models:
$ python -m spacy download de_core_news_sm
```

if you want to use a spacy model, import spacy and load the model (open same cmd window and type)

```
$ python
$ >> import spacy
$ nlp = spacy.load(‘de_core_news_sm’)
```

to deactivate python, type: exit() to deactivate venv, type: deactivate

Workflow

Step 1 - convert the “raw” text file to json format: Navigate to your virtualenv (venv) and activate the virtualenvironment

```
$ cd venv
$ .\Scripts\activate
```

Navigate back to root folder and start python

```
$ cd Scriptfolder
$ python
```

Tokenize your text file (see code from tokenize.py in Github) **note**: change the filename according to your file an make sure the file is in your root directory **note**: you can also change the filename of the output file

```
import spacy
import json
import sys
nlp = spacy.load('de_core_news_sm')
file_name = 'sandmann.txt'
file_text = open(file_name).read()
file_doc = nlp(file_text)
for token in file_doc:
    print ('Processing tokens.....')
    f = open('output_tokens.txt', 'a')
    print (token, token.idx, file=f)
    f.close()
...
```

press enter twice

Step 2 - Convert the tokenized output file to json using txt2Json.py (download the file from Github and save it in your root folder) open a new command prompt (without venv) or exit python and deactivate venv, then navigate in the root folder and run the txt2Json.py script by calling python txt2Json.py 'filename':

```
$ cd Scriptfolder
$ python txt2Json.py 'output_file_name'
```

- the script creates a json output file ('Output_tokens.json')
- **note**: when running the script multiple times, the output file will be replaced if you do not change the output file name.

Step 3: merge the annotations (tagged file) with the json file

Step 4: Implementation in Catma