Incremental Crawling Scripts Test

1.

```
./whole-web-crawling-incremental seeds 10 1
rm: seeds/it_seeds/urls: No such file or directory
Injector: starting at 2011-03-27 15:46:15
Injector: crawlDb: crawl/crawldb
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-27 15:46:31, elapsed: 00:00:15
Fetcher: starting at 2011-03-27 15:46:59
Fetcher: segment: crawl/segments/20110327154649
Fetcher: threads: 10
QueueFeeder finished: total 10 records + hit by time limit :0
fetching http://simple.wikipedia.org/wiki/%C2%A3sd
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
fetching http://simple.wikipedia.org/wiki/%2B44
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
fetching http://simple.wikipedia.org/wiki/%28What%27s_the_Story%29_Morning_Glory%3F
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
fetching http://simple.wikipedia.org/wiki/%C3%81lvaro_Mej%C3%ADa_P%C3%A9rez
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
fetching http://simple.wikipedia.org/wiki/%C3%81lvaro_Lopes_Can%C3%A7ado
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
fetching http://simple.wikipedia.org/wiki/%2703_Bonnie_&_Clyde
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233656322
  now           = 1301233656859
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
```

```
  nextFetchTime = 1301233662094
  now           = 1301233657867
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233662094
  now           = 1301233658939
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233662094
  now           = 1301233660020
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233662094
  now           = 1301233661025
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233662094
  now           = 1301233662032
  0. http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
  1. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  2. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  3. http://simple.wikipedia.org/wiki/%27N_Sync
fetching http://simple.wikipedia.org/wiki/%C3%81lvaro_Arbeloa
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233667900
  now           = 1301233663039
  0. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  1. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  2. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
```

```
  minCrawlDelay = 0
  nextFetchTime = 1301233667900
  now           = 1301233664285
  0. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  1. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  2. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233667900
  now           = 1301233665409
  0. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  1. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  2. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233667900
  now           = 1301233666415
  0. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  1. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  2. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233667900
  now           = 1301233667516
  0. http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
  1. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  2. http://simple.wikipedia.org/wiki/%27N_Sync
fetching http://simple.wikipedia.org/wiki/%27s-Hertogenbosch
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233668525
  0. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  1. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233669647
  0. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  1. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233670783
  0. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  1. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
```

```
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233671791
  0. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  1. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233672903
  0. http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
  1. http://simple.wikipedia.org/wiki/%27N_Sync
fetching http://simple.wikipedia.org/wiki/%60Abdu%27l-Bah%C3%A1
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233673363
  now           = 1301233673908
  0. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233678937
  now           = 1301233674914
  0. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233678937
  now           = 1301233675919
  0. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233678937
  now           = 1301233676925
  0. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233678937
  now           = 1301233677930
  0. http://simple.wikipedia.org/wiki/%27N_Sync
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
```

```
  nextFetchTime = 1301233678937
  now           = 1301233679037
  0. http://simple.wikipedia.org/wiki/%27N_Sync
fetching http://simple.wikipedia.org/wiki/%27N_Sync
-finishing thread FetcherThread, activeThreads=9
-finishing thread FetcherThread, activeThreads=8
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=6
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-27 15:48:04, elapsed: 00:01:04
CrawlDb update: starting at 2011-03-27 15:48:09
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20110327154649]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-27 15:48:19, elapsed: 00:00:09
LinkDb: starting at 2011-03-27 15:48:24
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/20110327154649
LinkDb: finished at 2011-03-27 15:48:32, elapsed: 00:00:07
SolrIndexer: starting at 2011-03-27 15:48:36
SolrIndexer: finished at 2011-03-27 15:48:54, elapsed: 00:00:17
Injector: starting at 2011-03-27 15:48:58
Injector: crawlDb: crawl/crawldb
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-27 15:49:15, elapsed: 00:00:16
Fetcher: starting at 2011-03-27 15:49:42
Fetcher: segment: crawl/segments/20110327154933
Fetcher: threads: 10
QueueFeeder finished: total 10 records + hit by time limit :0
fetching http://simple.wikipedia.org/wiki/%C3%81ngel_S%C3%A1nchez_%28baseball%29
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=9
fetching http://simple.wikipedia.org/wiki/%C3%81ngel_Javier_Arizmendi
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
fetching http://simple.wikipedia.org/wiki/%C3%81o_d%C3%A0i
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=7
fetching http://simple.wikipedia.org/wiki/%C3%82nderson_Polga
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
