Assignment 1 **Hussam Hallak**
CS834, Information Retrieval, Fall 2017 CS Master's Student
Old Dominion University, Computer Science Dept Prof: Dr. Nelson

Question 1:

Exercise 1.1:

Think up and write down a small number of queries for a web search engine. Make sure that the queries vary in length (i.e., they are not all one word). Try to specify exactly what information you are looking for in some of the queries. Run these queries on two commercial web search engines and compare the top 10 results for each query by doing relevance judgments. Write a report that answers at least the following questions: What is the precision of the results? What is the overlap between the results for the two search engines? Is one search engine clearly better than the other? If so, by how much? How do short queries perform compared to long queries?

Answer:

To answer this question, the following queries were tested on both Google and Bing.

1. Stent
2. Ureteral stent
3. Ureteral stent procedure
4. Do ureteral stents cause pain?

I chose these queries because, two weeks ago, I was told by the doctor that I need a "stent" to help pass a stone in my kidney. I had no idea what a stent is, so I asked what is a stent? and she explained. Let's pretend that she did not. I want to know what a stent is, the procedure of stenting, the removal of a stent, and whether or not stents cause pain.

1. Stent: I ran the first query "Stent" and found the following:

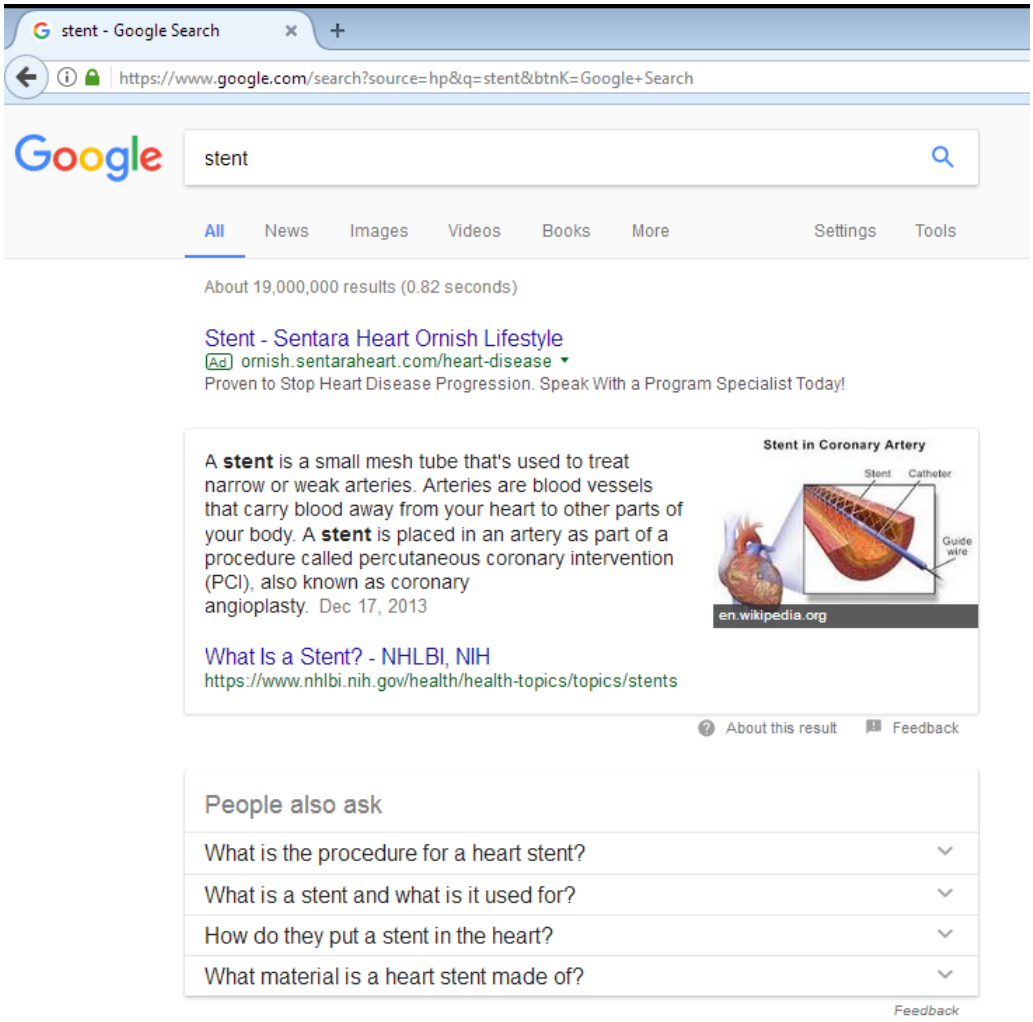
A. Google: Google found about 19,000,000 results. Skipping the ads and looking at the summary given by Google about a stent, I did not see that a stent is related to kidney stones. The summary stated that a stent is a mesh tube that is used to treat narrow or weak arteries. I opened all of the first 10 results returned by Google and found that only the 6th result mentioned something about ureteral stents. Obviously, there are different types of stents for different purposes. I am specifically looking for something related to kidney stones, but I was not specific enough for Google.

