CTAKESParser

Tika now has the ability to leverage Apache cTAKES for use in parsing biomedical information from text. This page documents how to get Tika working with cTAKES.

Installing cTAKES

The first step to getting the parser up and running is installing Apache cTAKES. Read on the following should work well on *nix systems. Windows directions are TODO. It's very important to install at least cTAKES version 3.2.2 or later.

1. `mkdir -p $HOME/src && cd $HOME/src`
2. `curl -O http://mirrors.ibiblio.org/apache//ctakes/ctakes-3.2.2/apache-ctakes-3.2.2-bin.tar.gz`
3. `tar xvzf *.tar.gz`
4. `export CTAKES_HOME=$HOME/src/apache-ctakes-3.2.2`

Now you have to download a separate resources package for cTAKES:

1. `cd $HOME/src`
2. `curl -Lo ctakes-resources-3.2.1.1-bin.zip "http://downloads.sourceforge.net/project/ctakesresources/ctakes-resources-3.2.1.1-bin.zip?r=http%3A%2F%2Fsourceforge.net%2Fprojects%2Fctakesresources%2F%3Fsource%3Dtyp_redirect&ts=1433609725&use_mirror=softlayer-dal"`
3. `mv *.zip apache-ctakes-3.2.2`
4. `cd apache-ctakes-3.2.2` 5. `unzip ctakes-resources-3.2.1.1-bin.zip`

After the above is done, cTAKES is installed.

Signing up for a UMLS account

To use cTAKES and the cTAKES Tika Parser you need a Unified Medical Language System (UMLS) account. You can sign up for one here. It can take up to 3 business days to get an account so be patient. Once your account is approved you can use the cTAKESParser and read on. Future improvements are to provide a means to include the offline UMLS dictionary.

Prepare your CTAKES configuration properties file

The cTAKESParser requires a configuration properties file. You can find an example here originally from TIKA-1645 and adapted and maintained in Github now in ctakesparser-utils.

Edit it as follows.

```
aeDescriptorPath=/ctakes-clinical-pipeline/desc/analysis_engine/AggregatePlaintextUMLSProcessor.xml
text=false
annotationProps=BEGIN,END,ONTOLOGY_CONCEPT_ARR
separatorChar=:
metadata=Study Title,Study Description
UMLSUser=your_username
UMLPass=your_password
```

NB the AggregatePlaintextUMLSProcessor may not extract all the medication mentions: if so, try the AggregatePlaintextFastUMLSProcessor

e.g., aeDescriptorPath=/ctakes-clinical-pipeline/desc/analysis_engine/AggregatePlaintextFastUMLSProcessor.xml

Be advised that if you do implement this change within a session you may find that the database gets locked temporarily. Clear the *.lck files, or just reboot the computer.

Analysis is performed on the extracted text and/or metadata from the AutoDetectParser. cTAKESParser decorates AutoDetectParser, and then takes the extracted metadata and/or text (or both) and then adds ctakes: prefixed metadata for procedure, medication, disease and other extracted information. To use the cTAKESParser, update the metadata property to be a comma separated list of metadata fields to search for medical terminology in. Then, if you would like the parser to also search the extracted text from Tika, set `text=true`.

The `annotationProps` is a comma separated list of what cTAKES properties to extract, and `separatorChar` is what to use to separate them in the extracted field. So, we are telling cTAKES to extract the begin and end of the found text (BEGIN,END), first, and then extract the Ontology concept array (ONTOLOGY_CONCEPT_ARR). UMLS uses identifiers for each term, that can be used to then search UMLS for more information about that term - this array includes the UMLS pointer to the term, and any also identified similar terms. An example of the extracted annotationProps would be:

```
mantle cell lymphoma:40592:40612:C0334634,C0334634,C0334634,C0334634
```

In this example, a cTAKES DiseaseDisorderMention of `mantle cell lymphoma` is identified, and then its associated annotation props (the text begins at position 40592 and ends at position 40612 (could be used for highlighting), and then an associated array of medical ontology concept identifiers are provided, i.e., C0334634,C0334634,C0334634,C0334634.

