Twitter dataset analysis

1. Locality and number of followers

Assumption

For the people who have large number of followers (like public figures), their followers will be more distributed, hence we need to write to all Datacenters for that kind of user.
The proof of this assumption contains two parts. Firstly, we need to prove that the followers number is different. Actually, the majority of users only have few hundreds of followers (figure 1) and only very few of users have number of followers more than one million (figure 2). Secondly, among those different followers numbers, we need to prove that the user with more followers, their follower region(state) is more distributed(figure 3).

Figure 1: the relationship between the follower number and the number of users who have that number of followers.

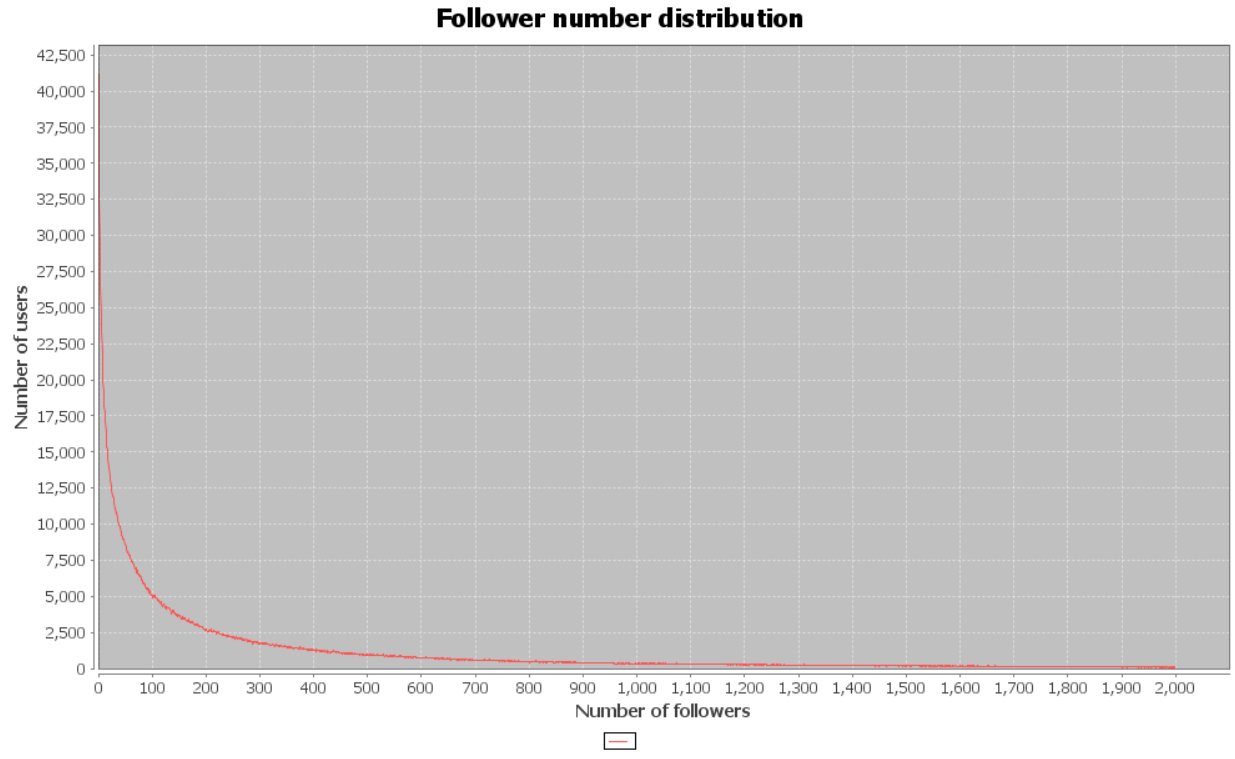


Figure 2: the relationship between the minimum follower number and the number of users whose follower is bigger than the minimum number.

In figure 2, the number of total users analyzed is 2633372, the number of users that at least have 10 followers is 2356424, the number of users at least have 100 followers is 1536687, the number of users at least have 1000 followers is 396813, the number of users at least have 10000 followers is 39861, the number

of users at least have 100000 followers is 3220 and the number of users at least have 1000000 followers is only 367.

