# Web Mining Digital Assignment 2

**Name: Om Ashish Mishra**

**Registration Number: 16BCE0789**

**Slot: F2**

Use the Raw Web Log of a domain of your choice and perform the following:-

- Breakdown of hits and HTML pages by hour (extra credit for graphing).
- Top 20 TLD (top-level domains) by hits and HTML pages. Extra credit for adding country names
- Top 20 (most requested) HTML pages
- Top 10 external referrer sites (not from direct access) by hits; also count direct entry (referrer = "-") hits.
- Top 10 IP addresses, including their user agent, by hits, and by HTML pages.
- Top 10 most frequently not found pages (status code 404)

## Answer:

## 1. Raw Web Log Used:-

HTTP/1.1 maya.cs.depaul.edu (Google Search Link)

# Web server logs

| | |
|---|---|
| 1 | 2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://dataminingresources.blogspot.com/ |
| 2 | 2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://maya.cs.depaul.edu/~classes/cs589/papers.html |
| 3 | 2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200<br>318814 HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)<br>http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey |
| 4 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/ |
| 5 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/announce.html |
| 6 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/announce.html |

From Above Web Log

Hourly Basis:-

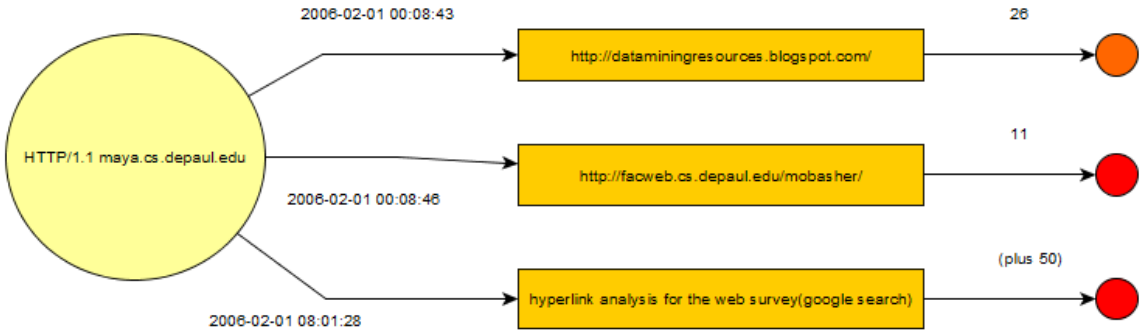
| Part 1: | Hits and HTML | |
|---|---|---|
| 2006-02-01 00:08:43 | http://dataminingresources.blogspot.com/ | (26) |
| 2006-02-01 00:08:46 | http://facweb.cs.depaul.edu/mobasher/ | (11) |
| 2006-02-01 08:01:28 | hyperlink analysis for the web survey(google search) | (50+) |

| Part 2: | Hits and HTML | |
|---|---|---|
| 2006-02-02 19:34:45 | http://facweb.cs.depaul.edu/mobasher/ | (11) |
| 2006-02-02 19:34:45 | http://facweb.cs.depaul.edu/mobasher/ | (11) |
| 2006-02-02 19:34:45 | http://facweb.cs.depaul.edu/mobasher/ | (11) |

The Brackets represent the hits took place in the particular page and therefore it shows the hyperlinks and no of download links as total.

