

XMLFlattener in Scala

Hesiquio ballines

Description

The XMLFlattener is designed to run on IntelliJ. What it does is it uses end tags to put one tag and its contents into one line. The original intention is for use with Hadoop Map/Reduce so that you can use the default input line format. For some it may be easier to process it one line at a time as there might be line by line readers that would like one entry per line.

With that being said if you are using Hadoop Map/Reduce with XML I would recommend creating your own XMLInputFormat or using the Mahout implementation here:

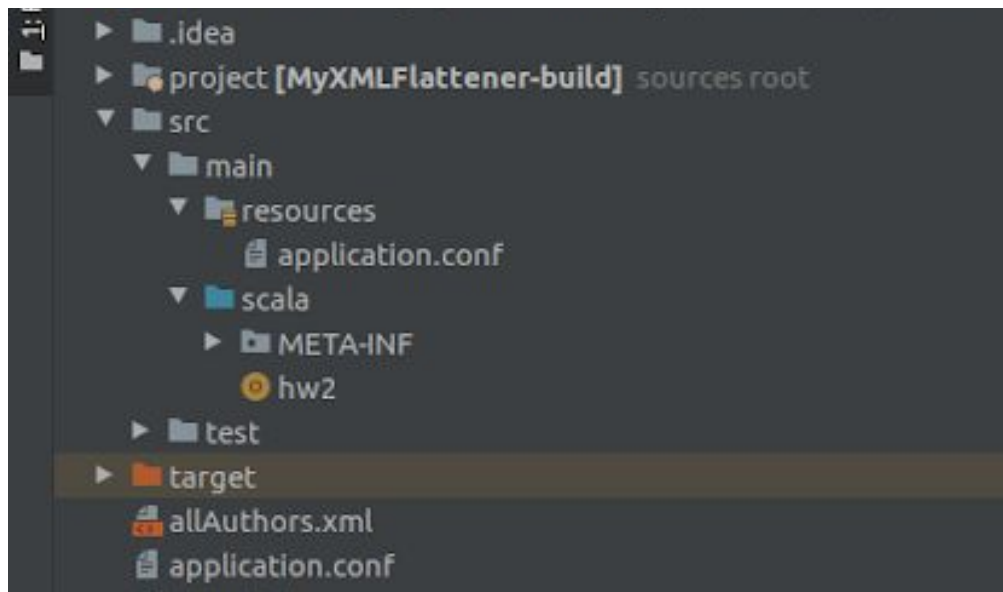
<https://github.com/apache/mahout/blob/66f164057e322d2e63ea02c35c9e30c3969e80b1/integration/src/main/java/org/apache/mahout/text/wikipedia/XMLInputFormat.java>

However, a limitation with Mahout's implementation is the lack of support for multiple tag parsing. For multiple tags I would recommend using Mohammed's Implementation here:

<https://github.com/Mohammed-siddiq/hadoop-XMLInputFormatWithMultipleTags>

Finally I have included example from and to files so you can see what the program produces based on the input and config files.

Project Structure



XML Processing

The XMLFlattener works simply by looking for specific end tags that can be entered in the config file as "tags". In order for the flattener to work properly there can be a max of one end tag per line and the XML must be well formed. The best format would be one tag but one end tag per line should work.

Config Format

Both filename and output can be directories but they should be xml files, text files are accepted.

tags : "</endtag1>|</endtag2>|...</endtagN>"

filename: "input.xml"

outputfile: "output.xml"

XML Input Example

"Input.xml", a bit larger than what is shown but this is an example taken from DBLP Public Dataset

