Page Rank Algorithm

PageRank is an algorithm used by Google to rank the web pages for the search results on Google. The basic idea behind this ranking is that the pages that have a lot of links coming from other pages should be more important than others. Before continuing of the explanation here are some definitions that will be useful.

- Inlink: A reference that a page recieves from another webpage. For example, if the Webpage A has a link to webpage B at somepoint we say B has an inlink from A
- Outlink: A reference to another webpage. For instance, in the previous case, we could also say A has an outlink to B.

The ranks are calculated within a certain amount of iteration, in my example, there are 10 iterations. We are given 2 text file named inlinks and titles.

- Each line on inlinks has a number followed by a colon (:) and at least 1 or more number. This implies the numbers after the colon has inlinks from the first number or the first number has outlinks to numbers that follow the colon. For example 2: 3,4,5 means 2 has a link to 3,4 and 5.
- The titles text file has the string name of each page ordered from 1 to N., For example, ABCDE means A is the 1st B is the 2nd page etc..

Starting from an initial rank for each page, at each iteration we update the rank of each page based on how many inlinks they have but also weighing the contribution of each inlinks by the number of outlinks they have. We also use a damping factor of $d = 0.85$ Here are the formal steps for the algorithm:

1. Initialize PR_0 at $\frac{100}{N}$ for each page i
2. At each iteration for each page update:

   $PR_i = 100\frac{1-d}{N} + d\sum_{j \in I_i} \frac{PR_l}{\#Out_j}$ Where:
   I_i: is the set of pages that the page i has inlinks from and

   $\#Out_j$: is the number of outlinks each page j has.
3. After the final iteration normalizes the outputs to have a sum of 100.