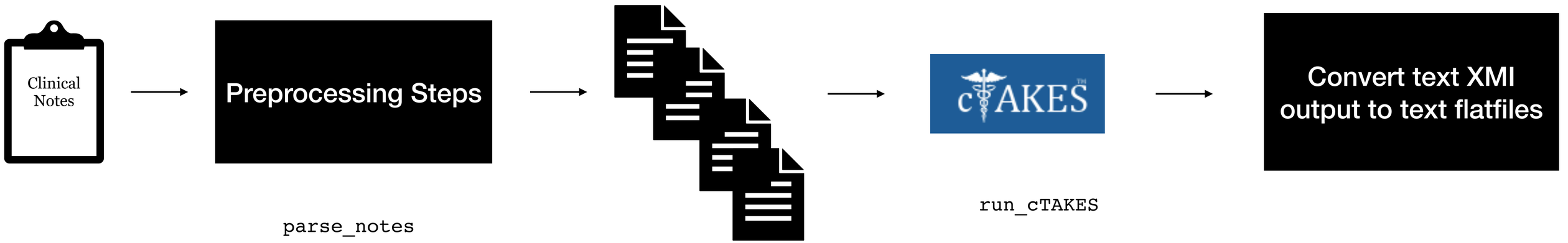
cTAKES Workflow

File Processing, CUI output, and flatfile generation

Overview

- Bird’s-eye view of workflow
- Preprocessing
- Running cTAKES
- Post-processing cTAKES output
- Input and Output side-by-side (interactive)

Bird’s-eye view of workflow



`parse_notes`

`sort_by_size`

`get_file_size`

`run_cTAKES`

`setup_for_flatfile_generation`

`generate_flatfiles`

Preprocessing

Requirements for input files

- One record per file
- CSV format with a header
- Unique patient ID
- Unique encounter ID
- Timestamp
- Unique document ID (this should be the file name)
- Document type
- Dummy INDEX Column

