Quick Start Guide

Christian-Albrechts-Universität zu Kiel

Quick Start:

1. Prerequisite: Download and install Java 8+.
2. Download the implementation from
   https://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction
3. Unzip the archive into a folder of your choice.
   1. For Tesseract OCR: Download the language data (e.g., eng.trainedata) and place it into a folder called „tessdata“ next to the JAR (Windows) or install it as a package using *sudo apt-get install tesseract-ocr* (Unix).
   2. For Ocropy OCR (Unix only): Follow Instructions on the next slide.
4. Create or download a XML file that specifies the pipeline structure (see Slide 3).
5. Execute the pipeline using the command:
   java –jar TX.jar <Your-XML-File>

Quick Start Guide

Christian-Albrechts-Universität zu Kiel

Installation of Ocropy (Unix only):

1. Download and install Git and Python (if not already installed).
2. Download and install Ocropy using the following commands:

mkdir workingcopy/

cd workingcopy/

git init .

git pull https://github.com/tmbdev/ocropy

sudo apt-get install $(cat PACKAGES)

wget -nd http://www.tmbdev.net/en-default.pyrnn.gz

mv en-default.pyrnn.gz models/

sudo python setup.py install

Quick Start Guide

Christian-Albrechts-Universität zu Kiel

XML File Creation:

Either download the provided XML files from

https://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction

and use/modify them or Create a new XML-File:

1. Create an empty XML-File and add a <container></container> tag.
2. Define an input stream inside the container-tag.
3. Define a processor that processes the stream.
4. List all relevant methods inside the processor in their sequential order.
5. Specify the input and output for each method.

An example XML-file is on the next slide.

Quick Start Guide

Christian-Albrechts-Universität zu Kiel

Example XML File:

```
<container>
	<stream id="data" class="de.cau.cs.kd.source.GUIFileChooser" />
	<process input="data">
		<de.cau.cs.kd.processor.open.OpenImage _fileKey="file" imageKey_="image" />
		<de.cau.cs.kd.processor.view.ImageView _imageKey="image" />
		<de.cau.cs.kd.processor.image.grey.ATSCLuminanceConversion _imageKey="image" imageKey_="imageGrey" />
		<de.cau.cs.kd.processor.image.binary.OtsuHierarchical _imageKey="imageGrey" imageKey_="imageBinary" _edgeBinarizationThreshold="32" />
		<de.cau.cs.kd.processor.image.label.ConnectedComponentAnalysis _imageKey="imageBinary" regionsKey_="regions" />
		<de.cau.cs.kd.processor.features.FeaturesFromCC _regionsKey="regions" featuresKey_="features"/>
		<de.cau.cs.kd.processor.filter.Dots _size="8" _featuresKey="features" _regionsKey="regions" />
		<de.cau.cs.kd.processor.filter.Size _relativeMax="0.6" _featuresKey="features" _regionsKey="regions" />
		<de.cau.cs.kd.processor.filter.Area _upperratio="0.8" _featuresKey="features" _regionsKey="regions" />
		<de.cau.cs.kd.processor.view.ImageView _imageKey="imageBinary" _featuresKey="features" _regionsKey="regions" />
		<de.cau.cs.kd.processor.clustering.DBSCAN _minPts="1" _adaptive="2,3" _separateNoise="true" _featuresKey="features" clusterKey_="cluster" />
		<de.cau.cs.kd.processor.clustering.OMST _angle="60,60" _featuresKey="features" _clusterKey="cluster" clusterKey_="subcluster" />
		<de.cau.cs.kd.processor.features.ComputeAngle  _method="hough" _clusterKey="subcluster" _featuresKey="features" _regionsKey="regions" anglesKey_="angles"  />

		<de.cau.cs.kd.processor.image.ImagesFromCluster itemsKey_="images" imageKey_="subimage" _regionsKey="regions" _imageKey="image" _clusterKey="subcluster" _anglesKey="angles" />
		<stream.flow.ForEach key="images">
			<de.cau.cs.kd.processor.ocr.Tesseract _imageKey="subimage" textKey_="ocredtext" _PSMmode="PSM_SINGLE_BLOCK" _language="eng" />
			<de.cau.cs.kd.processor.text.NormalizeWhitespaces _textKey="ocredtext" textKey_="cleanedtext" />
		</stream.flow.ForEach>
		<de.cau.cs.kd.processor.Condense _itemsKey="images" itemKey_="OCR" _mergeKey="cleanedtext,x,y,width,height" _append="true" />
		<de.cau.cs.kd.sink.CreateCSV _fileKey="file" _incremental="true" _itemsKey="x,y,width,height,cleanedtext" _separator="\t"/>
		<stream.flow.Delay time="3000ms"/>
	</process>
</container>
```