fetching http://simple.wikipedia.org/wiki/%C3%81lvaro_Recoba
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=5
```

```
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
fetching http://simple.wikipedia.org/wiki/%C3%81lvaro_Sabor%C3%ADo
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233819220
  now           = 1301233819888
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233825026
  now           = 1301233820895
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233825026
  now           = 1301233821902
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233825026
  now           = 1301233823027
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233825026
  now           = 1301233824032
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
```

```
  minCrawlDelay = 0
  nextFetchTime = 1301233825026
  now           = 1301233825039
  0. http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
  1. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  2. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  3. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
fetching http://simple.wikipedia.org/wiki/%C3%81ttila_de_Carvalho
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
  now           = 1301233826047
  0. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  1. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  2. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
  now           = 1301233827053
  0. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  1. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  2. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
  now           = 1301233828058
  0. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  1. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  2. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
  now           = 1301233829165
  0. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  1. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  2. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
  now           = 1301233830170
  0. http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
  1. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  2. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
fetching http://simple.wikipedia.org/wiki/%C3%81stor_Piazzolla
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233830697
```

```
  now           = 1301233831176
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233836264
  now           = 1301233832271
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233836264
  now           = 1301233833402
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233836264
  now           = 1301233834407
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233836264
  now           = 1301233835414
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233836264
  now           = 1301233836420
  0. http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
  1. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
fetching http://simple.wikipedia.org/wiki/%C3%82nderson_Lima_Veiga
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233841867
  now           = 1301233837520
  0. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233841867
  now           = 1301233838633
```

```
  0. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233841867
  now           = 1301233839667
  0. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233841867
  now           = 1301233840700
  0. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://simple.wikipedia.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301233841867
  now           = 1301233841923
  0. http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
fetching http://simple.wikipedia.org/wiki/%C3%81ngel_de_Saavedra,_Duke_of_Rivas
-finishing thread FetcherThread, activeThreads=9
-finishing thread FetcherThread, activeThreads=8
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=6
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-27 15:50:47, elapsed: 00:01:04
CrawlDb update: starting at 2011-03-27 15:50:52
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20110327154933]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-27 15:51:03, elapsed: 00:00:10
LinkDb: starting at 2011-03-27 15:51:08
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/20110327154649
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/20110327154933
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2011-03-27 15:51:27, elapsed: 00:00:18
SolrIndexer: starting at 2011-03-27 15:51:31
SolrIndexer: finished at 2011-03-27 15:51:54, elapsed: 00:00:22
```

2.

```
$ ./whole-web-crawling-incremental urls-input/MR6 5 2
rm -r crawl
```

```
rm: urls-input/MR6/it_seeds: No such file or directory
2 urls to crawl
rm: urls-input/MR6/it_seeds/urls: No such file or directory

bin/nutch inject crawl/crawldb urls-input/MR6/it_seeds
Injector: starting at 2011-03-27 15:28:07
Injector: crawlDb: crawl/crawldb
Injector: urlDir: urls-input/MR6/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-27 15:28:22, elapsed: 00:00:15

generate-fetch-updatedb-invertlinks-index-merge iteration 0:

bin/nutch generate crawl/crawldb crawl/segments -topN 5
Generator: starting at 2011-03-27 15:28:29 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 5 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/20110327152839 Generator: finished at 2011-03-27 15:28:45, elapsed: 00:00:15

bin/nutch fetch crawl/segments/20110327152839
Fetcher: starting at 2011-03-27 15:28:49
Fetcher: segment: crawl/segments/20110327152839
Fetcher: threads: 10
QueueFeeder finished: total 2 records + hit by time limit :0
fetching http://localhost:8080/qui/2.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232536012
  now           = 1301232538470
  0. http://localhost:8080/qui/1.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232543848
  now           = 1301232539474
  0. http://localhost:8080/qui/1.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232543848
  now           = 1301232540479
  0. http://localhost:8080/qui/1.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232543848
  now           = 1301232541514
  0. http://localhost:8080/qui/1.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232543848
  now           = 1301232542619
```

```
  0. http://localhost:8080/qui/1.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301232543848