DOC001.csv

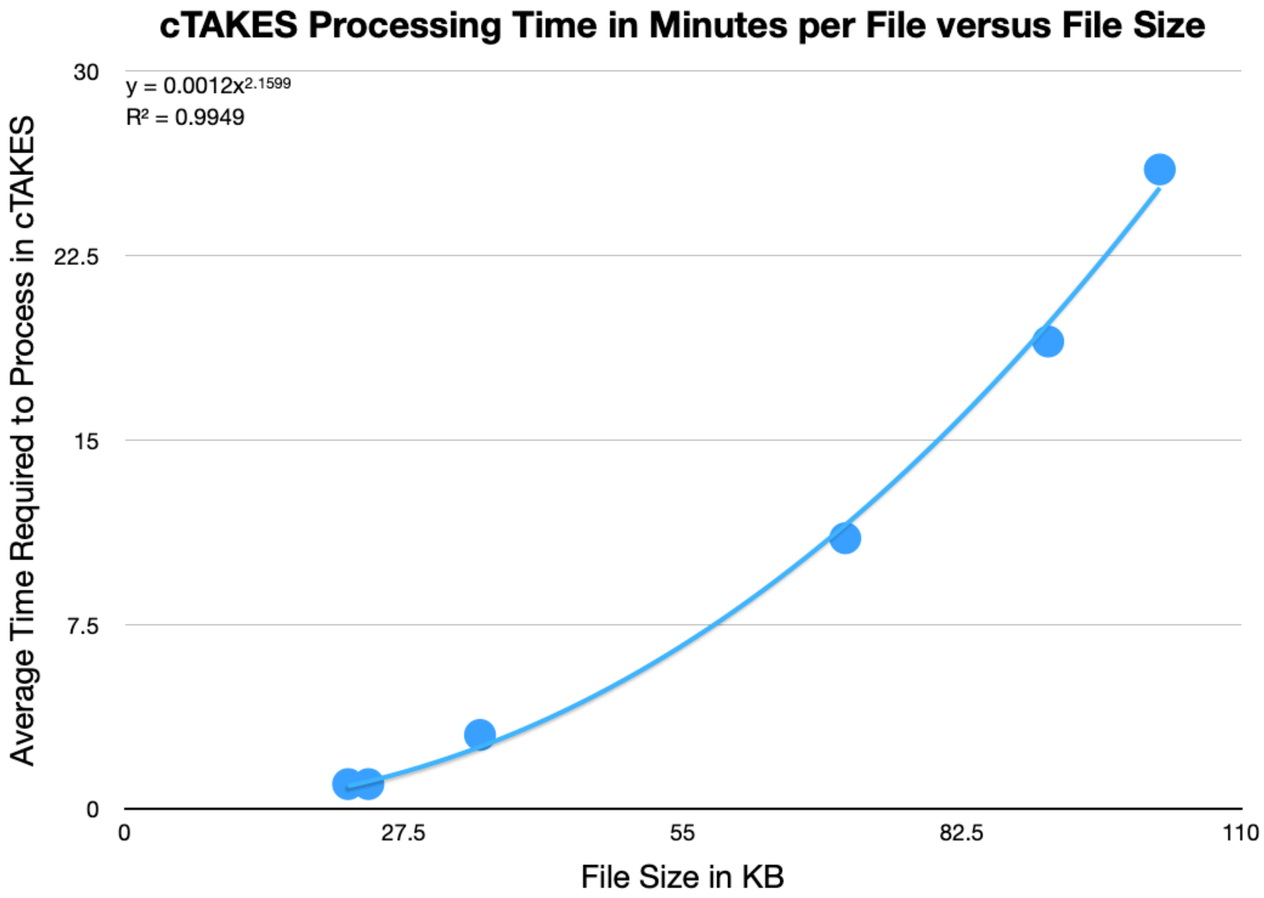
INDEX,DOCUMENT_ID,PATIENT_ID,HADMID,DOCUMENT_TYPE,TIMESTAMP,TEXT

0,DOC001,P0001,E0001,H&P,01-01-2015,“History and Pathology Physician Note: 35 yo Male was admitted to the ER with shortness of breath, chest pain, and numbness in left arm. Patient does not smoke and denied past cardiac history.”

Preprocessing

Additional Requirements

- Sort by file size



Preprocessing

Code Requirements

1. `parse_notes` -> provide input filename with concatenated notes, and term to split on

2. `getFileSizes` -> output directory from previous step and an output file name

3. `sortBySize` -> output file from previous step and location of directory with files from step 2

```
# MODIFY ME
# provide an input filename that contains clinical notes
NOTES_FILE=''
# provide a term that can be used to split the data into manageable groups
# for example, you could split the data by encounter ID
SPLIT_ID=''
# end MODIFY ME
```

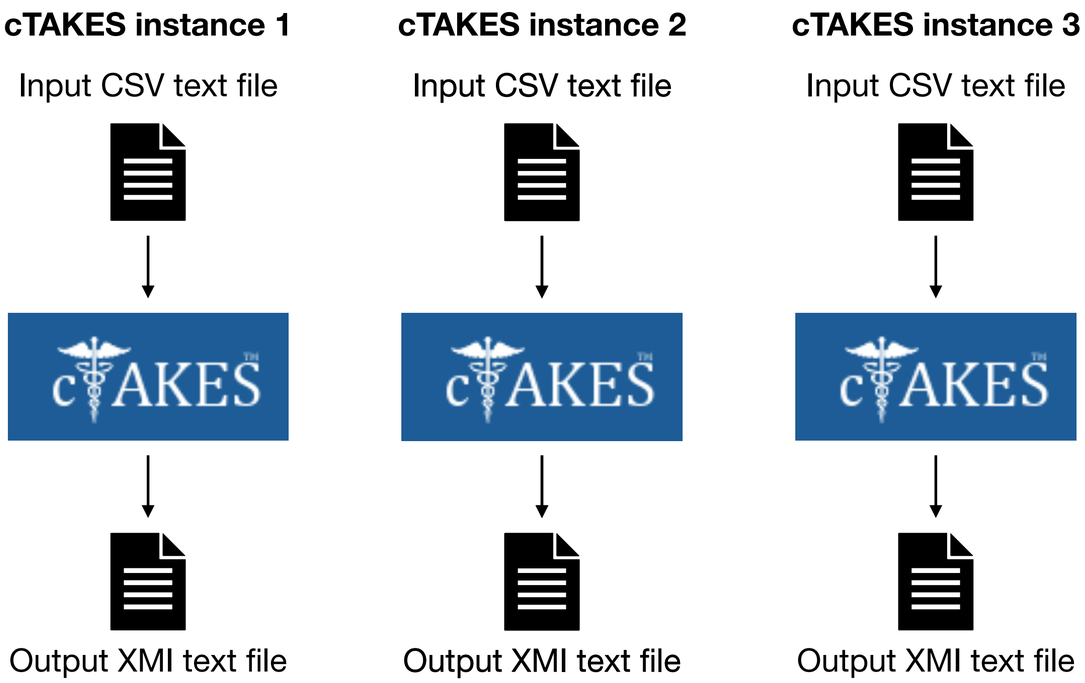
```
parser = argparse.ArgumentParser()
parser.add_argument('--dir', '-d', help='directory name to scan files', required=True)
parser.add_argument('--out', '-o', help='output file name', required=True)
args = parser.parse_args()
```

```
parser = argparse.ArgumentParser()
parser.add_argument('--infile', '-i', help='fileSizeList file name', required=True)
parser.add_argument('--dir', '-d', help='directory where files are located', required=True)
args = parser.parse_args()
```

