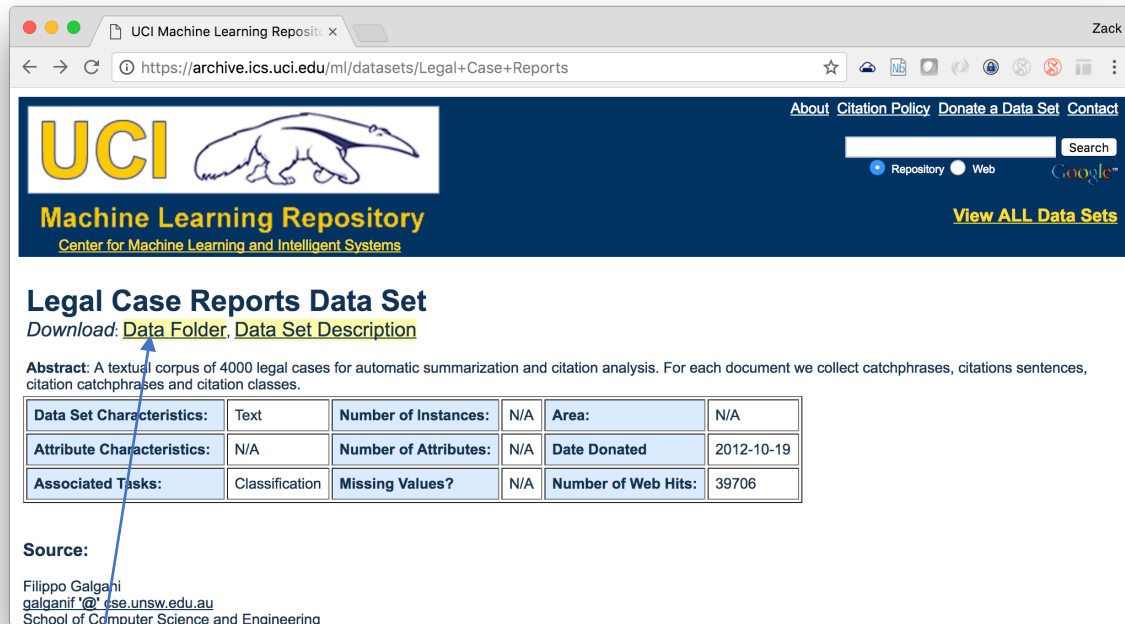


# Instructions for retrieving the Australian legal case report data set

Visit <https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports>



UCI Machine Learning Repository

**Legal Case Reports Data Set**

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** A textual corpus of 4000 legal cases for automatic summarization and citation analysis. For each document we collect catchphrases, citations sentences, citation catchphrases and citation classes.

<b>Data Set Characteristics:</b>	Text	<b>Number of Instances:</b>	N/A	<b>Area:</b>	N/A
<b>Attribute Characteristics:</b>	N/A	<b>Number of Attributes:</b>	N/A	<b>Date Donated</b>	2012-10-19
<b>Associated Tasks:</b>	Classification	<b>Missing Values?</b>	N/A	<b>Number of Web Hits:</b>	39706

**Source:**  
Filippo Galgani  
galgani.f@ese.unsw.edu.au  
School of Computer Science and Engineering

Click on the Data Folder link to download the corpus.

Unzip corpus.zip.

Name	Date Modified	Size	Kind
▶ citations_class	Jun 28, 2012, 8:47 PM	--	Folder
▶ citations_summ	Jun 28, 2012, 8:50 PM	--	Folder
▶ fulltext	Jun 28, 2012, 8:50 PM	--	Folder
readme.txt	Oct 10, 2012, 3:19 AM	6 KB	Plain Text

Open the readme.txt file to see a description of the files in the 3 directories.

The citation network can be built from the data in the **citations\_class** directory.

