Information Retrieval and Extraction: Assignment 2

Q1. What is the number of people with access to the internet today? Do all of them have equal and unrestricted access to the internet?

Q2. How many pages/websites exist as part of the internet? What percent of them belong to the surface web and the Deep/Dark web?

Q3. What is the volume of internet traffic that flows through the internet daily?

Q4. What is the size of the internet in total data today? (Include all media content + text as well as other data in your estimation)

Q5. What are the most popular 3-4 web server frameworks for hosting websites? What is their market share?

A1) "An estimated 4.1 billion people are using the Internet in 2019, reflecting a 5.3 per cent increase compared with 2018." ~ Reference[1]
According to Reference[2], this number is around 4.7 billion.

Everyone does not have access to higher bandwidths, or mobile broadband.

From Reference[3]:
Based on the countries for which data are available, it appears that mobile phone ownership is correlated with income levels. Also, the lowest mobile phone ownership rates are found in Africa and South Asia, the highest rates are in Europe, with Latin America in between. China uses specific rubrics to monitor and control internet search results and as such, the Chinese population may not have access to a set of pages which their government does not wish them to view.

A2) 1,800,545,200+ websites exist as part of the internet according to Reference[4].

Reference[5] estimates the size of the dark web at around 100 pages. It is difficult to reach a consensus on this it seems.

[1] https://itu.foleon.com/itu/measuring-digital-development/internet-use/
[2] https://www.internetlivestats.com/
[3] https://itu.foleon.com/itu/measuring-digital-development/internet-use/
[4] https://www.internetlivestats.com/
[5] https://www.cyberscoop.com/dark-web-marketplaces-research-recorded-future/

According to Reference[6], “It's hard to estimate just how big the deep web is, but the *commonly cited research* (albeit from 2001) puts the deep web at 400 to 550 times the size of the surface web.”

“According to worldwidewebsize.com, currently, in April 2020, search engines have indexed at least 5.53 billion pages. But did you know, these 5.53 billion pages make only 4% of the whole web? Yes, the number of pages may seem to be a lot but it is actually the “surface” of the ocean across the internet.” - from Reference[7]

According to Wikipedia, ‘the Surface Web only consists of 10 percent of the information that is on the internet.’

Thus, roughly 95% of the entire web would be the deep/dark web and only about 5% would be the surface web.

A3) I used Reference[8] for estimating the amount of data flowing through the internet in real time. I calculated the amount of data passing through in a minute (note that this is just an estimate). It came to 1,449,600 GB for one minute. A day has 1440 minutes. Therefore, 1,449,600 x 1440 GB = 2087424000 GB = 2.08 EB. Cisco places the estimate at around 5.3 EB/day. Thus we can safely guess that it is of the order of a few exabytes.
Refer here for the Cisco report: Reference[9]

A4) According to the post here at Reference[10]: “In 2017, we reported that there was 2.7 Zettabytes (ZB) of data in our digital universe. PwC believes that this reached 4.4 ZB in 2019, but more staggeringly predicts that this will grow to 44ZB of data this year (2020). In fact, IDC predicts the world’s data will grow to 175 ZB by 2025!”

Thus, the size of the web can be estimated to be of the order of 10s of Zettabytes.

[6] https://www.cnet.com/news/darknet-dark-web-101-your-guide-to-the-badlands-of-the-internet-tor-bitcoin/

[7] https://www.kratikal.com/blog/surface-web-and-dark-web-exploring-layers-of-web/

[8] https://www.webfx.com/internet-real-time/

[9] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiTzbyxy_DrAhXljuYKHX2kA6gQFjABegQIChAD&url=https%3A%2F%2Fwww.cisco.com%2Fc%2Fdam%2Fm%2Fen_us%2Fsolutions%2Fservice-provider%2Fvni-forecast-highlights%2Fpdf%2FGlobal_2020_Forecast_Highlights.pdf&usg=AOvVaw08PEQ_P2Ci3PmaVwLQqtLe

[10] https://www.nodegraph.se/how-much-data-is-on-the-internet/

A5) According to Reference[11], the present most popular web servers with corresponding market shares are as follows:

1. nginx: 36.55%
2. Apache: 25.45%
3. Microsoft: 11.36%
4. Google: 3.59%

According to Reference[12], it is as follows:

1. Apache: 36.3%
2. Nginx: 32.4%
3. Cloudfare server 15.9%
4. Microsoft-IIS 7.8%
5. LiteSpeed 7.1%
6. Google Servers 1.2%

- This is not the most current version. Netcraft's time series data seems more reliable here.

BONUS)
Metric to estimate size of the web: The metric will depend highly on what we need to use it for. I have an idea relating the size of the web to the amount of actual available knowledge. We know that we extract information from data, and assimilated information leads to knowledge. A schema which could potentially index the entire web (including deep and dark web) and build a knowledge graph with all possible annotations, and with references to all sorts of digital information would serve as a ginormous domain independent knowledge base. The only catch here is that there should be no redundancy and and repetitions whatsoever. Now, the amount of disk space required to store such a knowledge base would be my 'size of the web metric'.

Recall the notion of tokens and spans of text and how it relates to 'sense'. This metric in a way captures how much space is required to store the full knowledge available on the web OR what is the required size of the knowledge index such that the web itself makes perfect 'sense'.

[11] https://news.netcraft.com/archives/2020/08/26/august-2020-web-server-survey.html
[12] https://w3techs.com/technologies/overview/web_server