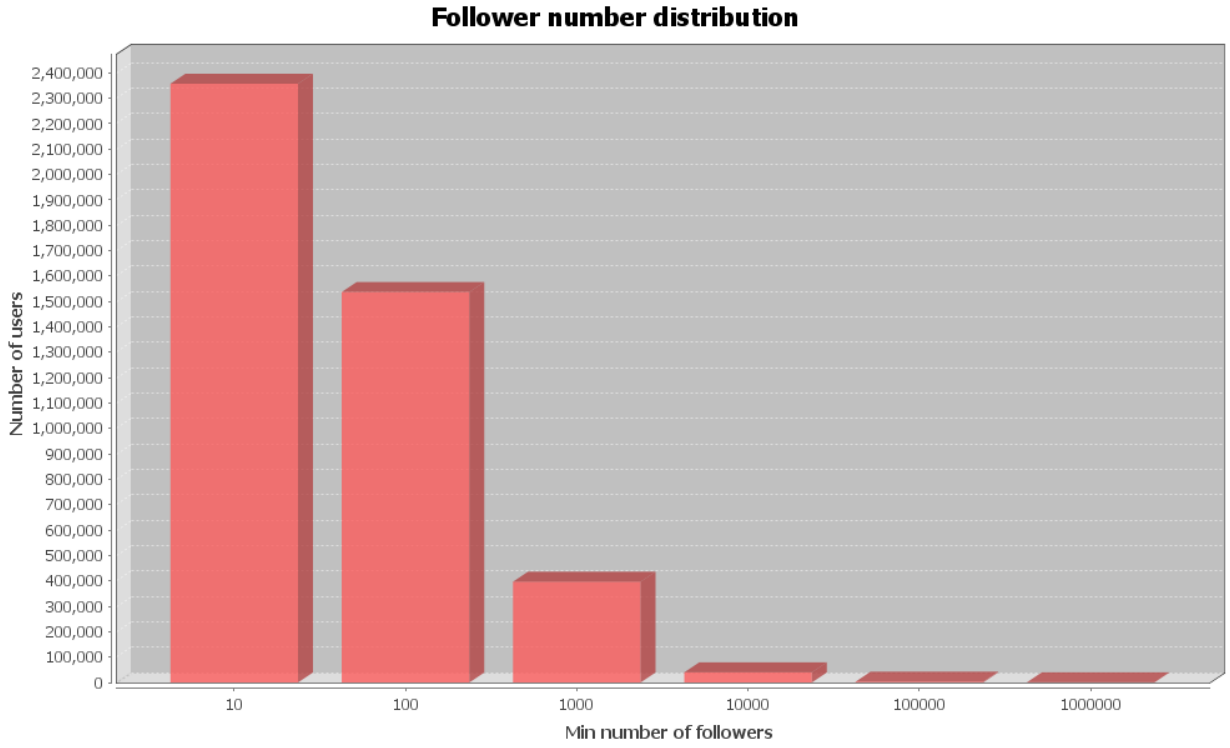
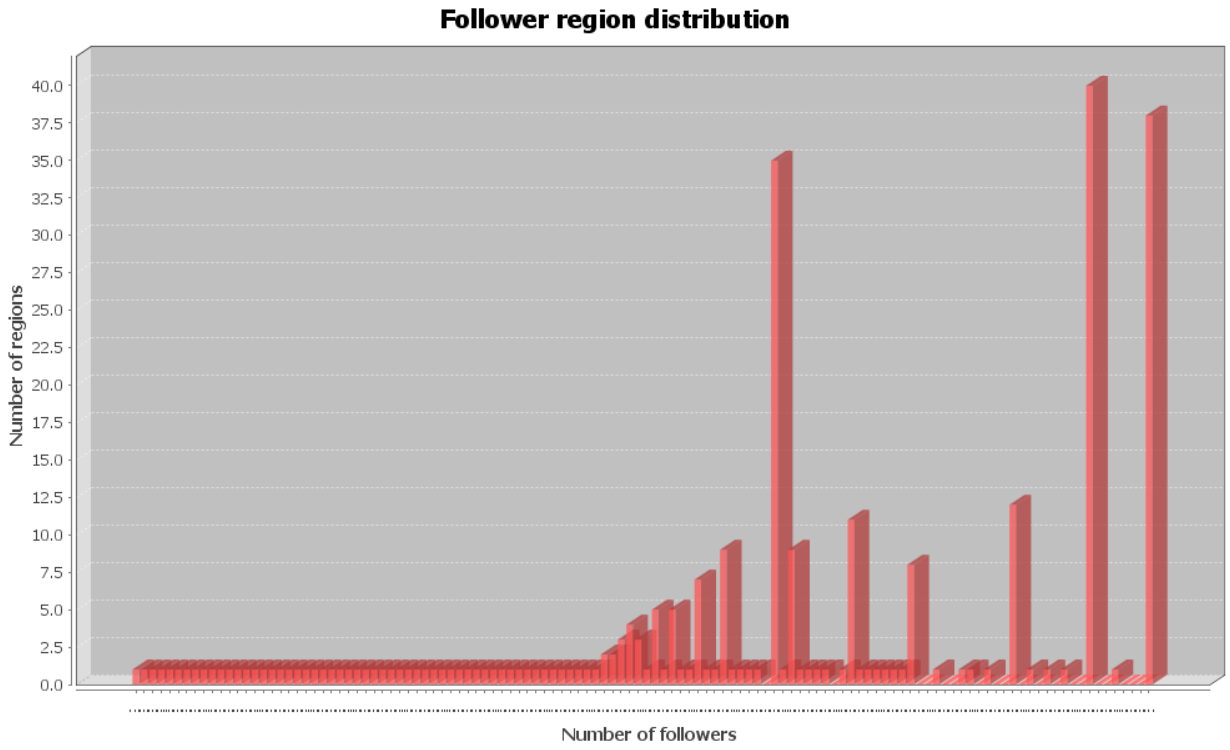