  now           = 1301232543640
  0. http://localhost:8080/qui/1.html
fetching http://localhost:8080/qui/1.html
-finishing thread FetcherThread, activeThreads=9
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=6
-finishing thread FetcherThread, activeThreads=8
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-27 15:29:07, elapsed: 00:00:17

bin/nutch updatedb crawl/crawldb crawl/segments/20110327152839
CrawlDb update: starting at 2011-03-27 15:29:12
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20110327152839]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-27 15:29:22, elapsed: 00:00:09

bin/nutch invertlinks crawl/linkdb -dir crawl/segments
LinkDb: starting at 2011-03-27 15:29:27
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/20110327152839
LinkDb: finished at 2011-03-27 15:29:34, elapsed: 00:00:06

rm: crawl/new_indexes: No such file or directory
bin/nutch index crawl/new_indexes crawl/crawldb crawl/linkdb crawl/segments/20110327152839
Indexer: starting at 2011-03-27 15:29:39
Indexer: finished at 2011-03-27 15:29:57, elapsed: 00:00:18

bin/nutch merge crawl/temp_indexes/part-1 crawl/indexes crawl/new_indexes
IndexMerger: starting at 2011-03-27 15:30:03
IndexMerger: merging indexes to: crawl/temp_indexes/part-1
Adding file:/Users/simpatico/nutch-1.2/crawl/new_indexes/part-00000
IndexMerger: finished at 2011-03-27 15:30:05, elapsed: 00:00:02

rm: crawl/indexes: No such file or directory

generate-fetch-updatedb-invertlinks-index-merge iteration 1:

bin/nutch generate crawl/crawldb crawl/segments -topN 5
Generator: starting at 2011-03-27 15:30:10 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 5 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb -stats
CrawlDb statistics start: crawl/crawldb
Statistics for CrawlDb: crawl/crawldb
TOTAL urls:        2
retry 0:        2
min score:        1.0
avg score:        1.0
max score:        1.0
```

```
status 2 (db_fetched):        2
CrawlDb statistics: done

bin/nutch mergedb crawl/temp_crawldb crawl/crawldb
CrawlDb merge: starting at 2011-03-27 15:30:37
Adding crawl/crawldb
CrawlDb merge: finished at 2011-03-27 15:30:44, elapsed: 00:00:07

rm: crawl/allcrawldb: No such file or directory

rm: crawl/allcrawldb/dump: No such file or directory
bin/nutch readdb crawl/allcrawldb -dump crawl/allcrawldb/dump
CrawlDb dump: starting
CrawlDb db: crawl/allcrawldb
CrawlDb dump: done

CrawlDb statistics start: crawl/allcrawldb
Statistics for CrawlDb: crawl/allcrawldb
TOTAL urls:        2
retry 0:        2
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        2
CrawlDb statistics: done
```

3.

```
./whole-web-crawling-incremental
Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

2 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 18:59:03
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 18:59:32, elapsed: 00:00:29


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 10
Generator: starting at 2011-03-29 18:59:48 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 10 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        2
retry 0:        2
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        2
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
```

```
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
rmr: cannot remove seeds/it_seeds: No such file or directory.
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

2 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:01:19
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:01:40, elapsed: 00:00:21


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
rmr: cannot remove crawl/segments/0: No such file or directory.
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 10
Generator: starting at 2011-03-29 19:02:00 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 10 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        2
retry 0:        2
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        2
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incrementa^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
rmr: cannot remove seeds/it_seeds: No such file or directory.
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

2 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:07:31
```

```
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:07:51, elapsed: 00:00:20


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
rmr: cannot remove crawl/segments/0: No such file or directory.
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 10
Generator: starting at 2011-03-29 19:08:05 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 10 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329190824 Generator: finished at 2011-03-29 19:08:33, elapsed: 00:00:28

bin/nutch fetch crawl/segments/0/20110329190824
Fetcher: starting at 2011-03-29 19:08:42
Fetcher: segment: crawl/segments/0/20110329190824
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest3.html
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:08:59, elapsed: 00:00:17

bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329190824
CrawlDb update: starting at 2011-03-29 19:09:03
CrawlDb update: db: crawl/crawldb/0
CrawlDb update: segments: [crawl/segments/0/20110329190824]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:09:14, elapsed: 00:00:10

bin/nutch invertlinks crawl/linkdb -dir crawl/segments/0
LinkDb: starting at 2011-03-29 19:09:20
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329190824
LinkDb: finished at 2011-03-29 19:09:29, elapsed: 00:00:08

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/0 crawl/linkdb crawl/segments/0/20110329190824
SolrIndexer: starting at 2011-03-29 19:09:33
SolrIndexer: finished at 2011-03-29 19:09:50, elapsed: 00:00:16

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329190824

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        3
retry 0:        3
min score:        1.0
avg score:        1.0
```

```
max score:        1.0
status 2 (db_fetched):        3
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental -f
bin/hadoop dfs -rmr crawl
Deleted file:/Users/simpatico/nutch-1.2/crawl

curl --fail http://localhost:8080/solr/update?commit=true -d '<delete><query>*:*</query></delete>'
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">20</int></lst>
</response>

rmr: cannot remove seeds/it_seeds: No such file or directory.