Precision is the fraction of retrieved documents that are relevant to the query. In other words, it is the number of correct results divided by the number of all returned results.

$$precision = \frac{|\{relevant\ documents\} \cap \{returned\ documents\}|}{|\{returned\ documents\}|}$$

$$precision = \frac{1}{10} = 0.1$$

Figure 1: Query: Stent, Search Engine: Google



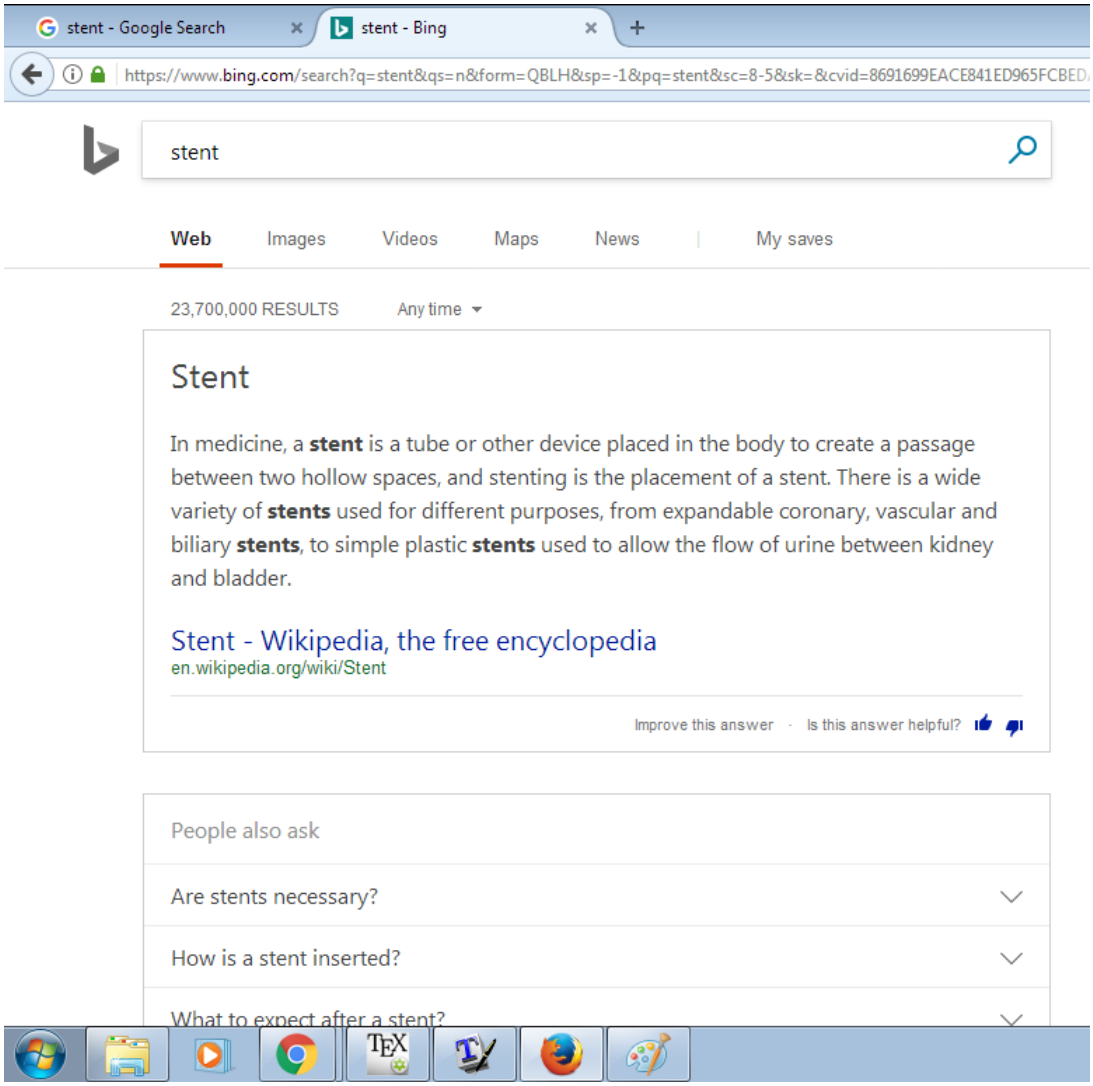
B. Bing: The query returned 23,700,000 results. The summary returned by Bing gave a general definition of a stent. The summary contains exactly what I am looking for "stents used to allow the flow of urine between kidney and bladder". This summary, returned by Bing, came from and the document in the 6th result, from wikipedia.org: https://en.wikipedia.org/wiki/Stent

