

CS286A Metadata Crawler Installation and Usage Guide

Kyle Dillon, Ian Juch, Eric Tu

System Requirements

- Need a LINUX machine
- At least 1 GB of RAM

Environment Setup

- install inotify-tools (version 3.14)
 - `sudo apt-get install inotify-tools` (on Ubuntu)
 - `sudo yum install inotify-tools` (on Redhat/Fedora)
 - DOES NOT WORK ON A MAC OR WINDOWS MACHINE (because inotify relies on a Linux system call)
- install other dependencies
 - `sudo apt-get install python-webpy` (version 0.37)
 - `sudo apt-get install curl`
- install Java (version "1.7.0_11")
 - need a JDK (in addition to JRE)
 - make sure `$JAVA_HOME` is set correctly
- set the correct permissions on the ssh key
 - `cd cs286A/crawler/; chmod 400 ec2.pem`
- ssh to ec2 machine and say yes (once for each new computer/account)

Code Setup

- clone repo
- `cd cs286A/crawler/gobblin/`
- `./compile_and_unpack.sh`
- `cd cs286A/crawler/`

Runtime instructions

- specify job frequency and data mover target in `config.py`
 - For demo purposes, the default frequency is set to every 30 seconds, and the default target machine/directory is the demo EC2 instance
 - if specifying a new machine besides the provided ec2 machine, you will need to get a new `.pem` key and replace the one in the `cs286A/crawler/` directory
- `./crawler.py start`
- `./crawler.py stop`