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "small.dtd">
<dblp>
<article mdate="2018-01-07" key="tr/meltdown/s18" pubtype="informal">
  <author>Paul Kocher</author>
  <author>Daniel Genkin</author>
  <author>Daniel Gruss</author>
  <author>Werner Haas</author>
  <author>Mike Hamburg</author>
  <author>Moritz Lipp</author>
  <author>Stefan Mangard</author>
  <author>Thomas Prescher 0002</author>
  <author>Michael Schwarz 0001</author>
  <author>Yuval Yarom</author>
<title>Spectre Attacks: Exploiting Speculative Execution.</title>
<journal>meltdownattack.com</journal>
<year>2018</year>
<ee>https://spectreattack.com/spectre.pdf</ee>
</article>
<article mdate="2018-01-07" key="tr/meltdown/m18" pubtype="informal">
  <author>Moritz Lipp</author>
  <author>Michael Schwarz 0001</author>
  <author>Daniel Gruss</author>
  <author>Thomas Prescher 0002</author>
  <author>Werner Haas</author>
  <author>Stefan Mangard</author>
  <author>Paul Kocher</author>
  <author>Daniel Genkin</author>
  <author>Yuval Yarom</author>
  <author>Mike Hamburg</author>
<title>Meltdown</title>
<journal>meltdownattack.com</journal>
<ee>https://meltdownattack.com/meltdown.pdf</ee>
<year>2018</year>
</article>
<book mdate="2019-05-27" key="tr/acm/CS2013">
<title>Computer Science Curricula 2013</title>
<publisher>ACM Press and IEEE Computer Society Press</publisher>
<year>2013</year>
<ee>https://doi.org/10.1145/2534860</ee>
<isbn>978-1-4503-2309-3</isbn>
</book>
<article mdate="2017-06-08" key="tr/gte/TR-0263-08-94-165" pubtype="informal">
<author>Frank Manola</author>
<title>An Evaluation of Object-Oriented DBMS Developments: 1994 Edition.</title>
<journal>GTE Laboratories Incorporated</journal>
<volume>TR-0263-08-94-165</volume>
<month>August</month>
<year>1994</year>
<url>db/labs/gte/index.html#TR-0263-08-94-165</url>
</article>
<article mdate="2019-05-27" key="tr/gte/TR-0222-10-92-165" pubtype="informal">
<author>Michael L. Brodie</author>
```

XML Output Example