You will need to place the CTAKESConfig.properties file in a classpath directory, e.g., org/apache/tika/parser/ctakes and include it on the classpath when calling the parser. Follow these steps:

1. `mkdir -p $HOME/src/ctakes-config/org/apache/tika/parser/ctakes && cd $HOME/src/ctakes-config/org/apache/tika/parser/ctakes`
2. `curl -kO "https://raw.githubusercontent.com/chrismattmann/ctakesparser-utils/master/config/org/apache/tika/parser/ctakes/CTAKESConfig.properties"`

Setting up the Tika Config file

You will need a custom Tika configuration file for the parser. You can find one here. The reason is that since cTAKESParser decorates AutoDetectParser, in reality, cTAKESParser can handle *any* kind of file type that it can. But you have to make cTAKESParser intercept the mime types you want it to extract biomedical information from. So if you want Tika and its cTAKESParser to etxract biomedical information from application/pdf files, you will need this custom config and to add application/pdf as a mime that the parser can deal with. The default config provided looks like:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<properties>
  <parsers>
    <parser class="org.apache.tika.parser.ctakes.CTAKESParser">
      <mime>application/x-isatab</mime>
      <parser class="org.apache.tika.parser.DefaultParser"/>
    </parser>
  </parsers>
</properties>
```

This maps the cTAKESParser to IsaTab biomedical files, and then the associated default CTAKESProperties.config file described above searches in Study Title and StudyDescription metadata fields from that. However, you could easily add:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<properties>
  <parsers>
    <parser class="org.apache.tika.parser.ctakes.CTAKESParser">
      <mime>application/x-isatab</mime>
      <mime>application/pdf</mime>
      <parser class="org.apache.tika.parser.DefaultParser"/>
    </parser>
  </parsers>
</properties>
```

To parse PDF files, and then set `text=true` in CTAKESProperties.config to parse PDF files and extract biomedical information.

To download and set up the custom Tika config, do the following.

1. `cd $HOME/src/ctakes-config`
2. `curl -kO "https://raw.githubusercontent.com/chrismattmann/ctakesparser-utils/master/config/tika-config.xml"`

Putting it all together: Tika-App

With all of the above information set and provided, you can call the cTAKESParser in Tika. Below is an example of how to use it on a downloaded PDF file from PubMed. You can find an example PDF file here.

The cTAKES parser can then be invoked from tika-app as follows:

```
java -classpath $HOME/src/ctakes-config:${CTAKES_HOME}/lib/opennlp-tools-1.5.3.jar:${TIKA_HOME}/tika-app/target
/tika-app-X.Y-SNAPSHOT.jar:${CTAKES_HOME}/desc:${CTAKES_HOME}/resources:${CTAKES_HOME}/lib/\* org.apache.tika.
cli.TikaCLI --config=$HOME/src/ctakes-config/tika-config.xml -m Vose-2013-American_Journal_of_Hematology.pdf
```