Figure 3 the relationship between number of followers and the number of follower regions distributed.

From figure 3, we can find that in most of cases, if there are more followers for a user, then the followers will be come from more different regions.

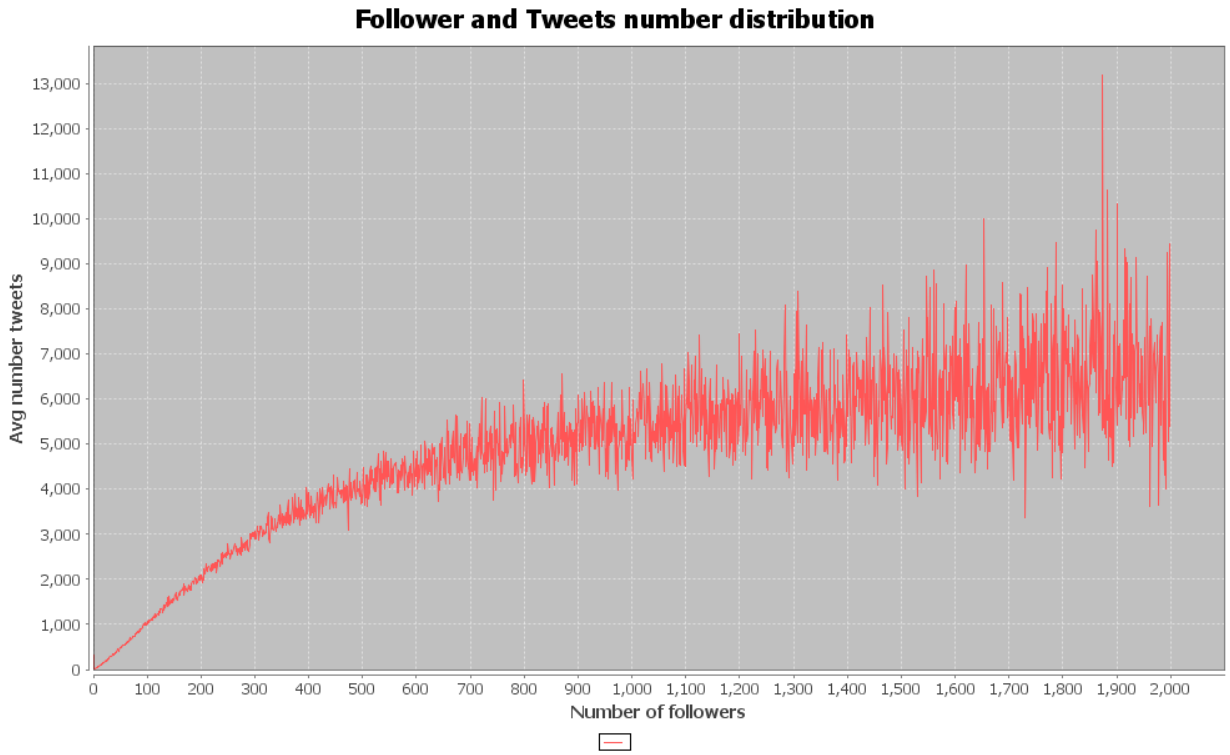


2. Number of tweets and number of followers

Assumption

If there are more followers for a user, there will be more tweets created by that user.
This assumption is that the one who is read intensive by others (who has many followers) is also write intensive (who will publish more tweets). Such as public figures or some news media.