Out of the 10 results returned by Bing, Only 3 talked about ureteral stents.

$$precision = \frac{|\{relevant\ documents\} \cap \{returned\ documents\}|}{|\{returned\ documents\}|}$$

$$precision = \frac{3}{10} = 0.3$$

Figure 2: Query: Stent, Search Engine: Bing



Overlap: The following table shows the ordered results returned by each search engine:

Order	Google	Bing
1	http://www.webmd.com/heart-disease/guide/stents-types-and-uses#1	https://medlineplus.gov/ency/article/002303.htm
2	https://www.heart.org/idc/groups/heart-public/@wcm/@hcm/documents/downloadable/ucm_300452.pdf	https://en.wikipedia.org/wiki/Coronary_stent

3	https://www.healthline.com/health/stent	http://secondscount.org/treatments/treatments-detail?cid=7709f984-f6a5-44bb-8c2f-d7114c5b4c0b
4	https://www.nhlbi.nih.gov/health/health-topics/topics/stents/	https://www.nhlbi.nih.gov/health/health-topics/topics/stents/
5	https://www.nhlbi.nih.gov/health/health-topics/topics/stents/after	http://www.webmd.com/heart-disease/guide/stents-types-and-uses
6	https://en.wikipedia.org/wiki/Stent	https://www.healthline.com/health/stent
7	https://www.nhlbi.nih.gov/health/health-topics/topics/stents/risks	http://www.mayoclinic.org/tests-procedures/coronary-angioplasty/home/ovc-20241582
8	https://myheart.net/articles/stent-save-life/	https://en.wikipedia.org/wiki/Stent
9	http://www.livemint.com/Industry/HMU54RjTrKBHQKPj4QSv0O/Boston-Scientific-may-withdraw-its-highend-stent-Synergy.html	http://medical-dictionary.thefreedictionary.com/stent

10	https://medlineplus.gov/ency/article/002303.htm	https://www.heart.org/idc/groups/heart-public/@wcm/@hcm/documents/downloadable/ucm_300452.pdf

From the table, it is clear that the overlap between the results for the two search engines is six results out of 10 or 60%.

$$G_1 \equiv B_5$$
$$G_2 \equiv B_{10}$$
$$G_3 \equiv B_6$$
$$G_4 \equiv B_4$$
$$G_6 \equiv B_8$$
$$G_{10} \equiv B_1$$