Which will produce (after much printing and output):

```
Content-Length: 457115
Content-Type: application/pdf
Creation-Date: 2013-11-20T13:24:11Z
Last-Modified: 2013-11-22T14:13:25Z
Last-Save-Date: 2013-11-22T14:13:25Z
WPS-ARTICLEDOI: 10.1002/ajh.23615
WPS-JOURNALDOI: 10.1002/(ISSN)1096-8652
WPS-PROCLEVEL: 2
X-Parsed-By: org.apache.tika.parser.CompositeParser
X-Parsed-By: org.apache.tika.parser.ctakes.CTAKESParser
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.pdf.PDFParser
access_permission:assemble_document: true
access_permission:can_modify: true
access_permission:can_print: true
access_permission:can_print_degraded: true
access_permission:extract_content: true
access_permission:extract_for_accessibility: true
access_permission:fill_in_form: true
access_permission:modify_annotations: true
created: Wed Nov 20 05:24:11 PST 2013
ctakes:AnatomicalSiteMention: Cell:189:193:C0007634,C1269647
ctakes:AnatomicalSiteMention: Media:432:437:C0162867
ctakes:AnatomicalSiteMention: Media:593:598:C0162867
ctakes:AnatomicalSiteMention: cell:967:971:C0007634,C1269647
ctakes:AnatomicalSiteMention: cell:1045:1049:C0007634,C1269647
ctakes:AnatomicalSiteMention: cell:1124:1128:C1269647,C0007634
ctakes:AnatomicalSiteMention: Media:2134:2139:C0162867
ctakes:AnatomicalSiteMention: cell:3716:3720:C1269647,C0007634
ctakes:AnatomicalSiteMention: cell:3839:3843:C1269647,C0007634
ctakes:AnatomicalSiteMention: lymph nodes:3920:3931:C0024204,C1269047
ctakes:AnatomicalSiteMention: spleen:3933:3939:C1278932,C0037993
ctakes:AnatomicalSiteMention: blood:3941:3946:C0005767
ctakes:AnatomicalSiteMention: bone marrow:3952:3963:C0005953,C0005953,C0005953
ctakes:AnatomicalSiteMention: bone:3952:3956:C1266909,C1266908,C0262950,C0262950,C0391978,C0391978
ctakes:AnatomicalSiteMention: lymph node:4093:4103:C1269047,C0024204
ctakes:AnatomicalSiteMention: bone marrow:4105:4116:C0005953,C0005953,C0005953
ctakes:AnatomicalSiteMention: bone:4105:4109:C0391978,C0391978,C0262950,C1266909,C0262950,C1266908
ctakes:AnatomicalSiteMention: arm:17428:17431:C1140618,C1269078,C0446516,C1140618,C1140618,C1269612
ctakes:AnatomicalSiteMention: R:17505:17506:C0035639
ctakes:AnatomicalSiteMention: arm:17520:17523:C1140618,C1269078,C1140618,C1269612,C1140618,C0446516
ctakes:AnatomicalSiteMention: R:17568:17569:C0035639
ctakes:AnatomicalSiteMention: R:17629:17630:C0035639
ctakes:AnatomicalSiteMention: R:17989:17990:C0035639
ctakes:AnatomicalSiteMention: R:18023:18024:C0035639
ctakes:AnatomicalSiteMention: stem cell:18458:18467:C0038250,C0018956,C0038250
ctakes:AnatomicalSiteMention: cell:18463:18467:C0007634,C1269647
ctakes:AnatomicalSiteMention: arm:18561:18564:C1269078,C1140618,C1140618,C1269612,C0446516,C1140618
ctakes:AnatomicalSiteMention: arm:18613:18616:C1140618,C1140618,C0446516,C1269078,C1140618,C1269612
ctakes:AnatomicalSiteMention: arm:18644:18647:C1269078,C1140618,C0446516,C1140618,C1140618,C1269612
ctakes:AnatomicalSiteMention: white blood cells:18942:18959:C0023516,C0023516
ctakes:AnatomicalSiteMention: blood cells:18948:18959:C0005773
ctakes:AnatomicalSiteMention: blood:18948:18953:C0005767
ctakes:AnatomicalSiteMention: cells:18954:18959:C1269647,C0007634
ctakes:AnatomicalSiteMention: WBC:19039:19042:C0023516,C0023516
ctakes:AnatomicalSiteMention: R:19995:19996:C0035639
ctakes:AnatomicalSiteMention: R:20250:20251:C0035639
ctakes:AnatomicalSiteMention: R:21066:21067:C0035639
ctakes:AnatomicalSiteMention: R:21138:21139:C0035639
...
ctakes:AnatomicalSiteMention: stem cell:21200:21209:C0038250,C0018956,C0038250
ctakes:AnatomicalSiteMention: cell:21205:21209:C1269647,C0007634
```

```
ctakes:AnatomicalSiteMention: stem cell:21331:21340:C0038250,C0018956,C0038250
ctakes:DiseaseDisorderMention: MCL:12646:12649:C0334634,C0334634,C0334634,C0334634
ctakes:DiseaseDisorderMention: proliferation:12709:12722:C0334094
ctakes:DiseaseDisorderMention: mens:12922:12926:C0027662,C0027662,C0027662
ctakes:DiseaseDisorderMention: graft-versus host disease:26944:26969:C0018133,C0018133
ctakes:DiseaseDisorderMention: disease:26962:26969:C0012634
ctakes:MedicationMention: lenalidomide:24088:24100:C1144149,C1144149
ctakes:SignSymptomMention: OS:4921:4923:C0232275
ctakes:SignSymptomMention: OS:4972:4974:C0232275
ctakes:SignSymptomMention: OS:4995:4997:C0232275
ctakes:SignSymptomMention: education:40715:40724:C0013658,C0013658,C0013658,C0424927,C0013658,C0013658,C0013658,
C0013658
ctakes:schema: coveredText:start:end:ontologyConceptArr
...
date: 2013-11-22T14:13:25Z
dc:format: application/pdf; version=1.5
dc:title: JW-AJH#130002 1083..1088
dcterms:created: 2013-11-20T13:24:11Z
dcterms:modified: 2013-11-22T14:13:25Z
meta:creation-date: 2013-11-20T13:24:11Z
meta:save-date: 2013-11-22T14:13:25Z
modified: 2013-11-22T14:13:25Z
pdf:PDFVersion: 1.5
pdf:encrypted: false
producer: PDFlib PLOP 2.0.0p6 (SunOS)/Adobe LiveCycle PDFG ES
resourceName: Vose-2013-American_Journal_of_Hematology.pdf
title: JW-AJH#130002 1083..1088
xmp:CreatorTool: Arbortext Advanced Print Publisher 9.0.114/W
xmpTPg:NPages: 7
```

Will this work from Tika Server?

