CSE199 Internet and Data Homework 1 Fall 2023

Internet Search

The lecture materials on the "Internet and Data" unit give short shrift to the major topic of Internet search engines—Google's in particular. This homework aims to remedy that. Because this unit is less specific and more expository than others, this homework—being due in the middle of the first of two weeks—is also largely independent of the lectures. It is graded out of 9 pts.; the score may be converted to one out of 3 pts. [**Date fixed to 2023.**]

The three main ingredients of an Internet search engine are:

1. A comprehensive index of publicly-accessible webpages. Needs computer space to store it and computer power to process it.
2. A basic scoring metric of the relevance of a particular webpage to a particular search query.
3. An algorithm for further processing of webpages and their scores.

Although the details of Google's algorithm—for point 3 in particular—have changed significantly over the quarter century since its inception, the essence is preserved enough for the algorithm to retain its original name, *PageRank*, even though the patents under that name have expired. By neat coincidence, the conceiver of this essence was named (Larry) Page. There are other algorithms, but Google's is definitely the elephant in the room. Like the 2,500-year-old parable of the blind men and the elephant, different sources will give you different takes on what this essence *is*. Finding it is the first objective of this homework.

Part 1

Enter simply `PageRank` into Google. One word, no space, capital `R`. Chances are your top hit will be Wikipedia's article of that name. **Do not** read it. The objective is to read the takes on *PageRank* by people trying to be simpler and crisper than Wikipedia. And we will go further by trying to pull just a few words off each one that characterize its take on *PageRank*. Your first two pages of hits will likely include the following, not necessarily in this order:

- A page from `Semrush.com`, a large company that sells tools to businesses to enhance traffic to their webpages.
- A page from `searchengineland.com`.
- A page from `Geeks for Geeks`, which is a coder's version of "How Things Work."
- A page from `AHrefs.com` (a company like Semrush) saying "PageRank is NOT Dead."
- An article explaining it from Amrani Armine of `towarddatascience.com`.
- A three-page handout from Stanford's CS54N class.
- A one-page definition from `WhatIs.com`.

There will also be some PageRank checker apps—we can ignore those—and some more technical sources. You are welcome to consult the latter as well if you wish.[1]

Your task: Pick *five* of the above or similar hits (again, not Wikipedia), and for each one, find *three* short phrases or single words that encapsulate how it describes *PageRank*. Your words can be quoted verbatim or paraphrased. If a concept like "Markov chain" seems to be important but you don't know what it is, you can quote it without having to look it up. If you see an equation of the form $PR(A) = \ldots$, pick a phrase on the page that describes what the equation does. For example, here is a more-technical hit I did not include above: https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/, for which one might say:

- Neo4j: (a) "A page is only as important as the pages that link to it." (b) "equation is used to iteratively update a candidate solution." (c) Bad stuff occurs when groups of pages have no outgoing links.

Then write a short paragraph as an "executive summary" of what you get about *PageRank* from these sources. If terms that are common to two or more of your sources strike you as important, be sure to include them.

Finally, give your vote as to which of the hits (not Wikipedia's) is the best single source to gain a quick and serviceable understanding of *PageRank*. (The "meta" aspect here is your functioning as a human page-ranker. 6 pts. total: 3 for lists and 3 for paragraph and vote.)

Part 2

This is a short application, to see if the understanding gained from part 1 is enough for insight. In my original 2017 edition of the "Internet and Data" slides, I felt on firm enough ground to quote "30%" as the percentage of Web *traffic* consisting of pornography. I linked a secondary source https://ourworldindata.org/internet, from which all mention of porn seems now to have disappeared, but I noted its reference to a 2012 article by Sebastian Anthony of Extreme Tech. Let's see if this figure is still tenable now a further five years from its origination.

Content Note: The assignment has been structured to avoid getting sexual content in hits, but my runs turned up a cartoon and infographic with some "sketchy" elements. The cartoon, from `straightdope.com` with 2005 date, depicts people dressed for "S&M" in a silly way. The infographic from `PaintBottle` (which was a porn site, now apparently vanished) is included in whole or part by a *Huffington Post* hit and by *IT Voice* and *Digit News* in India, both with 2013 date. It has racy-but-fully-clothed cartoon figures and some suggestive (but not obscene) language. The dates and the fact of the infographic having the line "30% of all the data transferred across the Internet is porn" are all you need to note, so you need not click on those hits. Search results are variable by person and time, however; if anything else/worse comes up, please let us know. The words `traffic` and `bandwidth` are synonymous

[1] I'll add that my own PhD graduate Dr. Arun K. Jagota writes copiously for Toward Data Science, and he has one PageRank example: https://medium.com/towards-data-science/pagerank-illustrated-c056a45a2f60. I've co-written an article on PageRank myself, https://rjlipton.wpcomstaging.com/2014/07/21/shifts-in-algorithm-design/, which has some Pythonic homor. But these do not show high in hits.

with `data` for our purposes, and the former seems *not* to bring up pages with "sex trafficking," but I've avoided it anyway. Look for all three terms but don't care as much about the count of *websites*, because one website could generate a lot of traffic.

1. Enter the search `Internet data percent porn extremetech` (without quotes) into Google. You should see the Extreme Tech link and a BBC link above or below it. Note how the BBC story says that the Extreme Tech article is "regularly quoted for calculating that 30% of all net traffic is generated by porn sites." Count at least 3 other hits that reference Extreme Tech—you can tell from the capsules and need not click on them.

2. Now click the Extreme Tech article. *Skim* it, but note the hedging in the last few sentences on page 2 of 2.

3. Now enter the search `Internet data percent porn`, that is, without `extremetech`. See from the capsules how many hits say "30%" and whether they reference PaintBottle/Huffington Post and/or Anthony/Extreme Tech. (I don't know if the former drew from the latter.) Then see if inserting `30` before `percent` in the search changes much. Note the years on the hits, whether any are later than 2017.

Finally, write a paragraph on what you observed about hits referencing each other, either with links or not, and their dates. Give your thoughts on how the nature of Google search may have influenced what you got as hits, and on how reliable you think the "30 percent" figure is now. (3 pts., for 9 total)