Document Classification

Modules

Document classification task is base on the three modules as follow

Scrapping

Scrapping module (scraper.py) scrap the documents from Wikipedia. It iterates over the link provided by csv file and extract main content of that URL. Lastly it saves the data in word format in data/category folder.

Training

Training module (training.py) get the documents from data folder and train 3 machine learning models. It also saves model for future prediction.

Testing

Testing module (testing.py) take the document in pdf/word format and makes prediction on the all-trained model. Lastly, it shows the prediction results on console.

Command Line Instructions

All the modules can be run using command line in follow the below instructions.

Training with Existing classes

1. Install the requirement file in your virtual environment

   ```
   pip install -r requirments.txt
   ```

2. After installing all the requirements, start the training by the following command
   ```
   python3 training.py
   ```

Training with New Classes

For the training of the models with new classes, the documents of the new classes will be required by scrapping module.

1. To Run the scrapping module, prepare a CSV file with two columns URL and Name (case sensitive) that contain the URL and its appropriate name respectively. (**Hint:** CSV files in data folder.)
2. Place the newly prepared CSV file in data folder and make a folder with name of your documents type.
3. Run the command to scrap data of newly added category
   ```
   python3 scraper.py --csv <filename> --category <folder-name>
   ```
4. Repeat all the above steps for each new class.
5. After this, make sure the folders name with new classes are present in data folder and contains the related documents.
6. Run the training module by the following command
   ```
   python3 training.py --classes <”class1” “class2” “class3” …. “Class n”]> --pdf True
   ```

Hint: class name will be the exact folder name in data folder and in double quotes. And change –pdf True if pdf not available other wise False.

Testing New Document

1. For the testing of the document, run the following command if the document is located in project directory
   ```
   python3 testing.py --path <relative-document-path>.
   ```
2. If the document outside the project directory, then run the following command.
3. python3 testing.py --abs True --path <relative-document-path>.

Good Luck!