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

2 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
test: File crawl/crawldb/0 does not exist.

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
^C

generate-fetch-updatedb-invertlinks-index-merge iteration 0:
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
^C
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental -f
bin/hadoop dfs -rmr crawl
Deleted file:/Users/simpatico/nutch-1.2/crawl

curl --fail http://localhost:8080/solr/update?commit=true -d '<delete><query>*:*</query></delete>'
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">21</int></lst>
</response>

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
test: File crawl/crawldb/0 does not exist.

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:12:50
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:13:07, elapsed: 00:00:17


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
rmr: cannot remove crawl/segments/0: No such file or directory.
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 10
Generator: starting at 2011-03-29 19:13:35 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 10 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329191349 Generator: finished at 2011-03-29 19:13:55, elapsed: 00:00:20

bin/nutch fetch crawl/segments/0/20110329191349
Fetcher: starting at 2011-03-29 19:14:01
```

```
Fetcher: segment: crawl/segments/0/20110329191349
Fetcher: threads: 10
QueueFeeder finished: total 3 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest1.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301418852588
  now           = 1301418856274
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301418852588
  now           = 1301418857301
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301418862874
  now           = 1301418858311
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301418862874
  now           = 1301418859322
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301418862874
  now           = 1301418860325
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
^C
bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329191349
^C
bin/nutch invertlinks crawl/linkdb -dir crawl/segments/0
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental -f
bin/hadoop dfs -rmr crawl
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists
```

```
bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:15:43
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:16:02, elapsed: 00:00:19


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 10
Generator: starting at 2011-03-29 19:16:19 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 10 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329191632 Generator: finished at 2011-03-29 19:16:38, elapsed: 00:00:18

bin/nutch fetch crawl/segments/0/20110329191632
Fetcher: starting at 2011-03-29 19:16:45
Fetcher: segment: crawl/segments/0/20110329191632
Fetcher: threads: 10
QueueFeeder finished: total 3 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest1.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419017701
  now           = 1301419021893
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419017701
  now           = 1301419022895
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419028802
  now           = 1301419023913
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419028802
  now           = 1301419024961
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
```

```
  minCrawlDelay = 0
  nextFetchTime = 1301419028802
  now           = 1301419025964
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419028802
  now           = 1301419026970
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419028802
  now           = 1301419027978
  0. http://localhost:8080/nutch/scoringtest3.html
  1. http://localhost:8080/qui/scoringtest.html
fetching http://localhost:8080/nutch/scoringtest3.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419033940
  now           = 1301419028983
  0. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419033940
  now           = 1301419029985
  0. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419033940
  now           = 1301419030988
  0. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419033940
  now           = 1301419031990
  0. http://localhost:8080/qui/scoringtest.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1301419033940
  now           = 1301419032994
  0. http://localhost:8080/qui/scoringtest.html
```

```
fetching http://localhost:8080/qui/scoringtest.html
-finishing thread FetcherThread, activeThreads=8
-finishing thread FetcherThread, activeThreads=8
-activeThreads=8, spinWaiting=7, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=6
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
^C
bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329191632
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:18:24
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:18:43, elapsed: 00:00:19


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 1
Generator: starting at 2011-03-29 19:19:03 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329191920 Generator: finished at 2011-03-29 19:19:26, elapsed: 00:00:22

bin/nutch fetch crawl/segments/0/20110329191920
Fetcher: starting at 2011-03-29 19:19:30
Fetcher: segment: crawl/segments/0/20110329191920
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest1.html
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:19:48, elapsed: 00:00:17

bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329191920
```

```
CrawlDb update: starting at 2011-03-29 19:19:56
CrawlDb update: db: crawl/crawldb/0
CrawlDb update: segments: [crawl/segments/0/20110329191920]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:20:09, elapsed: 00:00:12

bin/nutch invertlinks crawl/linkdb -dir crawl/segments/0
LinkDb: starting at 2011-03-29 19:20:17
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329191920
LinkDb: finished at 2011-03-29 19:20:28, elapsed: 00:00:10

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/0 crawl/linkdb crawl/segments/0/20110329191920
SolrIndexer: starting at 2011-03-29 19:20:36
SolrIndexer: finished at 2011-03-29 19:21:01, elapsed: 00:00:24

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329191920

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        3
retry 0:        3
min score:        1.0
avg score:        1.0
max score:        1.0
status 1 (db_unfetched):        2
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
test: File crawl/crawldb/1 does not exist.