Yes, it will! However it's a little tricky since cTAKES also includes its own version of Apache CXF and the jar version numbers are different than the version Tika Server uses, causing the classpath we generated before to fail (the one for tika-app). To obviate this, follow the below steps:

First, generate a script that will build us a classpath from `$CTAKES_HOME/lib` that doesn't include the cxf jar files. This would look something like this:

```
#!/bin/bash

CLASSPATH=""
for f in $(ls ${CTAKES_HOME}/lib/*.jar); do
    if [[ $f != *"cxf"* ]]; then
        CLASSPATH+=$f
        CLASSPATH+=":"
    fi
done

echo $CLASSPATH
```

Save this script as `gen-server-classpath.sh`. Then, start Tika-server like so:

```
java -classpath ./config:${CTAKES_HOME}/lib/opennlp-tools-1.5.3.jar:${TIKA_HOME}/tika-server/target/tika-server-
X.Y-SNAPSHOT.jar:${CTAKES_HOME}/desc:${CTAKES_HOME}/resources:`./gen-server-classpath.sh` org.apache.tika.
server.TikaServerCli --config tika-config.xml
```

With Tika server started, let's post that biomedical PDF file to it and see what happens!

Example Request

```
curl -T Vose-2013-American_Journal_of_Hematology.pdf -H "Content-Disposition: attachment;filename=Vose-2013-
American_Journal_of_Hematology.pdf" http://localhost:9998/rmeta
```