Running cTAKES

Code Requirements

- Install cTAKES
- Create UMLS Account
- Modify XML properties file to include username and API key
- Customize Piper file
- Customize Bash command/script
- Multiple parallel instances



Post-Processing cTAKES Output

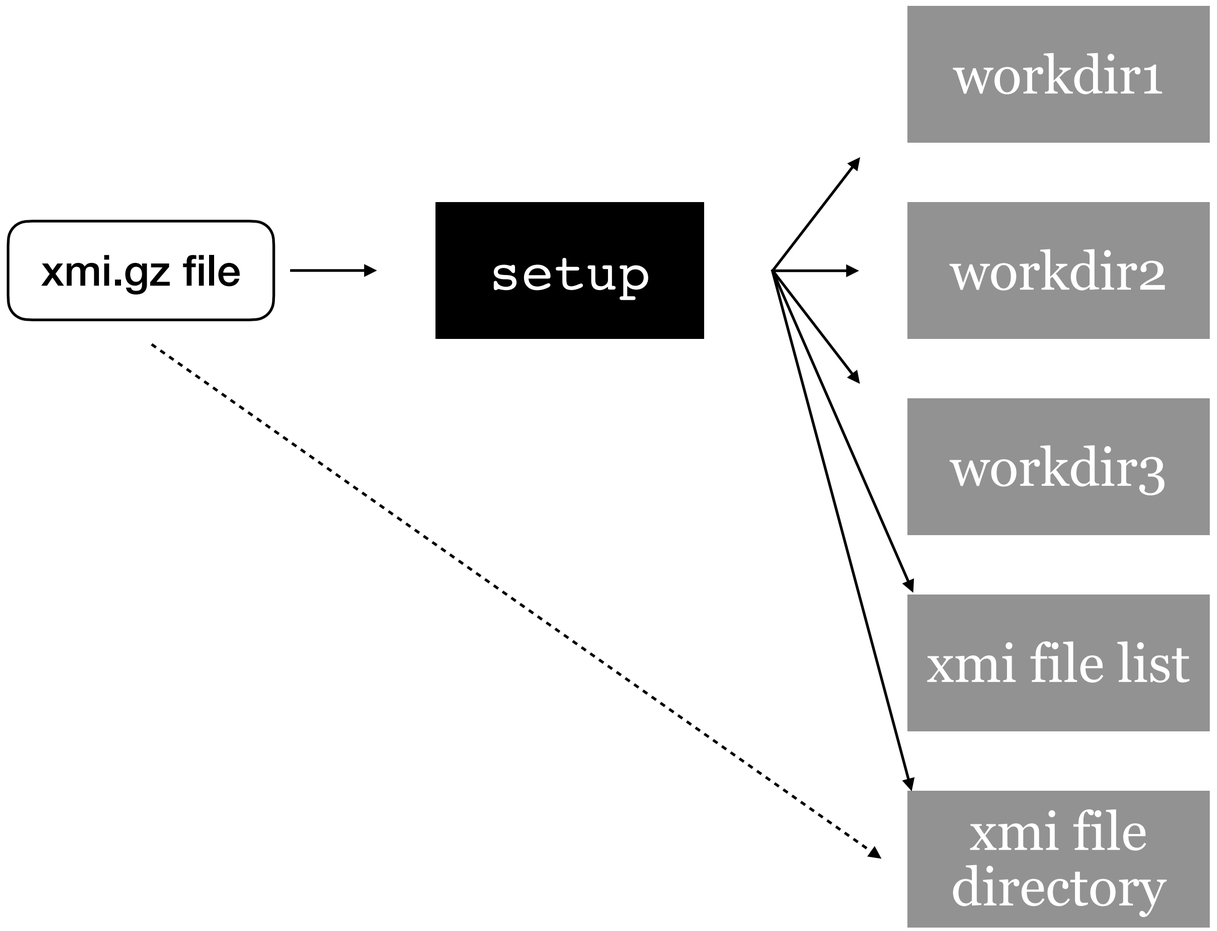
Overview

- (Default) Output from cTAKES is an XMI file, which is a variant of XML
- XMI files are **BIG** and must be gz-compressed!
- `setup_for_flatfile_generation` -> runs setup for up to 3 parallel instances of flatfile_generator
- `generate_flatfiles` -> generates flatfiles from compressed XMI output


```
<?xml version="1.0" encoding="UTF-8"?>
<xmi:XMI xmlns:util="http:///org/apache/ctakes/typesystem/type/util.ecore" xmlns:tcas="http:///uima/tc
as.ecore" xmlns:xmi="http://www.omg.org/XMI" xmlns:cas="http:///uima/cas.ecore" xmlns:type10="http:///
org/cleartk/syntax/constituent/type.ecore" xmlns:ne="http:///org/cleartk/type/ne.ecore" xmlns:textsem=
"http:///org/apache/ctakes/typesystem/type/textsem.ecore" xmlns:types2="http:///org/apache/ctakes/asse
rtion/zoner/types.ecore" xmlns:type6="http:///org/apache/ctakes/smokingstatus/i2b2/type.ecore" xmlns:r
efsem="http:///org/apache/ctakes/typesystem/type/refsem.ecore" xmlns:type11="http:///org/cleartk/synta
x/dependency/type.ecore" xmlns:type="http:///de/tudarmstadt/ukp/dkpro/core/api/metadata/type.ecore" xm
lns:type14="http:///org/cleartk/util/type.ecore" xmlns:assertion="http:///org/apache/ctakes/typesystem
/type/temporary/assertion.ecore" xmlns:type8="http:///org/cleartk/score/type.ecore" xmlns:syntax="http
:///org/apache/ctakes/typesystem/type/syntax.ecore" xmlns:type9="http:///org/cleartk/srl/type.ecore" x
mlns:type2="http:///org/apache/ctakes/constituency/parser/uima/type.ecore" xmlns:types="http:///org/ap
ache/ctakes/assertion/medfacts/types.ecore" xmlns:type7="http:///org/apache/ctakes/smokingstatus/type.