bin/nutch inject crawl/crawldb/1 seeds/it_seeds
Injector: starting at 2011-03-29 19:21:44
Injector: crawlDb: crawl/crawldb/1
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:22:53, elapsed: 00:01:09


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
rmr: cannot remove crawl/segments/1: No such file or directory.
rm: cannot remove crawl/crawldb/1/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/1/..locked.crc: No such file or directory.
bin/nutch generate crawl/crawldb/1 crawl/segments/1 -topN 1
Generator: starting at 2011-03-29 19:23:19 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/1/20110329192332 Generator: finished at 2011-03-29 19:23:40, elapsed: 00:00:20

bin/nutch fetch crawl/segments/1/20110329192332
Fetcher: starting at 2011-03-29 19:23:46
Fetcher: segment: crawl/segments/1/20110329192332
^C
bin/nutch updatedb crawl/crawldb/1 crawl/segments/1/20110329192332
^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ^C
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
```

```
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:24:51
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:25:22, elapsed: 00:00:30


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 1
Generator: starting at 2011-03-29 19:25:58 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329192612 Generator: finished at 2011-03-29 19:26:23, elapsed: 00:00:24

bin/nutch fetch crawl/segments/0/20110329192612
Fetcher: starting at 2011-03-29 19:26:31
Fetcher: segment: crawl/segments/0/20110329192612
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest3.html
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:26:52, elapsed: 00:00:21

bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329192612
CrawlDb update: starting at 2011-03-29 19:27:07
CrawlDb update: db: crawl/crawldb/0
CrawlDb update: segments: [crawl/segments/0/20110329192612]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:27:28, elapsed: 00:00:20

bin/nutch invertlinks crawl/linkdb -dir crawl/segments/0
LinkDb: starting at 2011-03-29 19:27:39
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329192612
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2011-03-29 19:28:05, elapsed: 00:00:25

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/0 crawl/linkdb crawl/segments/0/20110329192612
SolrIndexer: starting at 2011-03-29 19:28:10
SolrIndexer: finished at 2011-03-29 19:28:33, elapsed: 00:00:23

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329192612
```

```
bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        3
retry 0:        3
min score:        1.0
avg score:        1.0
max score:        1.0
status 1 (db_unfetched):        1
status 2 (db_fetched):        2
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
mkdir: cannot create directory crawl/crawldb/1: File exists

bin/nutch inject crawl/crawldb/1 seeds/it_seeds
Injector: starting at 2011-03-29 19:29:28
Injector: crawlDb: crawl/crawldb/1
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:29:58, elapsed: 00:00:30


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/1
rm: cannot remove crawl/crawldb/1/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/1/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/1 crawl/segments/1 -topN 1
Generator: starting at 2011-03-29 19:30:25 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/1/20110329193040 Generator: finished at 2011-03-29 19:30:52, elapsed: 00:00:26

bin/nutch fetch crawl/segments/1/20110329193040
Fetcher: starting at 2011-03-29 19:31:02
Fetcher: segment: crawl/segments/1/20110329193040
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest1.html
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:31:31, elapsed: 00:00:28

bin/nutch updatedb crawl/crawldb/1 crawl/segments/1/20110329193040
CrawlDb update: starting at 2011-03-29 19:31:43
CrawlDb update: db: crawl/crawldb/1
CrawlDb update: segments: [crawl/segments/1/20110329193040]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:31:57, elapsed: 00:00:14
```

```
bin/nutch invertlinks crawl/linkdb -dir crawl/segments/1
LinkDb: starting at 2011-03-29 19:32:06
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/1/20110329193040
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2011-03-29 19:32:23, elapsed: 00:00:16

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/1 crawl/linkdb crawl/segments/1/20110329193040
SolrIndexer: starting at 2011-03-29 19:32:27
SolrIndexer: finished at 2011-03-29 19:32:44, elapsed: 00:00:16

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/1/20110329193040

bin/nutch readdb crawl/crawldb/1 -stats
CrawlDb statistics start: crawl/crawldb/1
Statistics for CrawlDb: crawl/crawldb/1
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
test: File crawl/crawldb/2 does not exist.