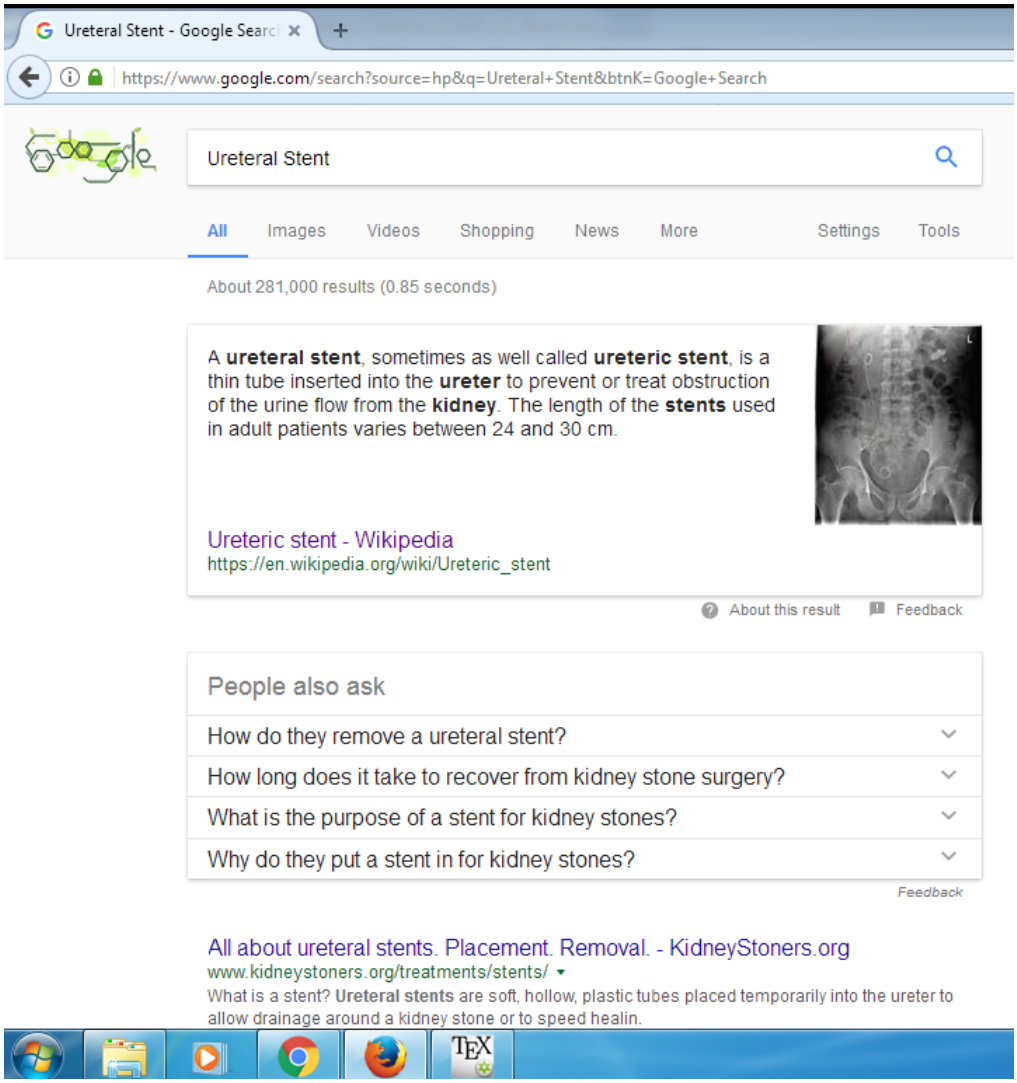
Where:
G_i denotes the ith result returned by Google.
B_j denotes the jth result returned by Bing.