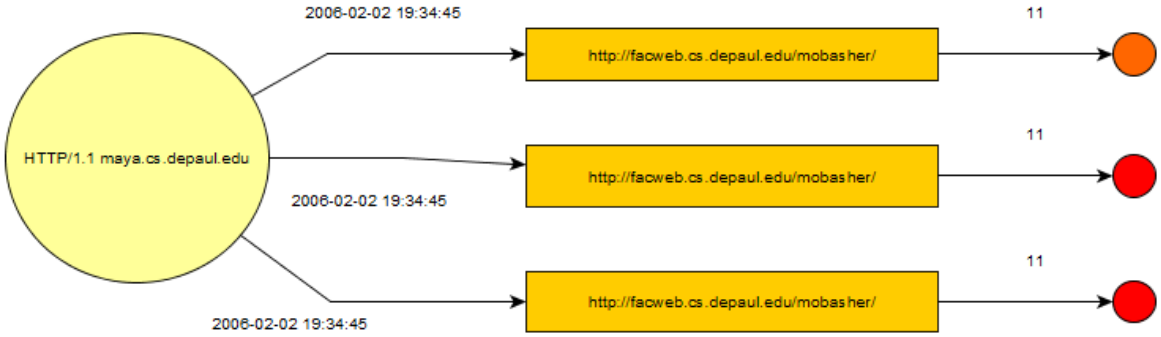
The Graph:

Part 1:




In part two the different page site is visited at the different time from 3 different people from the Raw Log.

Part 2:




In part two the same page site is visited at the same time from 3 different people from the Raw Log.

## 2. Top 20 Links available are:-

The Links are:-

2. https://books.google.co.in/books?isbn=3642194605 INDIA
   facweb.cs.depaul.edu/mobasher/classes/ect584/Lectures/12-web-usage-mining.pdf INDIA
4. www.dais.unive.it/~dm/New_Slides/10_WUM.pdf USA
5. web.iiit.ac.in/~bharath_kumar/.../data_1/Research%20Web%20Mining_id_url.txt INDIA
6. https://www.cs.helsinki.fi/u/langohr/graphmining/slides/chp4a.pdf ISREAL
7. www.jatit.org/volumes/Vol34No2/11Vol34No2.pdf INDIA
8. mail.im.tku.edu.tw/~myday/teaching/1011/WM/1011WM12_Web_Mining.pdf CHINA
9. www.ajms.co.in/sites/ajms2015/index.php/ajms/article/viewFile/515/441 INDIA
10. https://www.coursehero.com › Illinois Institute Of Technology › ITMD › ITMD 525 ILLINOIS
11. https://slideplayer.com/slide/6219382/ UK
12. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.304.5891&rep=rep1...pdf INDIA
13. shodhganga.inflibnet.ac.in/bitstream/10603/44185/16/16_references.pdf INDIA
14. https://www.slideshare.net/.../web-usage-mining-temas-avanzados ARGENTINA
15. https://www.researchgate.net/.../259575782_Fuzzy_Co-Clustering_of_Web_Documents
    INDIA
16. meri.edu.in/meri/wp-content/uploads/2017/01/Mooc-on-Weka.pdf INDIA
17. www.worldcomp-proceedings.com/proc/proc2015/DMIN15.../DMIN15_Papers.pdf INDIA
18. kb.psu.ac.th/psukb/bitstream/2553/2100/3/281913_bibli.pdf UK
19. iopscience.iop.org/article/10.1088/1757-899X/166/1/012031/pdf INDIA
20. www.cs.joensuu.fi/pages/whamalai/sciwri/sciwri.pdf FINLAND
21. nlp.uned.es/WebMining/Tema1.Introducción/ FRANCE

## 3. Top 20 HTML available are:-

**Same as above.** The pages are arranged according to SEO(Search Engine Optimization) and are made as per click bits by users and Page Ranking too. Therefore they above links represents the most visited and most requested HTML links available to get the data most accurately close to the term given or **Raw Website Log** given by the user.

