1) Breakdown of hits and HTML pages by hour (extra credit for graphing).
2) Top 20 TLD (top-level domains) by hits and HTML pages. Extra credit for adding country names

```
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '$7 ~ ".html$" && $11!="\"-\""' da-11-16.ip
ntld.log | head -20 | cut -d " " -f11
"http://discount-blah1.professional-doctor.com/"
"http://www.google.com/search?hl=en&q=use+of+data+cleaning+in+data+mining&spell=1"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/software/"
"http://search.yahoo.com/search?ei=utf-8&fr=slv1-wave&p=html+surveys"
"http://www.kdnuggets.com/dmcourse/index.html"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/"
"http://nas.cl.uh.edu/boetticher/CSCI5931%20Data%20Mining.html"
"http://www.kdnuggets.com/software/index.html"
"http://www.kdnuggets.com/software/index.html"
"http://www.google.ca/search?hl=en&q=reporting+solution&btnG=Google+Search&meta="
"http://www.kdnuggets.com/faq/data-mining.html"
"http://proj.moeaidb.gov.tw/KMPP/refererce/main.html"
"http://www.kdnuggets.com/fdsearch/search.pl?nocpp=1&Match=1&Realm=News_2004_2005&Terms=ret
rieval"
"http://www.kdnuggets.com/polls/2000/dm_tools_oct_2000.htm"
"http://www.kdnuggets.com/fdsearch/search.pl?nocpp=1&Match=1&Realm=News_2004_2005&Terms=ret
rieval"
"http://www.kdnuggets.com/index.html"
"http://www.kdnuggets.com/fdsearch/search.pl?nocpp=1&Match=1&Realm=News_2004_2005&Terms=ret
rieval"
```

3) Top 20 (most requested) HTML pages

```
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '$7 ~ ".html$"' da-11-16.ipntld.log | head
-20 | cut -d " " -f7
/gpspubs/sigkdd-kdd99-panel.html
/news/99/n23/i12.html
/dmcourse/data_mining_course/assignments/assignment-3.html
/news/2001/n10/15i.html
/aps/bt4-a.sol_crm.re.html
/news/2004/n24/18i.html
/aps/r-pur-1.c2.re.html
/news/2001/n21/28i.html
/news/2004/n24/20i.html
/news/2004/n07/15i.html
/news/2004/n24/21i.html
/dmcourse/index.html
/aps/f-salf-rf.sof_t.re.html
/jobs/2002-09-23_siemensmedical_4_software.html
/jobs/index.html
/jobs/index.html
/publications/surveys.html
/dmcourse/index.html
/news/2004/n24/23i.html
/aps/x-salf-tdm.c11_t.re.html
```

4) Top 10 external referrer sites (not from direct access or www.kdnuggets.com) by hits; also count direct entry (referrer = "-") hits.

```
File  Edit  View  Search  Terminal  Help
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '{if ($7~"html$" && $11!="\"-\"") print $11
}' da-11-16.ipntld.log | head -10 | cut -d " " -f7
"http://discount-blah1.professional-doctor.com/"
"http://www.google.com/search?hl=en&q=use+of+data+cleaning+in+data+mining&spell=1"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/software/"
"http://search.yahoo.com/search?ei=utf-8&fr=slv1-wave&p=html+surveys"
"http://www.kdnuggets.com/dmcourse/index.html"
"http://www.kdnuggets.com/"
"http://www.kdnuggets.com/"
"http://nas.cl.uh.edu/boetticher/CSCI5931%20Data%20Mining.html"
```

5) Top 10 IP addresses, including their user agent, by hits, and by HTML pages.
Top 10 based on the HTML pages

```
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '$7 ~ ".html$"' da-11-16.ipntld.log | head
-10 | cut -d " " -f1
ip1664.com
ip1115.unr
ip2283.unr
ip1946.com
ip992.unr
ip1253.com
ip2213.net
ip1886.com
ip1250.com
ip1919.com
```

Top 10 IP address based on hits

```
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '$6 ~ "GET" && $7 ~ ".html$"' da-11-16.ipnt
ld.log | head -10 | cut -d " " -f1
ip1664.com
ip1115.unr
ip2283.unr
ip1946.com
ip992.unr
ip1253.com
ip2213.net
ip1886.com
ip1250.com
ip1919.com
```

6) Top 10 most frequently not found pages (status code 404)

```
pranay@ubuntu:~/Desktop/Data Mining Week3$ gawk '$9 == 404 && $7 ~ ".html$"' da-11-16.ipntl
d.log | tail -10 | cut -d " " -f7
/sift/foil.html
/sift/alice_isoft.html
/sift/snob.html
/jobs/2005-06-17_amazon_8_technical.html
/solutions/crm.html%22%3ECRM%3C/a%3E,%20%3Ca%20href=%22solutions/web-mining.html
/websites/bioinformatics.html%22%3EBio%3C/a%3E,%20%3Ca%20href=%22websites/data-mining.html
/consulting.html%22%3EConsulting%3C/a%3E,%20%3Ca%20href=%22gpspubs/index.html
/jobs/2004-11-05_vistaprint_4_sas.html
/jobs/2005-06-10_ebay_3_quantitative.html
/jobs/2004-06-03_paradigmgenetics_4_data.html
```