2. Ureteral Stent: I ran the query "Ureteral Stent" in the same manners as the first query and found the following results:

A. Google: About 281,000 results are returned by Google. All of the first 10 results returned by Google, obviously, were related to ureteral stents. Only one of the links did not allow me to read the entire article until I sign up and login to the website, which I did not. I will consider this result not to be a good one because Google returned a result for which the representation of the resource was not retrievable without being a member of the website. All of the remaining 9 results contained information about the procedure, side effects, etc.

$$precision = \frac{9}{10} = 0.9$$

Figure 3: Query: Ureteral Stent, Search Engine: Google

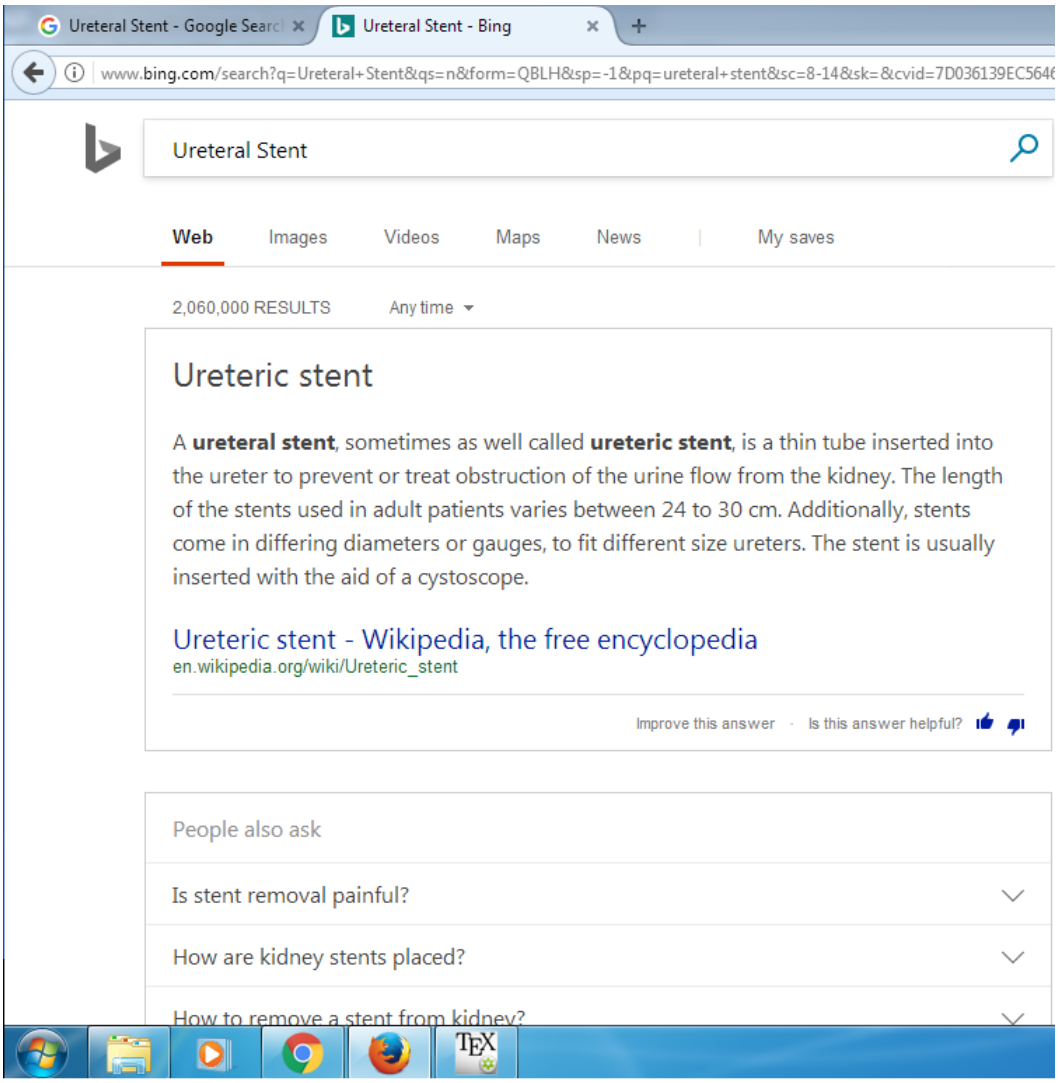


B. Bing:

The query returned 2,060,000 result. 8 out of the first 10 retrieved documents contained almost everything I needed to know about ureteral stents. One document listed the types/shapes of ureteral stents, but no other information. One document briefly explained what a ureteral stent is, but did not provide any information about the procedure, complications, side effects, etc.

$$precision = \frac{8}{10} = 0.8$$

Figure 4: Query: Ureteral Stent, Search Engine: Bing



Overlap: The overlap between the results for Google and Bing is 5 out of 10 or 50%.

