Sheetal Parikh
Problem Set – Module 9
605.744.81 Information Retrieval
Fall 2021

1.) *Give a short definition or explanation of the following concepts:*

a.) Broder's taxonomy

Broder's taxonomy was developed by scientist Andrei Broder to demonstrate the different requirements of a search engine depending on a user's search request. User needs are broken down into 3 major types of queries: 1.) Information queries, 2.) Transactional queries and 3.) Navigational queries. Information queries are normal queries in which the user is searching for information about a topic. Transactional queries help the user perform a transaction or obtain a service such as buying movie tickets or making travel reservations. With navigational queries, the user is trying to reach a particular site. The taxonomy helps demonstrate how we can generate better search results by focusing on the needs of the user which can overall help produce a better ranked list.

References:
Lecture 9D
https://medium.com/@seokai/broders-classification-of-keywords-16ddb1015a3

b.) in-degree

The in-degree of a web page represents the number of in-links or incoming hyperlinks to that particular page. In several studies, a web page's in-degree ranges from roughly 8 to 15. The in-degree of a web page is important for calculating the PageRank score for a page because the score is dependent on a page's in-degree and the pages that point to it.

References:
Pages 425: Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press

c.) learning to rank (in the context of web retrieval)

In the context of web retrieval, learning to rank is the application of using machine learning techniques, particularly supervised classifiers, to help in ranking or reranking the top of a ranked list of documents for queries. Supervised machine learning was generally not used for ranked retrieval because the number of queries can be infinite and training data would not be available. However, this topic has become more popular we now have access to plenty of user data due to sites mining user info, explicitly provided user information from account registrations, as well as metadata such as a user's IP address. Implicit info is also available based on a user's behavior such as evaluating how long a user looks at a page before taking an action or seeing what link a user clicks on which can serve as positive reinforcement for that page and negative reinforcement for other pages. Overall, many features can be obtained from all of the sources mentioned above which then can be used in supervised classification for approaches such as regression, pairwise preference, and list ordering that can help improve or re-rank document and give a better experience for the user.

References:
Lecture 9E

d.) robots exclusion protocol

The robots exclusion protocol is a voluntary honor code that tells web masters how to control behavior for compliant web crawlers. A robots.txt file is a protocol file used to give robots and web crawlers instructions about which files or directories should not be downloaded or indexed. Web crawlers can connect to numerous internet hosts and websites have robots that have too many connections or consume too much bandwidth. Therefore, the protocol helps a host to place a certain portion of his or her website off-limits to crawling. If the robots.txt file does not exist, robots and crawlers can assume that all parts of the website can be accessed.

References:
Lecture 10C

e.) web spam

Web spam is similar to email spam and is the manipulation of web page content by having pages that are not meant to be useful to the user, to appear high up in search results for selected keywords from queries. These pages are intended to make the spammers money by monetizing page usage by appearing to be good results by being ranked highly even though they do not provide any value to the user. Search engines are now able to detect large amounts of spam by screening out repetitions of certain keywords. Spammers have developed techniques such as cloaking in which the content presented to the web crawler is manipulated so that the content presented to the user's browser is different than that of the web crawler.

References:
Pages 426 – 429: *Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press*

Lecture 9D

2.) *Describe in your own words the process described in the course text to efficiently identify near duplicate documents in a large collection.*

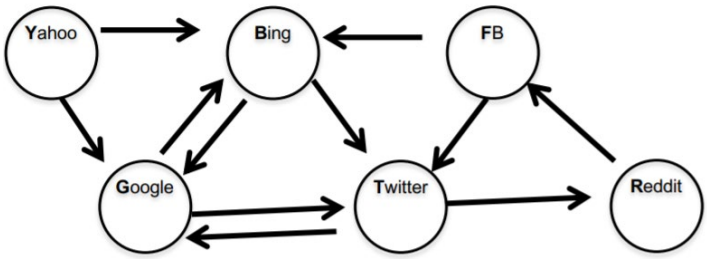
The web contains numerous copies of the same content and in many cases the content is very similar but not identical. These documents that are very similar but differ slightly on content are called near duplicates. Because it is not feasible to exhaustively compare all pairs of web pages, a technique called shingling is used. In the text, shingles are consecutive overlapping of terms, similar to word n-grams. The k-value for the shingles have to be chosen. A typical value used for identifying near-duplicate web pages is k = 4. Two documents are near duplicates if two documents have similar sets of shingles. To determine the similarity, the Jaccard Similarity metric is used which is calculating by taking the size of the intersection of shingles of two documents divided by the size of the union of shingles of two documents.

