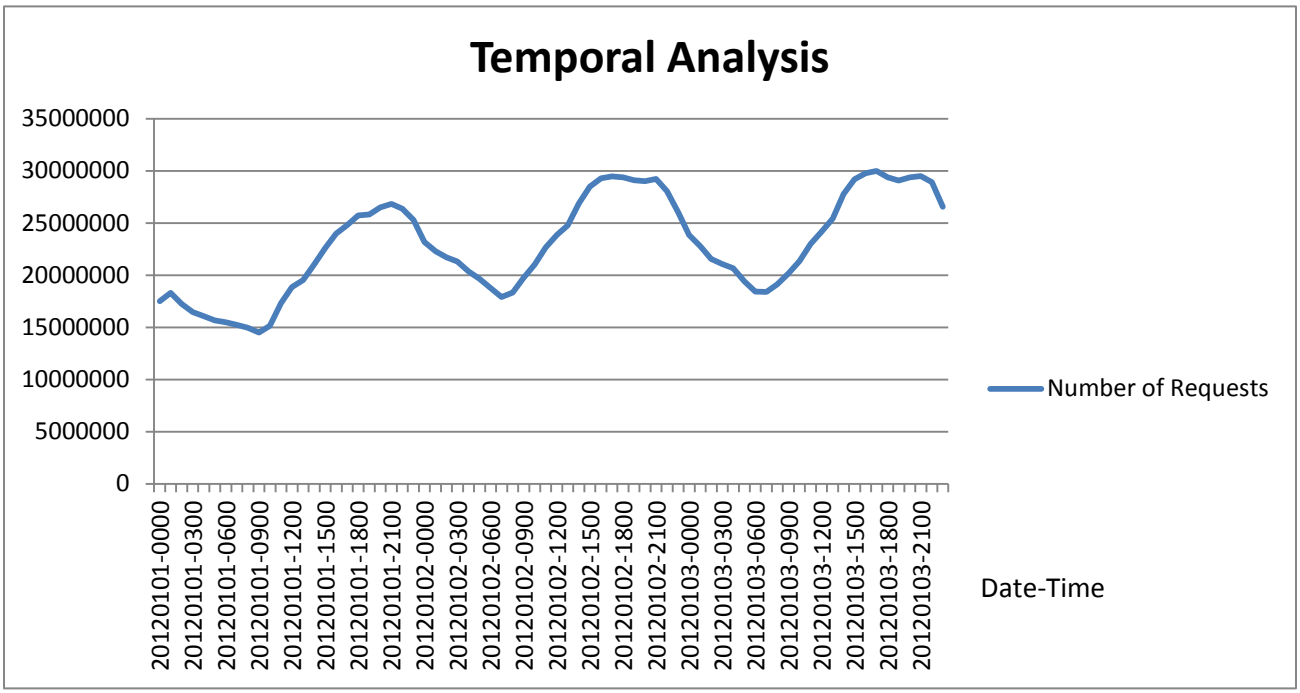
Analysis of Wikimedia Logs for Traffic Load and Popularity

Anant Pradhan

Results

1.) Temporal analysis on total number of requests per hour

The graph below shows the results obtained for number of requests per hour for the period from 01-01-2012 to 01-03-2012. These results indicate that the various Wikimedia projects experience peak traffic in the evening between 3pm and 11pm.



2.) Most popular Wikimedia project based on total views per hour per project:

The table below shows the most popular project per hour followed by the number of requests made to that project during that hour. Not surprisingly, the results show that the most popular Wikimedia project during every hour was the English language project.

en 20120101-0000	8423089	en 20120102-0000	10815033
en 20120101-0100	8580221	en 20120102-0100	10802542
en 20120101-0200	8496640	en 20120102-0200	10887829
en 20120101-0300	8217992	en 20120102-0300	10876979
en 20120101-0400	8131432	en 20120102-0400	10623767
en 20120101-0500	7844288	en 20120102-0500	10206642
en 20120101-0600	7613594	en 20120102-0600	9620189
en 20120101-0700	7217803	en 20120102-0700	8794938
en 20120101-0800	6706189	en 20120102-0800	8153236
en 20120101-0900	5988738	en 20120102-0900	7586260

en 20120101-1000 5827343 en 20120101-1100 5922991 en 20120101-1200 6297662 en 20120101-1300 6562193 en 20120101-1400 7294628 en 20120101-1500 7955580 en 20120101-1600 8794280 en 20120101-1700 9350437 en 20120101-1800 10086466 en 20120101-1900 10338036 en 20120101-2000 10817220 en 20120101-2100 11123238 en 20120101-2200 11298605 en 20120101-2300 11411446	en 20120102-1000 7516017 en 20120102-1100 7728178 en 20120102-1200 8156299 en 20120102-1300 8377859 en 20120102-1400 9098020 en 20120102-1500 9804324 en 20120102-1600 10568730 en 20120102-1700 11253286 en 20120102-1800 11741494 en 20120102-1900 11896023 en 20120102-2000 12028608 en 20120102-2100 12294397 en 20120102-2200 12229542 en 20120102-2300 11863221
en 20120103-0000 11446779 en 20120103-0100 11408114 en 20120103-0200 11064362 en 20120103-0300 10998201 en 20120103-0400 10848566 en 20120103-0500 10048841 en 20120103-0600 9239769 en 20120103-0700 8386079 en 20120103-0800 7838823 en 20120103-0900 7638054 en 20120103-1000 7814448 en 20120103-1100 8449698 en 20120103-1200 8595651 en 20120103-1300 8865926 en 20120103-1400 9820916 en 20120103-1500 10741574 en 20120103-1600 11498583 en 20120103-1700 12075122 en 20120103-1800 12122169 en 20120103-1900 12345738 en 20120103-2000 12698797 en 20120103-2100 13001706 en 20120103-2200 12894820 en 20120103-2300 12085523	