3. Ureteral Stent Procedure: Similarly, I ran the query "Ureteral Stent Procedure" and found the following results:

A. Google: About 153,000 results are returned by Google. One of the links did not allow me to read the entire article until I sign up and login to the website; the same link was returned running the previous query. One link in the results did not have enough

information about the procedure itself, but it had good information about the recovery, diet, etc. All of the remaining 8 results contained information about the procedure, side effects, etc.

$$precision = \frac{8}{10} = 0.8$$

B. Bing: The query returned 3,330,000 results. 8 out of the first 10 retrieved documents contained enough information about ureteral stent procedure. One document was about stenting pets dogs and cats. One document briefly explained what a ureteral stent is, but did not include the procedure.

$$precision = \frac{8}{10} = 0.8$$

Overlap: The overlap between the results for Google and Bing is 5 out of 10 or 50%.

4. Do ureteral stents cause pain?: Finally, I ran the query "Do ureteral stents cause pain?" and found the following results:

A. Google: About 908,000 results are returned by Google. All of the 10 results had useful information about the pain and ways to manage it.

$$precision = \frac{10}{10} = 1.0$$

B. Bing: The query returned 163,000,000 results. 8 out of the first 10 retrieved documents contained enough information about the pain associated with ureteral stent procedure and how to manage it. One document had questions from patients answered by doctors, but there was not any questions about the pain. One document briefly explained what a ureteral stent is, but did not include information about the pain.

$$precision = \frac{8}{10} = 0.8$$

Overlap: The overlap between the results for Google and Bing is 6 out of 10 or 60%.

Conclusion:

Bing did a better job on the single word query I ran, while Google got better as the query got longer and more specific.

Average Precision for Google is:

$$precision_{AVG} = \frac{0.1 + 0.9 + 0.8 + 1.0}{4} = 0.7$$

Average Precision for Bing is:

$$precision_{AVG} = \frac{0.3 + 0.8 + 0.8 + 0.8}{4} = 0.675$$

The average precision for Google is slightly higher than it is for Bing.

The average overlap between the results for Google and Bing is 55%:

$$Overlap_{AVG} = \frac{0.6 + 0.5 + 0.5 + 0.6}{4} = 0.55$$

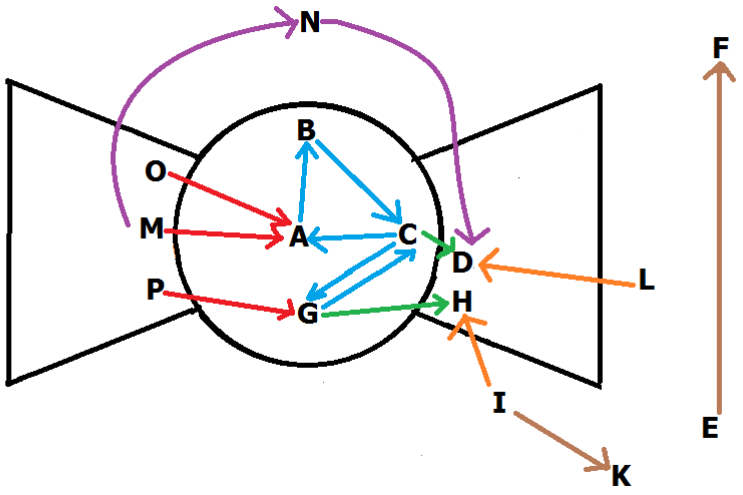
Question 2:

Exercise 3.8: Suppose that, in an effort to crawl web pages faster, you set up two crawling machines with different starting seed URLs. Is this an effective strategy for distributed crawling? Why or why not?

Answer:

Assuming that the number of machines is fixed. