Assignment 1 : Top-kSimilarPair using PThread

Input: A file that contain data. Each line is one record and each record is a comma-separated values “Object_id,Attr_1,Attr_2,Attr_3,…,Attr_n”.

For example,

```
BJP,Politics,Organization,
Modi,Person,Politics,Politician,PM,
Sachin,Cricketer,Bowler,Person
Rahul,Person,Politician,
…
```

Explanation of Input File
BJP is first object with two attributes: Politics and Organization
Modi is second object with four attributes Person, Politics,Politician and PM.
…

Given a pair of two objects, o1 and o2, the **Jaccard Similarity** calculates the number of attributes common in both the objects divided by the union of object attributes. For example, let us consider BJP and Modi:

Common attributes: {Politics}
Att attributes: {Person,Politics,Politician,PM,Organization}

Hence, Jaccard Similarity(o1,o2) = 1/5 = 0.20

Problem. The problem of discovering **Top-kSimilarPair** focus on discovering k pair of objects such that their similarity value is very high. In particualr, Given a *k* (=5 or 10 or 20 or 50 or 100), you need to write an algorithm to output *k* object pairs (Oi,Oj), Oi != Oj, such that, their similarity value lie in top k list.

1. Write Sequential Algorithm
2. Write Parallel Algorithm using Pthread

Compare the performance of both the algorithm.

Hint. The simple solution is to first discover the similarity between all pairs of the objects, and then identify the k pairs with the maximum value. However, there is an hidden optimization, You can think !!!!.