Figure 4 the relationship between number of followers of users and the average tweets created by the users. (0-2000 followers)

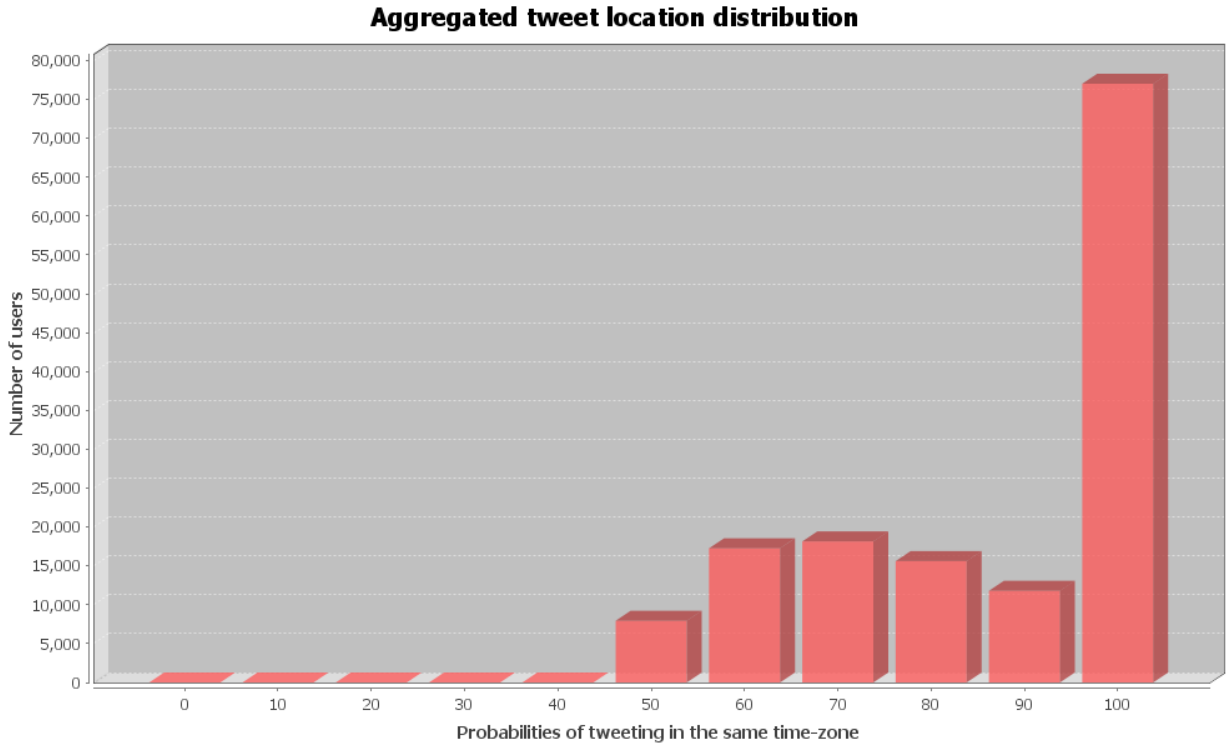


3. Users are always tweet from the same Data center.

Assumption

This assumption is that writes are always happens in the same datacenter. This assumption is for the migratable leader which we assume the leader is stable at most of the time.

Figure 5 the probability of the user tweet in the same region


Aggregated tweet location distribution
Number of users
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0
10
20
30
40
50
60
70
80
90
100
Probabilities of tweeting in the same time-zone