Example Response

And the output should be (much omitted from below):


```
[
    {
        "Content-Type": "application/pdf",
        "Creation-Date": "2013-11-20T13:24:11Z",
        "Last-Modified": "2013-11-22T14:13:25Z",
        "Last-Save-Date": "2013-11-22T14:13:25Z",
        "WPS-ARTICLEDOI": "10.1002/ajh.23615",
        "WPS-JOURNALDOI": "10.1002/(ISSN)1096-8652",
        "WPS-PROCLEVEL": "2",
        "X-Parsed-By": [
            "org.apache.tika.parser.CompositeParser",
            "org.apache.tika.parser.ctakes.CTAKESParser",
            "org.apache.tika.parser.DefaultParser",
            "org.apache.tika.parser.pdf.PDFParser"
        ],
        "X-TIKA:content": "\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nJW-
AJH#130002 1083..1088\n\n\nContinuing medical education activity\nin American Journal of Hematology\n\nCME
Author: Julie M. Vose, M.D., M.B.A.\nCME Editor: Ayalew Tefferi, M.D.\n\nCME Information: Mantle Cell Lymphoma:
2013 Update on\nDiagnosis, Risk-Stratification andClinical Management\nIf you wish to receive credit for this
activity, please refer\n\nto the website: www.wileyhealthlearning.com\n\nAccreditation and Designation
Statement:\n\nBlackwell Futura Media Services is accredited by the\nAccreditation Council for Continuing
Medical Education to\nprovide continuing medical education for physicians.\nBlackwell Futura Media Services
designates this journal-\n\nbased CME for a maximum of 1 AMA PRA Category 1\nCreditTM. Physicians should only
claim credit commensu-\nrate with the extent of their participation in the activity.\n\nEducational
Objectives\n\nUpon completion of this educational activity, participants\nwill be better able to:\n1. Identify
the histologic types and prognostic models\n\nused for mantle cell lymphoma\n2. Explain the different options
for initial therapy for man-\n\ntle cell lymphoma\n3. Explain the different options for treatment of
recurrent\n\nmantle cell lymphoma\n\nActivity Disclosures\n\nNo commercial support has been accepted related to
the\ndevelopment or publication of this activity.\nAuthor: Julie M. Vose, M.D., M.B.A., discloses
research\n\ngrant support from Allos Therapeautics/Spectrum, Bristol-\nMyers Squibb, Celgene, Genentech,
GlaxoSmithKline, Incyte\nCorp., Janssen Biotech, Millennium, Onyx Pharmaceuticals,\nPharmacyclics, Sanofi-
Aventis US, Inc., and US Biotest, Inc.\nCME \n",
        "X-TIKA:parse_time_millis": "366163",
        "access_permission:assemble_document": "true",
        "access_permission:can_modify": "true",
        "access_permission:can_print": "true",
        "access_permission:can_print_degraded": "true",
        "access_permission:extract_content": "true",
        "access_permission:extract_for_accessibility": "true",
        "access_permission:fill_in_form": "true",
        "access_permission:modify_annotations": "true",
        "created": "Wed Nov 20 05:24:11 PST 2013",
        "ctakes:AnatomicalSiteMention": [
            "Cell:189:193:C0007634,C1269647",
            "Media:432:437:C0162867",
            "Media:593:598:C0162867",
            "cell:967:971:C0007634,C1269647",
            "cell:1045:1049:C0007634,C1269647",
            "cell:1124:1128:C0007634,C1269647",
            "Media:2134:2139:C0162867",
            "cell:3716:3720:C0007634,C1269647",
            "cell:3839:3843:C0007634,C1269647",
            "lymph nodes:3920:3931:C1269047,C0024204",
            "spleen:3933:3939:C1278932,C0037993",
           ],
        "ctakes:DateAnnotation": [
            "October 2013:8713:8725:",
            "1/2:35241:35244:",
            "1/2:39154:39157:"
        ],
        "ctakes:DiseaseDisorderMention": [
            "Mantle Cell Lymphoma:182:202:C0334634,C0334634,C0334634,C0334634",
            "Lymphoma:194:202:C0024299,C0024299,C0024299,C0024299,C0024299,C0024299,C0024299,C0024299,C0024299",
            "ards:1673:1677:C0035222,C0035222,C0035222,C0035222,C0035222,C0035222",
            "dis:1955:1958:C0012634",
            "HEMATOLOGICAL MALIGNANCIES:3645:3671:C0376545,C0376545,C0376545",
```

```
        ],
        "ctakes:FractionAnnotation": [
            ".1088:19:24:",
            "2013.:8781:8786:",
            "10.1002:8857:8864:",
            ".2:11412:11414:",
            "11.:11849:11852:",
        ],
        "ctakes:MeasurementAnnotation": [
            "106 L:18964:18969:",
            "175 mg:23907:23913:"
        ],
        "ctakes:MedicationMention": [
            "duration:3987:3995:C2926735",
            "Bendamustine:5472:5484:C0525079,C0525079",
            "bortezomib:5609:5619:C1176309,C1176309",
            "Tyrosine:5735:5743:C0041485,C0041485",
            "formalin:12875:12883:C0949307,C0949307",
            "paraffin:12891:12899:C0030415",
        ],
        "ctakes:ProcedureMention": [
            "therapy:1025:1032:C0087111,C0087111,C0087111,C0087111",
            "treatment:1095:1104:C0087111,C0087111,C1533734,C0087111,C1533734,C0087111",
        ],
        "ctakes:RomanNumeralAnnotation": [
            "M:127:128:",
            "M:161:162:",
            "D:163:164:",
        ],
        "ctakes:SignSymptomMention": [
            "education:43:52:C0013658,C0013658,C0013658,C0424927,C0013658,C0013658,C0013658,C0013658",
            "M:127:128:C0024554,C0024554,C0024554,C0024554",
        ],
        "ctakes:schema": "coveredText:start:end:ontologyConceptArr",
        "date": "2013-11-22T14:13:25Z",
        "dc:format": "application/pdf; version=1.5",
        "dc:title": "JW-AJH#130002 1083..1088",
        "dcterms:created": "2013-11-20T13:24:11Z",
        "dcterms:modified": "2013-11-22T14:13:25Z",
        "meta:creation-date": "2013-11-20T13:24:11Z",
        "meta:save-date": "2013-11-22T14:13:25Z",
        "modified": "2013-11-22T14:13:25Z",
        "pdf:PDFVersion": "1.5",
        "pdf:encrypted": "false",
        "producer": "PDFlib PLOP 2.0.0p6 (SunOS)/Adobe LiveCycle PDFG ES",
        "resourceName": "Vose-2013-American_Journal_of_Hematology.pdf",
        "title": "JW-AJH#130002 1083..1088",
        "xmp:CreatorTool": "Arbortext Advanced Print Publisher 9.0.114/W",
        "xmpTPg:NPages": "7"
    }
]
```