# Region of Waterloo Event

## Region of Waterloo Demo, 24 November 2015

This is a notebook to demo how we're forseeing the rapid prototyping of work with web archives.

Note that we can begin to intersperse text with the code that we're writing, to enable the sharing of notebooks and research ideas.

```
:cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHO
```

3 items

| string value |
|---|
| /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar |
| file:/Users/ianmilligan1/Dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/ |
| file:/Users/ianmilligan1/dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/lib/common.common-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0.jar |

```
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
```

```
var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIV
var arcdir="/Users/ianmilligan1/dropbox/warcs-workshop";
```

/Users/ianmilligan1/dropbox/warcs-workshop

```
val r =
RecordLoader.loadArc(arc,
sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))
.countItems()
.take(10)
```

▦ 📊 ◕

```
al r =
ecordLoader.loadArc(arc,
c)
keepMimeTypes(Set("text/html"))
discardDate(null)
map(r => {
al t = ExtractRawText(r.getBodyContent)
ER3Classifier("/Users/ianmilligan1/dropbox/ner/stanford-ner-2015-04-20/
al entities = NER3Classifier.classify(t)
al len = 100
r.getCrawldate, r.getMimeType, entities, r.getUrl, if ( t.length > len
en) else t)})
collect()
```

1570 items (Out of 1570 items, only the 25 first items are shown)

| _1 | _2 | _3 |
|---|---|---|
| 20051004 | text/html | {"PERSON":["Bill Pay"],"ORGANIZATION":[],"LOCATION" |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |

| | | |
|---|---|---|
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":["Canada" |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":["Internet Information |
| 20051004 | text/html | {"PERSON":["Langley"],"ORGANIZATION":["Clean Start Government","Party","Party","Canadian Citizens","Un Party","Party","Party","Party","Party","Senate","Se Treasury","UIC","UIC","RCMP","ICOS Corp.","Darwin M Party","Surrey"],"LOCATION": ["Winnipeg","Canada","Canada","Canada","Canada","Qu States","Canada","Canada","Canada","Canada","Canada Care"]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":["Michel Garneau","Marie-Claude Charleboi Putes Parti Populaire des Putes Coalition Pour les des Putes","Coalition Pour les Droits des Travaille Workers","PPP","Ottawa Citizen","PPP Parti Populair |

| | | |
|---|---|---|
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":[],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":["Socialist Party of Ca |
| 20051004 | text/html | {"PERSON":["Hardial Bains"],"ORGANIZATION":["Marxis Volume","Communist Party of Canada -LRB- Marxist - Canada","Elections Canada","Marxist-Leninist Party Steel Industry Restructuring • Mittal","Blockades C ["South Africa","Indonesia","Amherst Street","Montr |
| 20051004 | text/html | {"PERSON":["Dave Trautman"],"ORGANIZATION":["Commun |
| 20051004 | text/html | {"PERSON":["Carol Taylor","Blair Longley"],"ORGANIZ Canada","Following Elections Canada Conferences","M District Association Become","Nunavut Electoral Dis |
| 20051004 | text/html | {"PERSON":["Harper","David Dingwall","Paul Martin", MacKaySeptember","Rona","Michael Chong"],"ORGANIZAT Party","Conservative Finance Critic Monte Solberg", Edward","Conservative Party of Canada","Agent of th Canada Shorter Waiting Times Tax Relief and Better ["Canada","Ottawa"]} |
| 20051004 | text/html | {"PERSON":["Benvenuto","Taeiôg"],"ORGANIZATION":["C Canada"],"LOCATION":[]} |
| 20051004 | text/html | {"PERSON":[],"ORGANIZATION":["WSM","WSM"],"LOCATION |
| | | |

| | | |
|---|---|---|
| 20051004 | text/html | {"PERSON":["Paul Russell Chef de l'UPCF"],"ORGANIZA l'espoir que les Canadiens"],"LOCATION":[]} |

```
def createClickableLink(url: String, date: String): String = {
"<a href='http://web.archive.org/web/" + date + "/" + url + "'>" +
url + "</a>"
}
```

```
val r =
RecordLoader.loadArc(arc,
sc)
.keepValidPages()
.map(r => {
val t = ExtractRawText(r.getBodyContent)
val len = 100
(r.getCrawldate, createClickableLink(r.getUrl,
r.getCrawldate), if ( t.length > len ) t.substring(0, len) else t)})
.collect()
```

