QuickStartparseChecker

Requirement

*install Java
*set JAVA_HOME
*install Apache Ant (brew install ant) if on Mac OSX, apt-get install ant if on Ubuntu/Linux

Steps

- git clone https://github.com/apache/nutch && cd nutch
- run

```
$ ant runtime && cd runtime/local/
```

- edit conf/nutch-site.xml
- add below code between <configuration> section and replace "Value_name" with the desire name

```
<property>
  <name>http.agent.name</name>
  <value>Value_name</value>
  <description>HTTP 'User-Agent' request header. MUST NOT be empty -
  please set this to a single word uniquely related to your organization.

  NOTE: You should also check other related properties:

    http.robots.agents
    http.agent.description
    http.agent.url
    http.agent.email
    http.agent.version

  and set their values appropriately.

  </description>
</property>
```

- run parsecheker for NASA JPL website for example by

```
./bin/nutch parsechecker -dumpText https://www.jpl.nasa.gov > jpl_out.txt
```