Linc Coding Test

We plan to build a product recommendation engine based on the similarity of purchase histories among shoppers. As part of a pilot project, you will use the Jaccard index to measure the similarity between two shoppers' purchase histories. The Jaccard index is defined as the size of the intersection divided by the size of the union of two sets:

$$J(A, B) = |A \cap B| \div |A \cup B|$$

For example:

$$J(\{3, 5, 7, 9\}, \{3, 6, 9\}) = |\{3, 9\}| \div |\{3, 5, 6, 7, 9\}| = 2 \div 5 = 0.4$$

Given a purchase history H_X and a product P where $P \notin H_X$, we then define their *recommendation index* as the sum of $J(H_X, H_i)$ for every H_i where $P \in H_i$, divided by n:

$$R(H_X, P) = (J(H_X, H_1) + J(H_X, H_2) + \dots + J(H_X, H_n)) \div n$$

For example, assume we have purchase history H_A, H_B, H_C, H_D and product 1 to 6:

$$H_A = \{1, 2, 3, 4\}, H_B = \{2, 3, 4, 5\}, H_C = \{1, 3, 5\}, H_D = \{2, 4, 6\}$$
$$R(H_A, 5) = (J(H_A, H_B) + J(H_A, H_C)) \div 2$$
$$R(H_A, 6) = J(H_A, H_D) \div 1$$
$R(H_A, 1)$ to $R(H_A, 4)$ are undefined

Products with higher recommendation indexes may be recommended to shoppers.

In addition to the above indexes, measurements like product popularity or purchase recency may also be taken into account, but it's out of our current scope.

Your Assignment

Write a program to process the purchase log in data.txt. Ignore repetitive purchase of the same product in a shopper's purchase history for now.

1. Find the two shoppers with the highest Jaccard index.
2. Find the top 3 products we should recommend to shopper "andrew".

Remember that we are a startup, so both the coding speed and the code quality matter.