Because comparing the shingles with each other for each pair of documents is still very costly, a subset of the shingles can be compared instead. This subset of shingles is taken from each document, which is called a sketch, and should be representative of the document. The Jaccard similarity can now be computed using the sketch and based on the similarity threshold that is set, we can determine whether we have a near-duplicate document. Because results can vary depending on the sketch (meaning a document could be considered a near-duplicate in one sketch and possibly not be considered a near-duplicate in another sketch) it is important to get a random permutation of shingles and see how many of them have matching results. Therefore, this process should be repeated for multiple permutations for the most representative results.

References:
Pages 437 - 441: Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press

https://medium.com/@jonathankoren/near-duplicate-detection-b6694e807f7a

3.) *For this problem work with the directed web graph shown below. In the graph there are six nodes: Y, B, F, G, T, R (for the websites Yahoo, Bing, Facebook, Google, Twitter, and Reddit). Use a teleport probability of 0.20. Assume no other pages or links exist beside those shown in the figure.*



(a) Provide (i.e., write) the six recurrence equations that indicate how to iteratively calculate the PageRank score of each page at time t given scores from time t-1.

$$PageRank\ Formula = PR(a) = \frac{q}{N} + (1-q)\sum_{i=1}^{n}\frac{PR(p_i)}{C(p_i)}$$

(1) $PR(Yahoo,\ t_i) = \frac{q}{N} + (1-q) * 0$
$= \frac{q}{N}$

(2) $PR(Google,\ t_i) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{i-1})}{2} + \frac{PR\ (Bing,\ t_{i-1})}{2} + \frac{PR\ (Twitter,\ t_{i-1})}{2}\right]$

(3) $PR(Bing,\ t_i) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{i-1})}{2} + \frac{PR\ (Google,\ t_{i-1})}{2} + \frac{PR\ (FB,\ t_{i-1})}{2}\right]$

(4) $PR(Twitter,\ t_i) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Google,\ t_{i-1})}{2} + \frac{PR\ (Bing,\ t_{i-1})}{2} + \frac{PR\ (FB,\ t_{i-1})}{2}\right]$

(5) $PR(FB,\ t_i) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Reddit,\ t_{i-1})}{1}\right]$

(6) $PR(Reddit,\ t_i) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Twitter,\ t_{i-1})}{2}\right]$

(b) Using the brute-force iterative method of calculation shown in the video lecture calculate two iterations of PageRank scores for each page in the graph. Be sure to show scores at times t=0, t=1, and finally at t=2. Report scores using three digits of precision (e.g., 0.247, not 0.2 or 0.24696485932). Show work and do not merely provide a table of values.

	Yahoo	Google	Bing	Twitter	FB	Reddit
t = 0	0.167	0.167	0.167	0.167	0.167	0.167
t = 1	0.033	0.234	0.234	0.234	0.167	0.100
t = 2	0.033	0.234	0.207	0.287	0.113	0.127

t = 0:

I set PR(x) = 1/6 ≈ 0.167 since we have 6 pages. Therefore, at t = 0, all 6 pages have a PageRank score of 0.167 as can be seen in the table above.

t = 1:

N = 6 pages

q = teleport probability = 0.20

(1) $PR(Yahoo,\ t_1) = \frac{q}{N} + (1-q) * 0$

$= \frac{0.20}{6}$

$= \mathbf{0.033}$

(2) $PR(Google,\ t_1) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{1-1})}{2} + \frac{PR\ (Bing,\ t_{1-1})}{2} + \frac{PR\ (Twitter,\ t_{1-1})}{2}\right]$

