Assignment P1

CS 6675: Advanced Internet Systems and Applications
Mher Arabian

Question 1

The three Internet-fueled innovations:

a. **Cloud Computing (Transforming the physical world to cyber world)**
Cloud Computing has played a major role in driving technological innovation in recent years. Individuals and companies no longer have to possess the technical expertise/have the required physical infrastructure to set up their own web servers. This enables engineers/companies around the world to quickly setup and deploy web apps on cloud providers such as GCP, AWS, and Azure. It has also allowed us to scale up our technologies much more efficiently; this has played a major role in the development of software technologies in the past decade (especially in the field of AI in recent years).

b. Smart-Watches (Combining the physical world and cyber world)
IoT devices like smart-watches, are a good example of technologies which have combined the physical and cyber worlds. They allow us to monitor our heart rate and sleep schedule, track our exercises; some even have the capability to automatically collect health data for healthcare professionals to monitor, or call emergency services in the event of a fall.

c. Community-based Chat Platforms (Enriching the physical world)
Platforms like Discord enhance physical interaction between people using online communities, where people interact with each other with chat/voice/video calls and create new physical experiences.

What they have in common:

- They are all trying to automate a task for the user in some way.. In the case of cloud computing, it's automating the need to do DevOps. For smart-watches, it's automating manual exercise tracking/health monitoring. In the case of Discord, it makes it easier for users to find like minded individuals, through its curated/labeled online communities.
- Interestingly all three involve cloud computing in some way.

Differences:

- Although smart-watches have a software component, they are actually a hardware innovation as well unlike the other two.
- Cloud Computing can be seen as a disruptive technology; but smart-watches are an entirely new type of device that don't necessarily replace traditional watches. Interestingly, they are still worn for style, fashion, or sentimental reasons.

Question 2

Surface Web Activity: Visiting the Wikipedia page entry for a historical figure to for educational purposes.

Deep Web Activity: Going on Amazon.com and querying for school supplies, or furniture.

Most deep web pages are hidden behind search/login forms. They are not served statically/hosted on one of the Internets' servers. Thus, they cannot be crawled and indexed by modern search engines. For example, when searching for school supplies on Amazon.com, it returns a unique dynamic URL that a crawler could not find/reach.

Building a deep web search engine would require crawling the deep web. Since many URLs are not statically served and locked behind login forms, it is not feasible to build a useful deep web crawler/search engine. One possibility is to allow websites to provide some of their most commonly visited dynamic (search) URLs (e.g. common product queries on Amazon.com). This would require public Internet infrastructure to be set up, in order to allow individuals/companies to make their dynamic URLs publicly available for crawlers to consume. Another potential solution is to allow website creators to provide guidance on how to access their dynamic URLs via their Robot.txt files.

Question 3

One limitation with the PULSE Crawler Project is that for every page, only the top 1000 characters are being crawled. The rest of the documents are being dropped. This achieves fasting crawling speed per page, but is not realistic in a real-world setting. This would effectively greatly reduce the search space of available URLs to crawl, since the system would ignore URLs that would be found past the 1000 character count.
The depth-first crawling strategy with a maximum depth of 25 is also a major constraint, since it never fully goes down to explore the breadth of the website's structure,

potentially missing critical information and links that could be crucial for a comprehensive understanding or index of the site.

To design a crawler for www.gatech.edu, I would approach it the same eway PULSE Crawler did, as his method seems to be a good demonstration on the capabilities of crawlers for small websites. Instead of prioritizing # pages/per time unit, I would focus on covering a wide range of URLs from the university domain. For this task, I would use BFS (breadth-first search), since it systematically covers all information on the website at each level; this ensures a well-covered crawl.

To achieve crawling 1 billion pages per year, the system would need to be able to crawl roughly 32 pages per second, which seems very demanding, even on a personal computing device.