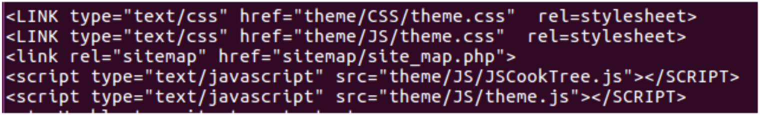
Big Data Architectures
For Machine Learning and Data Mining

April 26, 2017

Presentation of the data sets

Presentation of the data sets from ClueWeb

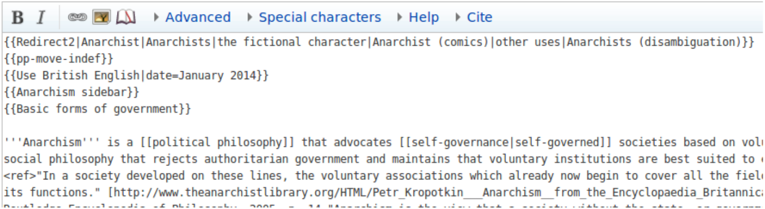
- Location: HDFS
- How to get there? **hadoop fs -ls /corpora/clueweb/09** or **hadoop fs -ls /corpora/clueweb/12**
- Example: In **/corpora/clueweb/09/ClueWeb09/_Arabic_1/ar0000/00.warc.gz** are websites as html
- Typical sizes of one file: 133.421.986 Bytes, for all in Clueweb09: 3.9 TByte, for all in Clueweb12: 4.5 TByte



```
<LINK type="text/css" href="theme/CSS/theme.css"  rel=stylesheet>
<LINK type="text/css" href="theme/JS/theme.css"  rel=stylesheet>
<link rel="sitemap" href="sitemap/site_map.php">
<script type="text/javascript" src="theme/JS/JSCookTree.js"></SCRIPT>
<script type="text/javascript" src="theme/JS/theme.js"></SCRIPT>
```

Presentation of the data sets from Wikipedia

- Location: HDFS
- How to get there? **hadoop fs -ls /corpora/enwiki/20160501/ enwiki-20160501-pages-articles.xml.bz2**
- Example: In **Wikipedia Anarchism** → **Edit** are texts with links
- Typical sizes: 12.896.598.324 Bytes



```
{{Redirect2|Anarchist|Anarchists|the fictional character|Anarchist (comics)|other uses|Anarchists (disambiguation)}}
{{pp-move-indef}}
{{Use British English|date=January 2014}}
{{Anarchism sidebar}}
{{Basic forms of government}}

'''Anarchism''' is a [[political philosophy]] that advocates [[self-governance|self-governed]] societies based on vol
social philosophy that rejects authoritarian government and maintains that voluntary institutions are best suited to
<ref>"In a society developed on these lines, the voluntary associations which already now begin to cover all the fiel
its functions." [http://www.theanarchistlibrary.org/HTML/Petr_Kropotkin___Anarchism__from_the_Encyclopaedia_Britannic
Routledge Encyclopedia of Philosophy. 2005. p. 14 "Anarchism is the view that a society without the state, or governm
```

Presentation of the data sets from the academic graph

- Location: HDFS
- How to get there? **hadoop fs -ls /corpora/corpus-microsoft-academic-graph**
- Example (shortened): graph in text
- Typical sizes: 26.6 GByte

Paper ID	Orig. paper title	Paper publish year	Paper publish date	Paper Docum. Object Identif.	Orig. venue name	Normal. venue name	Confer. ID mapped to venue name	Paper rank