3.) Top 10 most popular pages for a given day

The table below shows the 10 most popular pages for each day, followed by the day and number of requests made to that page during that day. It is interesting how this list is almost exactly the same for all three days. Even the numbers for requests per page are very close for the same pages across the three days.

en 20120103 34139742 ja 20120103 8832817 Special:Search 20120103 4803319 Main_Page 20120103 3989070 Special:Random 20120103 3476791 de 20120103 2664225 es 20120103 2198180 fr 20120103 1919822 ru 20120103 1519483 Wikipedia:Hauptseite 20120103 1422095
en 20120102 39724275 ja 20120102 8190750 Special:Search 20120102 4629597 Main_Page 20120102 3673809 Special:Random 20120102 3545396 de 20120102 3012014 es 20120102 2256454 fr 20120102 2136550 ru 20120102 1508374 Wikipedia:Hauptseite 20120102 1413705
en 20120101 37214831 ja 20120101 7603943 Special:Search 20120101 3801542 Special:Random 20120101 3380843 de 20120101 3334472 Main_Page 20120101 3044947 fr 20120101 2282884 es 20120101 2026012 it 20120101 1534497 ru 20120101 1354442

4.) Top 10 pages that returned the most content during a given day

The table below shows 10 pages per day that returned the most content, followed by the day and total content size in bytes transmitted by that page during the day. As expected, this list largely mirrors the list above which indicates that, in general, the most popular pages return the highest amount of data.

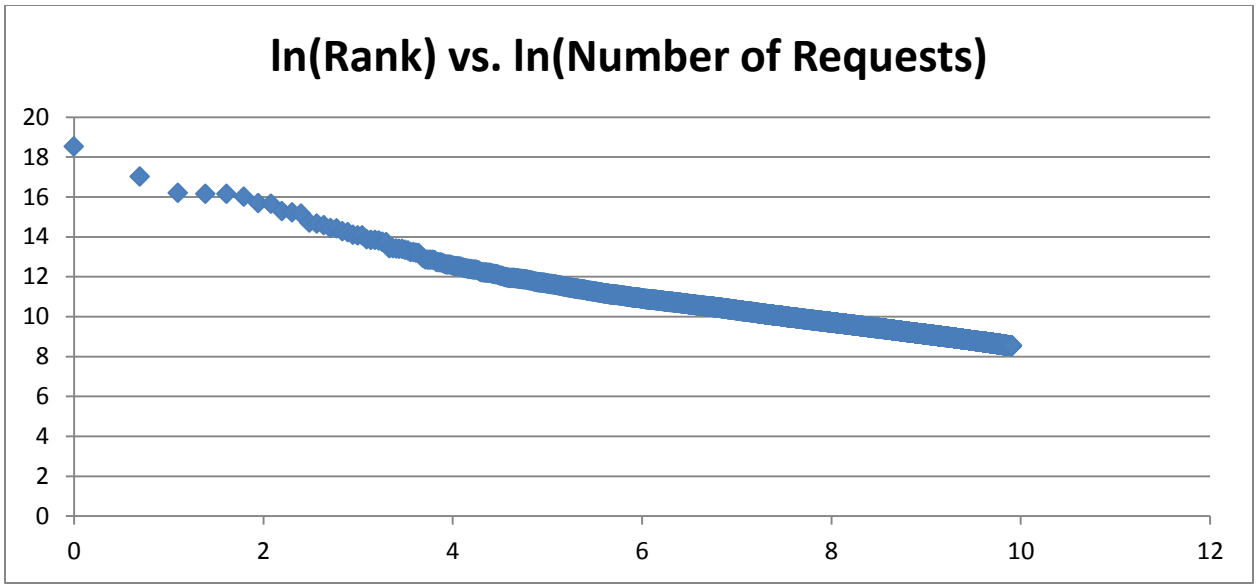
en 20120103 670178283332 ja 20120103 155316660459 Main_Page 20120103 69127197557 de 20120103 34537969687 Pagina_principale 20120103 31476492412

fr 20120103 29489184096 es 20120103 28747538141 ru 20120103 24283571907 List_of_Primeval_episodes 20120103 21177783404 Special:Search 20120103 17460030032
en 20120102 782776805885 ja 20120102 146725721599 Main_Page 20120102 64144791030 de 20120102 38716177362 fr 20120102 33062067055 Pagina_principale 20120102 31744957033 es 20120102 29362989151 ru 20120102 24244696571 it 20120102 16714244491 Special:Search 20120102 16683743790
en 20120101 746508052620 ja 20120101 136643793531 Main_Page 20120101 49467004353 de 20120101 41889693352 fr 20120101 36562097604 Pagina_principale 20120101 28720337102 es 20120101 27547608160 ru 20120101 22022286586 it 20120101 18312510792 Special:Search 20120101 13763875023

5.) Zipfian Distribution

Zipf's law states that the probability of occurrence of words or other items starts high for the most common ones and then tapers off. Thus, a few occur very often while many others occur rarely [3]. Zipf's law is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency).

To generate the Zipfian distribution, top 20000 records, out of 129 million, were used from a sorted list of the most popular pages. The graph below shows the natural logarithm of rank on x-axis versus the natural logarithm of the number of requests made to the page at that rank on y-axis. Despite the relatively small sample size, it appears that this data obeys Zipf's law since there are clearly a small number of extremely popular pages, followed by a very large number of moderately popular pages, followed by a small number of low popularity pages.



Note: The results presented above are a very small subset of the results obtained from Hadoop.