InjectOptions

bin/nutch inject

called java class

net.nutch.db.WebDBInjector

command line options

bin/nutch inject <db> (-urlfile <url_file> | -dmozfile <dmoz_file>) [-subset <subsetDenominator>] [-includeAdultMaterial] [-skew skew] [-noDmozDesc]

-urlfile <url_file>

Injects urls from a text file. Use a file with one url per line.

-dmozfile <dmoz_file>

Injects the urls from a dmoz content file. You can download the current content file from dmoz.org.

-subset <subsetDenominator>

Use this option if you want to inject only one of <subsetDenominator> urls. Injecting and fetching all urls from the open directory means to fetch over 4 million urls. Maybe for testing you would start with fewer urls. For example inject one out of every 4000 urls with -subset 4000, which whould be around 1000 urls injected. A random subset is selected: repeated calls with the same value will inject different urls.

-includeAdultMaterial

By default urls from the adult part of the open directory will not be included.

-skew skew

The seed for the randomization used by subsetDenominator. For debugging.

-noDmozDesc

If specified, the Open Directory description is **not** used as a link to the page.

config file options

db.score.injected

The score of new pages added by the injector. 2.0 by default.

db.default.fetch.interval

The number of days after each page injected is fetched that it should next be fetched. 30 by default.

– MatthiasJaekle - 13 Mar 2004