1510 items (Out of 1510 items, only the 25 first items are shown)

| _1 | _2 |
|---|---|
| 20051004 | http://geocities.com/CapitolHill/2823/ (http://web.archive.org/web/20051004/http://geocities.com/Cap |
| 20051004 | http://walnet.org/ppp/index2.html (http://web.archive.org/web |
| 20051004 | http://worldsocialism.org/canada/ (http://web.archive.org/web |
| 20051004 | http://cpcml.ca/ (http://web.archive.org/web/20051004/http:// |
| 20051004 | http://communist-party.ca/ (http://web.archive.org/web/200510 |
| 20051004 | http://partimarijuana.org/index.en.php3 (http://web.archive.org/web/20051004/http://partimarijuana.or |
| 20051004 | http://agoracosmopolite.com/ (http://web.archive.org/web/2005 |
| 20051004 | http://www.worldsocialism.org/404.html (http://web.archive.org/web/20051004/http://www.worldsocialis |
| 20051004 | http://geocities.com/upcf/ (http://web.archive.org/web/200510 |
| 20051004 | http://worldsocialism.org/canada/contents.htm (http://web.archive.org/web/20051004/http://worldsocialism.or |

| | |
|---|---|
| 20051004 | http://communist-party.ca/top.html (http://web.archive.org/we<br>party.ca/top.html) |
| 20051004 | http://members.aol.com/totarisse/ucd-dcu.html<br>(http://web.archive.org/web/20051004/http://members.aol.com/t |
| 20051004 | http://home.ican.net/%7Ealexng/ndp.html<br>(http://web.archive.org/web/20051004/http://home.ican.net/%7E |
| 20051004 | http://themis.geocities.yahoo.com/jsoff.css?thIP=207.241.225.<br>(http://web.archive.org/web/20051004/http://themis.geocities.<br>thIP=207.241.225.107&thTs=1128453183) |
| 20051004 | http://worldsocialism.org/canada/spchdr.htm<br>(http://web.archive.org/web/20051004/http://worldsocialism.or |
| 20051004 | http://westernblockparty.com/ (http://web.archive.org/web/200 |
| 20051004 | http://liberal.ca/ (http://web.archive.org/web/20051004/http: |
| 20051004 | http://freedomparty.ca/ (http://web.archive.org/web/20051004/ |
| 20051004 | http://communist-party.ca/current_banner.gif (http://web.arch<br>party.ca/current_banner.gif) |
| 20051004 | http://download.macromedia.com/pub/shockwave/cabs/flash/swfla<br>(http://web.archive.org/web/20051004/http://download.macromed |
| 20051004 | http://themis.geocities.yahoo.com/jsoff.css?thIP=207.241.225.<br>(http://web.archive.org/web/20051004/http://themis.geocities.<br>thIP=207.241.225.107&thTs=1128453185) |
| 20051004 | http://blocquebecois.org/fr/default.asp<br>(http://web.archive.org/web/20051004/http://blocquebecois.org |
| 20051004 | http://www.geocities.com/cgi-bin/counter/ccsp<br>(http://web.archive.org/web/20051004/http://www.geocities.com |
| 20051004 | http://members.aol.com/robots.txt/<br>(http://web.archive.org/web/20051004/http://members.aol.com/ |

| | |
|---|---|
| | (http://web.archive.org/web/20051004/http://members.aol.com/r |
| 20051004 | http://greenparty.ca/ (http://web.archive.org/web/20051004/ht |