Assignment 1 Analysis and Reporting

1. **Hardware Configuration:**

   CPU: 3.1 GHz Dual-Core Intel Core i5
   RAM: 8GB

2. **Data Modeling Assumptions:**

   Number of users: 10,000 (user_id from 1000001 to 1010000)
   Number of tweets per user: Randomly distributed
   Distribution of the number of followers per user:

   (a) For user_id between 1000001 and 1001000 (1,000 users): follows 100
   (b) For user_id between 1001001 and 1002000 (1,000 users): follows 300
   (c) For user_id between 1002001 and 1003000 (1,000 users): follows 500
   (d) For user_id between 1003001 and 1001000 (7,000 users): follows 10

   Total follows relation: 970,000 lines

   Time Stamp: Between 2021-01-01 and 2021-01-31
   Tweet_text: Randomly selected from upper and lower case letters and all punctuations with random length

3. **Results:**

   The number of tweets inserted per second:

   It takes 397.83 seconds to insert 1 million tweets. Since each tweet has different length of tweet_text, the inserting time may vary. Therefore, I calculated the average to minimize the error. The speed is **2513 tweets per second**.

   The number of random home timelines retrieved per second:

   (a)

   (1) Only retrieve tweet_id and tweet_ts (10 most recent):

   1 user: 0.34 second
   10 users: 3.25 seconds avg: 0.325 second
   30 users: 9.60 seconds avg: 0.32 second
   50 users: 16.31 seconds avg: 0.326 second

   By retrieving different number of users and calculating the average time to retrieve one user's timeline, the retrieval speed is about **0.323 second/user**

   (2) Retrieve tweet_id, tweet_ts, and tweet_text (10 most recent):

   1 user: 0.35 second
   10 users: 3.32 seconds avg: 0.332 second
   30 users: 9.65 seconds avg: 0.32 second
   50 users: 16.29 seconds avg: 0.325 second

   By retrieving different number of users and calculating the average time to retrieve one user's timeline, the retrieval speed is about **0.325 second/user**

(b)

(1) Only retrieve tweet_id and tweet_ts (20 most recent):

1 user: 0.327 second
10 users: 3.328 seconds avg: 0.328 second
30 users: 10.19 seconds avg: 0.34 second
50 users: 16.80 seconds avg: 0.336second

By retrieving different number of users and calculating the average time to retrieve one user's timeline, the retrieval speed is about **0.333 second/user**

(2) Retrieve tweet_id, tweet_ts, and tweet_text (20 most recent):

1 user: 0.374 second
10 users: 3.54 seconds avg: 0.354 second
30 users: 10.786 seconds avg: 0.36 second
50 users: 17.03 seconds avg: 0.34 second

By retrieving different number of users and calculating the average time to retrieve one user's timeline, the retrieval speed is about **0.357 second/user**

4. Analysis

	Retrieve without text (10 most recent)	**Retrieve with text (10 most recent)**	**Retrieve without text (20 most recent)**	**Retrieve with text (20 most recent)**
Speed (second/user)	0.323	0.325	0.333	0.357

Clearly, from the table we can see that when retrieve 10 most recent timelines, retrieve with or without tweet_text doesn't really affect the retrieval speed. When retrieve 20 most recent timelines, the speed difference is about 0.02 second for two different retrieval ways. Speed for retrieving 20 most recent timelines is slower than the 10.