bin/nutch inject crawl/crawldb/2 seeds/it_seeds
Injector: starting at 2011-03-29 19:33:46
Injector: crawlDb: crawl/crawldb/2
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:34:10, elapsed: 00:00:23


generate-fetch-updatedb-invertlinks-index-merge iteration 0:
rmr: cannot remove crawl/segments/2: No such file or directory.
rm: cannot remove crawl/crawldb/2/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/2/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/2 crawl/segments/2 -topN 1
Generator: starting at 2011-03-29 19:34:28 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/2/20110329193439 Generator: finished at 2011-03-29 19:34:45, elapsed: 00:00:16

bin/nutch fetch crawl/segments/2/20110329193439
Fetcher: starting at 2011-03-29 19:34:49
Fetcher: segment: crawl/segments/2/20110329193439
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/nutch/scoringtest3.html
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:35:01, elapsed: 00:00:12

bin/nutch updatedb crawl/crawldb/2 crawl/segments/2/20110329193439
```

```
CrawlDb update: starting at 2011-03-29 19:35:07
CrawlDb update: db: crawl/crawldb/2
CrawlDb update: segments: [crawl/segments/2/20110329193439]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:35:20, elapsed: 00:00:12

bin/nutch invertlinks crawl/linkdb -dir crawl/segments/2
LinkDb: starting at 2011-03-29 19:35:26
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/2/20110329193439
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2011-03-29 19:35:46, elapsed: 00:00:19

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/2 crawl/linkdb crawl/segments/2/20110329193439
SolrIndexer: starting at 2011-03-29 19:35:53
SolrIndexer: finished at 2011-03-29 19:36:14, elapsed: 00:00:20

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/2/20110329193439

bin/nutch readdb crawl/crawldb/2 -stats
CrawlDb statistics start: crawl/crawldb/2
Statistics for CrawlDb: crawl/crawldb/2
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
rmr: cannot remove seeds/it_seeds: No such file or directory.
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:38:32
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:38:48, elapsed: 00:00:16


generate-fetch-invertlinks-updatedb-index iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 1
Generator: starting at 2011-03-29 19:39:03 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: Partitioning selected urls for politeness. Generator: segment: crawl/segments
/0/20110329193913 Generator: finished at 2011-03-29 19:39:18, elapsed: 00:00:15

bin/nutch fetch crawl/segments/0/20110329193913
Fetcher: starting at 2011-03-29 19:39:22
Fetcher: segment: crawl/segments/0/20110329193913
Fetcher: threads: 10
QueueFeeder finished: total 1 records + hit by time limit :0
fetching http://localhost:8080/qui/scoringtest.html
```

```
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2011-03-29 19:39:35, elapsed: 00:00:13

bin/nutch updatedb crawl/crawldb/0 crawl/segments/0/20110329193913
CrawlDb update: starting at 2011-03-29 19:39:40
CrawlDb update: db: crawl/crawldb/0
CrawlDb update: segments: [crawl/segments/0/20110329193913]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2011-03-29 19:39:50, elapsed: 00:00:09

bin/nutch invertlinks crawl/linkdb -dir crawl/segments/0
LinkDb: starting at 2011-03-29 19:39:54
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329193913
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2011-03-29 19:40:08, elapsed: 00:00:14

bin/nutch solrindex http://localhost:8080/solr crawl/crawldb/0 crawl/linkdb crawl/segments/0/20110329193913
SolrIndexer: starting at 2011-03-29 19:40:13
SolrIndexer: finished at 2011-03-29 19:40:29, elapsed: 00:00:16

Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0/20110329193913

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        3
retry 0:        3
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        3
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
mkdir: cannot create directory crawl/crawldb/1: File exists

bin/nutch inject crawl/crawldb/1 seeds/it_seeds
Injector: starting at 2011-03-29 19:41:05
Injector: crawlDb: crawl/crawldb/1
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:41:20, elapsed: 00:00:15


generate-fetch-invertlinks-updatedb-index iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/1
rm: cannot remove crawl/crawldb/1/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/1/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/1 crawl/segments/1 -topN 1
Generator: starting at 2011-03-29 19:41:33 Generator: Selecting best-scoring urls due for fetch. Generator:
```

```
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/1 -stats
CrawlDb statistics start: crawl/crawldb/1
Statistics for CrawlDb: crawl/crawldb/1
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
mkdir: cannot create directory crawl/crawldb/2: File exists

bin/nutch inject crawl/crawldb/2 seeds/it_seeds
Injector: starting at 2011-03-29 19:42:14
Injector: crawlDb: crawl/crawldb/2
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:42:29, elapsed: 00:00:15


generate-fetch-invertlinks-updatedb-index iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/2
rm: cannot remove crawl/crawldb/2/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/2/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/2 crawl/segments/2 -topN 1
Generator: starting at 2011-03-29 19:42:42 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/2 -stats
CrawlDb statistics start: crawl/crawldb/2
Statistics for CrawlDb: crawl/crawldb/2
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
Gabriele-Kahlouts-MacBook:nutch-1.2 simpatico$ ./whole-web-crawling-incremental
rmr: cannot remove seeds/it_seeds: No such file or directory.
