HW-1.

In this assignment you have to compute relative frequency (likelihood) of each word given other word. Input is a text corpus and a parameter defining the window size. For example assume the text corpus contains only one document with the following text:

“BDA is an advanced course in which l am learning algorithms and technologies to manage huge data sets. The data set size can be in peta bytes. The course is of 4 credits.”

If the window size is 10 words. The relative frequency should consider 9 words in the left and 9 words in the right for a given target word. Let the target word be “course”, “BDA” is the left most word which is in the context and “to” is the word in right that will be in the context. We count each pair, consisting of the word in the context and target word, for computing the relative frequency. Therefore the pairs (BDA,course), (is, course)....., (technologies, course) will be counted. Relative frequency of “BDA” given “course” is 1/2. Relative frequency of “is” given “course” is 1.

You need to write a Apache Spark program to compute the relative frequencies.
Part-1: It has to be implemented on your machine.

Input: A corpus of 10 text documents. You can create these documents yourself.
Output: A file describing the top-100 relative frequencies. Each line will have a relative frequency of one word given other word.

Part-2: Deploy your program on Institute’s Apache Spark cluster and measure the performance for large input such as wikipedia. Report results with varying computing resources, number of cores, RAM, and dataset size.

Deadline: After 2 weeks, 15th Feburary 2019.