ecore" xmlns:relation="http:///org/apache/ctakes/typesystem/type/relation.ecore" xmlns:type12="http://
/org/cleartk/timeml/type.ecore" xmlns:type5="http:///org/apache/ctakes/sideeffect/type.ecore" xmlns:ty
pe15="http:///org/cleartk/ne/type.ecore" xmlns:type13="http:///org/cleartk/token/type.ecore" xmlns:str
uctured="http:///org/apache/ctakes/typesystem/type/structured.ecore" xmlns:textspan="http:///org/apach
e/ctakes/typesystem/type/textspan.ecore" xmlns:libsvm="http:///org/apache/ctakes/smokingstatus/type/li
bsvm.ecore" xmlns:type3="http:///org/apache/ctakes/coreference/type.ecore" xmlns:type4="http:///org/ap
ache/ctakes/drugner/type.ecore" xmi:version="2.0">
<cas:NULL xmi:id="0"/>
<tcas:DocumentAnnotation xmi:id="8" sofa="1" begin="0" end="535" language="x-unspecified"/>
<structured:DocumentID xmi:id="13" documentID="DOC001"/>
<structured:DocumentIdPrefix xmi:id="15" documentIdPrefix=""/>
<structured:DocumentPath xmi:id="17" documentPath="/LAB_SHARED/DATA/UW-Madison/Analytic/XMI_workdir/te
mplate_ctakes_4.0.0.1_instance/inputDir/DOC001.csv"/>
<textspan:Segment xmi:id="19" sofa="1" begin="0" end="535" id="SIMPLE_SEGMENT"/>
<textspan:Sentence xmi:id="25" sofa="1" begin="0" end="59" sentenceNumber="0"/>
<textspan:Sentence xmi:id="31" sofa="1" begin="61" end="144" sentenceNumber="1"/>
<textspan:Sentence xmi:id="37" sofa="1" begin="146" end="200" sentenceNumber="2"/>
<textspan:Sentence xmi:id="43" sofa="1" begin="202" end="274" sentenceNumber="3"/>
<textspan:Sentence xmi:id="49" sofa="1" begin="278" end="427" sentenceNumber="4"/>
<textspan:Sentence xmi:id="55" sofa="1" begin="431" end="533" sentenceNumber="5"/>
<syntax:NewlineToken xmi:id="61" sofa="1" begin="59" end="61" tokenNumber="19"/>
<syntax:NewlineToken xmi:id="69" sofa="1" begin="274" end="276" tokenNumber="73"/>
```