```
hw2.scala  application.conf  flatrest.xml  flatmini.xml
<article mdate="2018-01-07" key="tr/meltdown/s18" pubtype="informal"> <author>Paul Kocher</author> <author>Daniel Genkin</author> <author>Daniel Gruss</author> <author>Werner Haas</author>
<article mdate="2018-01-07" key="tr/meltdown/m18" pubtype="informal"> <author>Moritz Lipp</author> <author>Michael Schwarz 0001</author> <author>Daniel Gruss</author> <author>Thomas
<book mdate="2019-05-27" key="tr/acm/CS2013"><title>Computer Science Curricula 2013</title><publisher>ACM Press and IEEE Computer Society Press</publisher><year>2013</year><ee>https://doi.org/10.1145/3299061</ee></book>
<article mdate="2017-06-08" key="tr/gte/TR-0263-08-94-165" pubtype="informal"><author>Frank Manola</author><title>An Evaluation of Object-Oriented DBMS Developments: 1994 Edition.</title><journal>GTE Laboratories</journal></article>
<article mdate="2019-05-27" key="tr/gte/TR-0222-18-92-165" pubtype="informal"><author>Michael L. Brodie</author><author>Michael Stonebraker</author><title>DARWIN: On the Incremental Migration of Object-Oriented Applications to Relational Databases.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0174-12-91-165" pubtype="informal"><author>Mark F. Hornick</author><author>Joe D. Morrison</author><author>Farshad Nayeri</author><title>Integrating Object-Oriented Applications and Middleware with Relational Databases.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0149-06-89-165" pubtype="informal"><author>Frank Manola</author><title>Object Model Capabilities For Distributed Object Management.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0310-11-95-165" pubtype="informal"><author>Frank Manola</author><title>Integrating Object-Oriented Applications and Middleware with Relational Databases.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0146-06-91-165" pubtype="informal"><author>Alejandro P. Buchmann</author><author>M. Tamer Özsu</author><author>Dimitrios Georgakopoulos</author><title>Object-Oriented Database Management Systems: The State of the Art.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0231-08-93-165" pubtype="informal"><author>Frank Manola</author><author>Sandra Heiler</author><title>A 'RISC' Object Model for Object System Interfacing.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0244-12-93-165" pubtype="informal"><author>Frank Manola</author><title>MetaObject Protocol Concepts for a RISC Object Model.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0169-12-91-165" pubtype="informal"><author>Frank Manola</author><title>Object Data Language Facilities for Multimedia Data Types.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0332-11-90-165" pubtype="informal"><author>Frank Manola</author><author>Mark F. Hornick</author><author>Alejandro P. Buchmann</author><title>Object-Oriented Database Management Systems: The State of the Art.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0014-06-88-165" pubtype="informal"><author>Frank Manola</author><title>Distributed Object Management Technology.</title><journal>GTE Laboratories</journal></article>
<article mdate="2017-06-08" key="tr/gte/TR-0236-09-93-165" pubtype="informal"><author>Farshad Nayeri</author><author>Benjamin Hurwitz</author><title>Experiments with Dispatching in a Distributed Object-Oriented Database Machine.</title><journal>University of California</journal></article>
<article mdate="2017-06-08" key="tr/uc/erl-m79-28" pubtype="informal"><author>Michael Stonebraker</author><title>Muffin: A Distributed Database Machine.</title><journal>University of California</journal></article>
<article mdate="2018-06-28" key="tr/sql/X3H2-90-412" pubtype="informal"><author>David Beech</author><author>Cetin Ozbutun</author><title>Object Oriented DBMS as a Generalization of Relational DBMS.</title><journal>ANSI X3H2</journal><volume>X3H2-9</volume></article>
<article mdate="2018-06-28" key="tr/sql/X3H2-91-133rev1" pubtype="informal"><author>Krishna G. Kulkarni</author><author>Jim Melton</author><author>Jonathan Bauer</author><author>Mike Kelley</author><title>Object ADTs.</title><journal>ANSI X3H2</journal><volume>X3H2-9</volume></article>
<article mdate="2018-06-28" key="tr/sql/X3H2-90-292" pubtype="informal"><author>Phil Shaw</author><title>Modification of User Defined Types.</title><journal>ANSI X3H2</journal><volume>X3H2-9</volume></article>
<article mdate="2018-06-28" key="tr/sql/X3H2-91-083rev1" pubtype="informal"><author>Jim Melton</author><author>Jonathan Bauer</author><author>Krishna G. Kulkarni</author><title>Object ADTs.</title><journal>ANSI X3H2</journal><volume>X3H2-9</volume></article>
<article mdate="2018-06-28" key="tr/sql/X3H2-92-062" pubtype="informal"><author>David Beech</author><title>Unification of Value and Object ADTs.</title><journal>ANSI X3H2</journal><volume>X3H2-9</volume></article>
<article mdate="2017-06-08" key="tr/ibm/TW85191" pubtype="informal"><author>Rolf Sanders</author><title>Die Repräsentation räumlichen Wissens und die Behandlung von Einbettungsproblemen.</title><journal>LILOG-Report</journal></article>
<article mdate="2017-06-08" key="tr/ibm/LILO659" pubtype="informal"><author>Thomas Ludwig 0001</author><title>Algebraical Optimization of FTA-Expressions.</title><journal>LILOG-Report</journal></article>
<article mdate="2017-06-08" key="tr/ibm/LILO615" pubtype="informal"><author>Werner Ende</author><author>Claus-Rainer Rollinger</author><title>Wissensrepräsentation und Maschinelle Inferenz.</title><journal>LILOG-Report</journal></article>
<article mdate="2017-06-08" key="tr/ibm/LILO640" pubtype="informal"><author>Christoph Beierle</author><author>Udo Pleter</author><author>Hans Uszkoreit</author><title>An Algebraic Characterization of Query Languages for Object-Oriented Databases.</title><journal>LILOG-Report</journal></article>
<www mdate="2012-04-19" key="homepages/49/11192"><author>Matthew R. Francis</author> <title>Home Page</title></www>
<www mdate="2009-06-09" key="homepages/49/2365"><author>Nicholas Chia-Yuan Chang</author> <title>Home Page</title></www>
<www mdate="2009-06-10" key="homepages/49/5773"><author>Mingcheng Qu</author> <title>Home Page</title></www>
```