Clarifying PageRank

I've received a few questions about PageRank and about the quiz, so this document is an attempt to clarify things.

I think part of the confusion stems from the fact that there are different PageRank definitions, as well as different ways to implement those definitions.

- Let variables u and v to refer to web pages,
- Let $B(u)$ be the "back links" of u; that is, the set of pages that link to u
- Let $F(u)$ be the "forward links" of u; that is, the set of pages that u links to

So, for a graph with pages $\{A, B, C\}$ and edges from $A \to \{B\}$, $B \to \{C\}$, and $C \to \{A, B\}$, we have:

- $B(A) = \{C\}, F(A) = \{B\}$
- $B(B) = \{A, C\}, F(B) = \{C\}$
- $B(C) = \{B\}, F(C) = \{A, B\}$

Simple PageRank, Synchronous Updates

Let's start with the simplest pagerank score, which we will call r_s (slide 19 from L23):

$$r_s(u) = \sum_{v \in B(u)} \frac{r_s(v)}{|F(v)|}$$

So, the score for page u is the sum of the scores for each page v that points to u, where each of these v scores are divided by the number of forward links of v. (Conceptually, the score emerging from v is divided equally among all the pages v points to.)

This is what I was asking you to compute on the last quiz. After initializing all scores to 1, we have:

$$\begin{aligned}
r_s(A) &= \frac{r_s(A)}{2} = \frac{1}{2} \\
r_s(B) &= \frac{r_s(A)}{1} + \frac{r_s(C)}{2} = 1 + \frac{1}{2} = \frac{3}{2} \\
r_s(C) &= \frac{r_s(B)}{1} = \frac{1}{1} = 1
\end{aligned}$$

Simple PageRank, Synchronous Updates

Note that the above computation uses *synchronous* updates, which means that to update all scores at iteration i, we use the scores from the prior iteration $i - 1$.

Alternatively, we can do an *asynchronous* update, which means that when updating scores at iteration i, we may use scores for nodes that have already been recomputed this iteration. Let's denote this function r_a. For example, if we update the scores in the same order above (A,B,C), then the asynchronous update results in:

$$\begin{aligned}
r_a(A) &= \frac{r_a(A)}{2} = \frac{1}{2} \\
r_a(B) &= \frac{r_a(A)}{1} + \frac{r_a(C)}{2} = \frac{\mathbf{\frac{1}{2}}}{\mathbf{1}} + \frac{1}{2} = 1 \\
r_a(C) &= \frac{r_a(B)}{1} = \frac{1}{1} = 1
\end{aligned}$$

Note that the value for $r_a(B)$ differs from the earlier value $r_s(B)$, since here we use the latest PageRank value for A when computing the update for B. (We can add superscripts like r_a^t to indicate iteration t, but I'm trying to keep the notation simpler.)

Note that in your final assignment, you should do an asynchronous update.

Teleport PageRank, Synchronous Updates

Finally, we can use the "random teleport" idea to deal with spider traps and deadends. This adds a new parameter β to interpolate the values of the simple pagerank with the teleport value. Let's call the resulting function r_t (for *t*eleport):

$$r_t(u) = (1-\beta)\frac{1}{N} + \sum_{v \in B(u)} \beta \frac{r_t(v)}{|F(v)|}$$

where N is the total number of pages

(See slide 44 of L23)

For our running example, if we initialize values to 1 and set $\beta = .8$, we get:

$$\begin{aligned}
r_t(A) &= (.2)\frac{1}{3} + (.8)\frac{r_t(A)}{2} = .466... \\
r_t(B) &= (.2)\frac{1}{3} + (.8)\frac{r_t(A)}{1} + (.8)\frac{r_t(C)}{2} = 1.266... \\
r_t(C) &= (.2)\frac{1}{3} + (.8)\frac{r_t(B)}{1} = .866...
\end{aligned}$$

Initialization and Normalization

When implementing any of these functions, we also need to decide how to initialize the values and how to normalize them. In the examples above, we initialized all values to 1. In practice, we would often initialize them to $\frac{1}{N}$. (I used one to make the examples simpler.)

Additionally, after each iteration, we would typically normalize the values. We can do this in one of two ways:

- Divide by the sum; e.g., $\sum_u r(u)$
- Divide by the square-root of the sum of squares: $\sqrt{\sum_u r(u)^2}$

Either is acceptable.

On the final, I'll be sure to specify these details if I ask you to compute any of these.