## 4. Top 10 external referral sites:-

1. https://scholar.google.co.in/scholar?rlz=1C1CHZL_enIN763IN763&biw=1242&bih=597&um=1&ie=UTF-8&lr&cites=17210335240485935579 (7)
2. https://www.researchgate.net/profile/Myra_Spiliopoulou/publication/265142499_Ein_Ausblick_zum_Forschungsgebiet_Web_Usage_Mining/links/5582e4f208ae12bde6e6380d/Ein-Ausblick-zum-Forschungsgebiet-Web-Usage-Mining.pdf (6)
3. http://komputika.tk.unikom.ac.id/_s/data/jurnal/v2no2/1.aprianti-memanfaatkan-big-data-untuk-mendeteksi-emosi.pdf/pdf/1.aprianti-memanfaatkan-big-data-untuk-mendeteksi-emosi.pdf (4)
4. https://www.researchgate.net/profile/A_Sukhov2/publication/308847761_Analysis_of_Internet_service_user_audiences_for_network_security_problems/links/59f70f57aca272607e2bee41/Analysis-of-Internet-service-user-audiences-for-network-security-problems.pdf (6)
5. https://onlinelibrary.wiley.com/doi/abs/10.1002/1532-2890(2000)9999:9999%3C::AID-ASI1066%3E3.0.CO;2-Y (5)
6. https://pdfs.semanticscholar.org/1e0c/5c6fe81bf965b179edf2c53c2f5c34d2efdb.pdf (5)
7. https://hal.archives-ouvertes.fr/hal-00560096/ (4)
8. https://dl.acm.org/citation.cfm?id=1526732 (3)
9. http://www-sop.inria.fr/dias/Theses/phd-73.pdf (3)
10. https://link.springer.com/chapter/10.1007/978-3-322-89871-5_7 (4)

The Brackets show **the number of referral** to a particular page and how the Raw Website Log is used as **the citation Link** for those pages.

## 5. Top 10 IP Addresses with user agent and HITS of HTML page:-

| IP | User Agent | HITS of HTML |
|---|---|---|
| **172.217.11.14** | lga25s60-in-f14.1e100.net | 12 |
| **140.192.36.128** | facweb.cs.depaul.edu | 24 |
| 18.34.67.123 | dais.unive.it | 13 |
| **14.139.82.8** | web.iiit.ac.in | 27 |
| **66.219.22.160** | jatit.org | 34 |
| **163.13.201.222** | mail.im.tku.edu.tw | 12 |
| **104.16.148.224** | coursehero.com | 17 |
| **182.50.130.87** | ajms.co.in | 25 |
| **14.139.116.20** | shodhganga.inflibnet.ac.in | 33 |
| **108.174.11.74** | slideshare.net | 29 |

The link used for the IP address and User Agent:-

https://ipinfo.info/html/ip_checker.php

## 6. Top 10 most frequently not found pages (404):-

https://flippingbook.com/404

https://www.behance.net/gallery/46128269/404-Page-Stranger-Things

`urlOfMy404errorPage.html`

https://weeblr.com/joomla-seo-analytics-security/404-page-keep-those-lost-visitors

https://www.daar-om.nl/404-paginas-en-moet-ermee/attachment/404/

https://aradbranding.ir/%D8%B1%D9%81%D8%B9-%D8%A7%DB%8C%D8%B1%D8%A7%D8%AF-%D9%88-%D8%A7%D8%B4%D8%AA%D8%A8%D8%A7%D9%87%D8%A7%D8%AA-%D8%B3%D8%A7%DB%8C%D8%AA-%D8%A8%D8%B1%D9%86%D8%AF%DB%8C%D9%86%DA%AF/

https://churchthemes.com/page-didnt-load-google-maps-correctly/

https://builtvisible.com/optimizing-your-404-page/

https://jp.123rf.com/photo_55885004_%E3%81%8A%E3%81%A3%E3%81%A8%E5%A3%8A%E3%82%8C%E3%81%9F%E9%89%9B%E7%AD%86-404-%E3%82%A8%E3%83%A9%E3%83%BC-%E3%83%9A%E3%83%BC%E3%82%B8%E3%80%81%E3%83%99%E3%82%AF%E3%83%88%E3%83%AB-%E3%83%86%E3%83%B3%E3%83%97%E3%83%AC%E3%83%BC%E3%83%88.html

https://kateko.bg/bg/4_404/