Qualifier Exam for Debjyoti Paul

Problem 1

Social media data is enormous, but semi-private. List relevant social media data sources, and explain what is known about their sizes (in terms of storage space, and number of records), including both what is (probably) privately controlled by companies, and what is available for sufficiently-motivated and -resourced academic researchers.

Explain the state-of-the-art (with references to research papers) for scraping such semi-publicly accessible data sets, and what are the largest bottlenecks for such tasks.

Predict (using a machine learning / data mining techniques on the data above) what the total number of social media records available to researchers will be in 2022.