The description from the readme file:

citations\_class: Contains for each case a list of labeled citations. Fields:

- <name> : name of the case
- <AustLII> : link to the austlii page from where the document was taken
- <citations> : contains a list of <citation> elements
  - <citation> : a citation to an older case, it has an id attribute and contains the following elements:
    - <class> : the class of the citation as indicated on the document
    - <tocase> : the name of the case which is cited
    - <AustLII> : the link to the document of the case which is cited
    - <text> : paragraphs in the cited case where the current case is mentioned

There is a slight problem with the citations\_class xml formatting. The citation tag has a malformed attribute name="value" entry. This appears to be a global problem.

```
1 <?xml version="1.0"?>
2 <case>
3 <name>Broad Construction Services (WA) Pty Ltd v The Construction, Forestry, Mining &
4 <AustLII>http://www.austlii.edu.au/au/cases/cth/FCA/2006/44.html</AustLII>
5 <citations>
6 <citation id=c0">
7 <class>cited</class>
8 <tocase>Broad Construction Services (WA) Pty Ltd v The Construction, Forestry, Mining &
9 <AustLII>http://www.austlii.edu.au/au/cases//cth/federal_ct/2005/613.html</AustLII>
10 <text>9 At the conclusion of the hearing on 9 May 2005, I granted an interim injunct.
```

This can be fixed by scanning files for "**id=(c[0-9]+)**" and replacing with **id="\1"**. You may need to alter the regular expressions to match the engine of your choice.

After fixing the XML issue, you can parse the files and begin to analyze the network.

Citations are noted in the citations\_class file of the case MAKING the citation. For instance, in file **09\_266.xml**, you can see the AustLII of the current case (ends with /2009/266.html) and you can see a case it cites under the <citations> tag. For instance, the first citation cites a case called *Metcash Trading Limited v Bunn (No 5) [2009] FCA 16*, which has an AustLII ending in /2009/16.html.

```

1 <?xml version="1.0"?>
2 <case>
3   <name>Metcash Trading Limited v Bunn (No 6) (Corrigendum 2 April 2009) [2009] FCA 266 (27 Ma
4   <AustLII>http://www.austlii.edu.au/au/cases/cth/FCA/2009/266.html</AustLII>
5   <citations>
6     <citation id="c0">
7       <class>cited</class>
8       <tocase>Metcash Trading Limited v Bunn (No 5) [2009] FCA 16</tocase>
9       <AustLII>http://www.austlii.edu.au/au/cases//cth/FCA/2009/16.html</AustLII>
10      <text>In Metcash Trading Limited v Bunn (No 5) [2009] FCA 16 I concluded that there

```

If you go to the file **09\_16.xml**, you'll see that matching AustLII for the cited case.

```

1 <?xml version="1.0"?>
2 <case>
3   <name>Metcash Trading Limited v Bunn (No 5) [2009] FCA 16 (20 January 2009)</name>
4   <AustLII>http://www.austlii.edu.au/au/cases/cth/FCA/2009/16.html</AustLII>
5   <citations>
6     <citation id="c0">
7       <class>cited</class>
8       <tocase>Metcash Trading Limited v Bunn [2006] FCA 322</tocase>
9       <AustLII>http://www.austlii.edu.au/au/cases//cth/FCA/2006/322.html</AustLII>
10      <text>2 On 24 February 2006, the applicant companies (which I will describe collec

```

There is another data issue – in the /case/citations/citation/AustLII data, there is an extra slash in the path which must be cleaned up, or you won't get a match:

Cited: <AustLII>http://www.austlii.edu.au/au/cases/cth/FCA/2009/16.html</AustLII>  
Citing: <AustLII>http://www.austlii.edu.au/au/cases//cth/FCA/2009/16.html</AustLII>

One that is corrected, these identifiers are the keys in building the citation network:

```

http://www.austlii.edu.au/au/cases/cth/FCA/2009/266.html
cites
http://www.austlii.edu.au/au/cases/cth/FCA/2009/16.html

http://www.austlii.edu.au/au/cases/cth/FCA/2009/16.html
cites
http://www.austlii.edu.au/au/cases/cth/FCA/2006/322.html

```

Additional information in the **citations\_class** files, as well as data from the **citations\_summ** or **fulltext** directories, can be used to augment the citation network as you see fit.

Good luck!