$= \frac{0.20}{6} + (0.80)\left[\frac{0.167}{2} + \frac{0.167}{2} + \frac{0.167}{2}\right]$

$= 0.0333 + 0.2004$

$= \mathbf{0.234}$

(3) $PR(Bing,\ t_1) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{1-1})}{2} + \frac{PR\ (Google,\ t_{1-1})}{2} + \frac{PR\ (FB,\ t_{1-1})}{2}\right]$

$= \frac{0.20}{6} + (0.80)\left[\frac{0.167}{2} + \frac{0.167}{2} + \frac{0.167}{2}\right]$

$= 0.0333 + 0.2004$

$= \mathbf{0.234}$

(4) $PR(Twitter,\ t_1) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Google,\ t_{1-1})}{2} + \frac{PR\ (Bing,\ t_{1-1})}{2} + \frac{PR\ (FB,\ t_{1-1})}{2}\right]$

$= \frac{0.20}{6} + (0.80)\left[\frac{0.167}{2} + \frac{0.167}{2} + \frac{0.167}{2}\right]$

$= 0.0333 + 0.2004$

$= \mathbf{0.234}$

(5) $PR(FB,\ t_1) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Reddit,\ t_{1-1})}{1}\right]$

$= \frac{0.20}{6} + (0.80)\left[\frac{0.167}{1}\right]$

$= 0.0333 + 0.1336$

$= \mathbf{0.167}$

(6) $PR(Reddit,\ t_1) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Twitter,\ t_{1-1})}{2}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.167}{2}\right]$$
$$= 0.0333 + 0.0668$$
$$= \mathbf{0.100}$$

t = 2:

(1) $PR(Yahoo,\ t_2) = \frac{q}{N} + (1-q) * 0$

$$= \frac{0.20}{6}$$

$$= \mathbf{0.033}$$

(2) $PR(Google,\ t_2) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{2-1})}{2} + \frac{PR\ (Bing,\ t_{2-1})}{2} + \frac{PR\ (Twitter,\ t_{2-1})}{2}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.033}{2} + \frac{0.234}{2} + \frac{0.234}{2}\right]$$
$$= 0.0333 + 0.2004$$
$$= \mathbf{0.234}$$

(3) $PR(Bing,\ t_2) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Yahoo,\ t_{2-1})}{2} + \frac{PR\ (Google,\ t_{2-1})}{2} + \frac{PR\ (FB,\ t_{2-1})}{2}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.033}{2} + \frac{0.234}{2} + \frac{0.167}{2}\right]$$
$$= 0.0333 + 0.1736$$
$$= \mathbf{0.207}$$

(4) $PR(Twitter,\ t_2) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Google,\ t_{2-1})}{2} + \frac{PR\ (Bing,\ t_{2-1})}{2} + \frac{PR\ (FB,\ t_{2-1})}{2}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.234}{2} + \frac{0.234}{2} + \frac{0.167}{2}\right]$$
$$= 0.0333 + 0.254$$
$$= \mathbf{0.287}$$

(5) $PR(FB,\ t_2) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Reddit,\ t_{2-1})}{1}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.100}{1}\right]$$
$$= 0.0333 + 0.0800$$
$$= \mathbf{0.113}$$

(6) $PR(Reddit,\ t_2) = \frac{q}{N} + (1-q)\left[\frac{PR\ (Twitter,\ t_{2-1})}{2}\right]$

$$= \frac{0.20}{6} + (0.80)\left[\frac{0.234}{2}\right]$$
$$= 0.0333 + 0.0936$$
$$= \mathbf{0.127}$$

(c) Which page (or pages) has/have the lowest PageRank score after two iterations?

After two iterations, Yahoo has the lowest PageRank score at 0.033.

(d) Which page (or pages) has/have the highest PageRank score after two iterations?

After two iterations, Twitter has the highest PageRank score at 0.287.

References:
Lecture 9C