Programming Assignment 10

The Jacquard index of two finite sets A, B is defined as $\frac{|A \cap B|}{|A \cup B|}$ (the ratio of the number of elements in the intersection over the number of elements in the union) and is a statistical measure for the similarity between two sets.

You are given a large csv file with the internet habits of 12,000 users. Use the enclosed Python implementation of LH to store a pair consisting of a website name and the list of users visiting it in a list indexed.

For example, Bucket 5 will contain a single entry, namely
`[('sophos.com', [10305, 7439, 1353, 9838, 3177, 7162, …])]`
Bucket 2 will contain two entries, namely
`[('linkedin.com', [6916, 5431, 4662, 5796, 8219, 8576, 11282, …], (, ('marquette.edu', [2479, 1677, 4297, 4190, 8858, …])].`
Here, I suppressed quite a number of user identifiers in the list.

Provides functions to the user that:

- Give the list of all user identifiers that accessed a given web-site.
- Calculate the number of different users that have accessed a given web-site.
- Calculates the Jacquard index of the user identifiers of two different web-sites.

You will notice that while the number of visitors to a website are not identical (see right), the Jacquard indices are almost constant, showing that the file is not a true trace, but the result of a poor simulation.