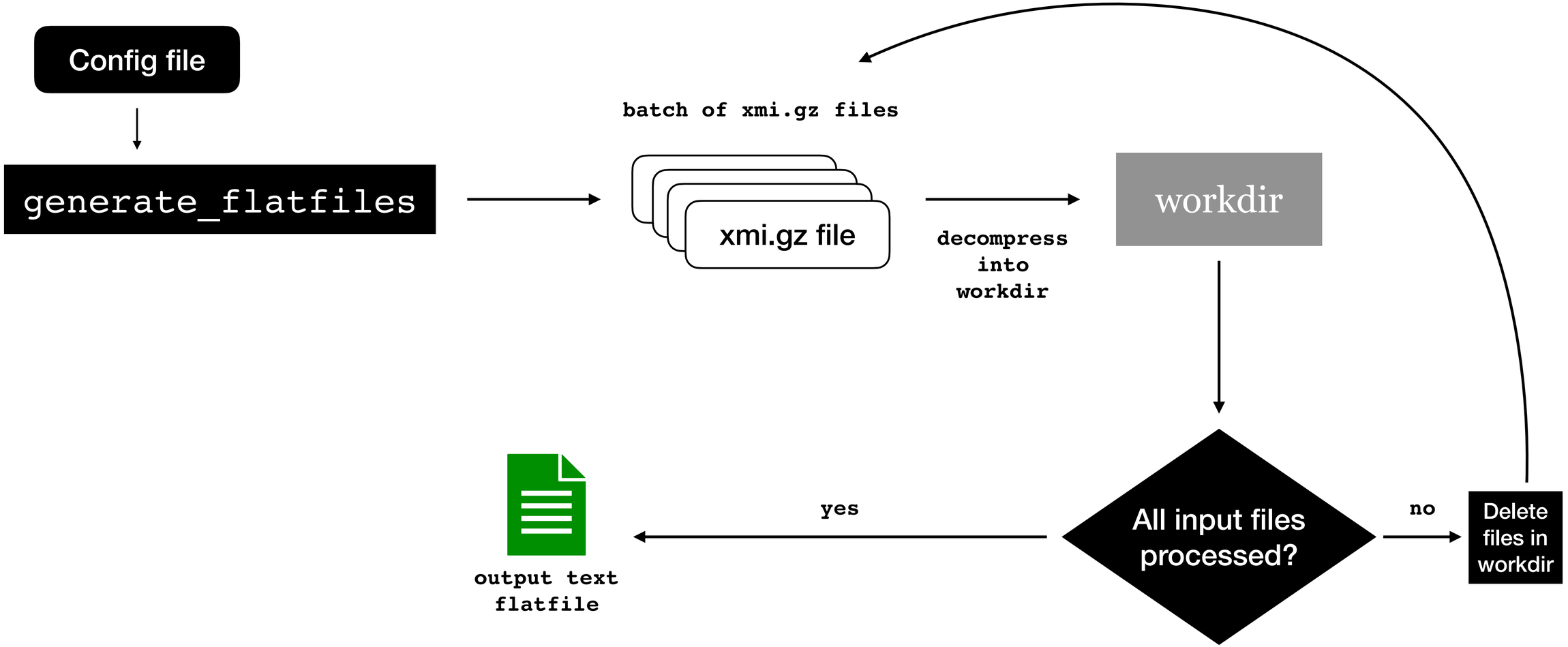
Post-Processing cTAKES Output

Overview



Post-Processing cTAKES Output

Overview



Post-Processing cTAKES Output

Code Requirements

- Generate config file for `generate_flatfiles`
  - Comma-separated file with the shown labels and indices
- Specify number of parallel instances to run, input/output labels, and name of config file

DOC001.csv

INDEX,DOCUMENT_ID,PATIENT_ID,HADMID,DOCUMENT_TYPE,TIMESTAMP,TEXT

0,DOC001,P0001,E0001,H&P,01-01-2015,“History and Pathology Physician Note: 35 yo Male was admitted to the ER with shortness of breath, chest pain, and numbness in left arm. Patient does not smoke and denied past cardiac history…”

config.txt

FileName,6
DocType,9
PatID,7
EncID,8
TS,10

Interactive Example

Visualized Example

- Input CSV
- Output XMI
- Output Flatfile