mkdir: cannot create directory crawl/crawldb: File exists
bin/hadoop dfs -get seeds/local-url seeds/urls-local-only

3 urls to crawl
rm: cannot remove seeds/it_seeds/urls: No such file or directory.
mkdir: cannot create directory crawl/crawldb/0: File exists

bin/nutch inject crawl/crawldb/0 seeds/it_seeds
Injector: starting at 2011-03-29 19:43:45
Injector: crawlDb: crawl/crawldb/0
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:44:01, elapsed: 00:00:16


generate-fetch-invertlinks-updatedb-index iteration 0:
Deleted file:/Users/simpatico/nutch-1.2/crawl/segments/0
rm: cannot remove crawl/crawldb/0/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/0/..locked.crc: No such file or directory.
```

```
bin/nutch generate crawl/crawldb/0 crawl/segments/0 -topN 1
Generator: starting at 2011-03-29 19:44:14 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/0 -stats
CrawlDb statistics start: crawl/crawldb/0
Statistics for CrawlDb: crawl/crawldb/0
TOTAL urls:        3
retry 0:        3
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        3
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
mkdir: cannot create directory crawl/crawldb/1: File exists

bin/nutch inject crawl/crawldb/1 seeds/it_seeds
Injector: starting at 2011-03-29 19:44:55
Injector: crawlDb: crawl/crawldb/1
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:45:11, elapsed: 00:00:15


generate-fetch-invertlinks-updatedb-index iteration 0:
rmr: cannot remove crawl/segments/1: No such file or directory.
rm: cannot remove crawl/crawldb/1/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/1/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/1 crawl/segments/1 -topN 1
Generator: starting at 2011-03-29 19:45:23 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...

bin/nutch readdb crawl/crawldb/1 -stats
CrawlDb statistics start: crawl/crawldb/1
Statistics for CrawlDb: crawl/crawldb/1
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds/urls
mkdir: cannot create directory crawl/crawldb/2: File exists

bin/nutch inject crawl/crawldb/2 seeds/it_seeds
Injector: starting at 2011-03-29 19:46:05
Injector: crawlDb: crawl/crawldb/2
Injector: urlDir: seeds/it_seeds
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2011-03-29 19:46:20, elapsed: 00:00:15


generate-fetch-invertlinks-updatedb-index iteration 0:
rmr: cannot remove crawl/segments/2: No such file or directory.
rm: cannot remove crawl/crawldb/2/.locked: No such file or directory.
rm: cannot remove crawl/crawldb/2/..locked.crc: No such file or directory.

bin/nutch generate crawl/crawldb/2 crawl/segments/2 -topN 1
Generator: starting at 2011-03-29 19:46:33 Generator: Selecting best-scoring urls due for fetch. Generator:
filtering: true Generator: normalizing: true Generator: topN: 1 Generator: jobtracker is 'local', generating
exactly one partition. Generator: 0 records selected for fetching, exiting ...
```

```
bin/nutch readdb crawl/crawldb/2 -stats
CrawlDb statistics start: crawl/crawldb/2
Statistics for CrawlDb: crawl/crawldb/2
TOTAL urls:        1
retry 0:        1
min score:        1.0
avg score:        1.0
max score:        1.0
status 2 (db_fetched):        1
CrawlDb statistics: done

Deleted file:/Users/simpatico/nutch-1.2/seeds/it_seeds
```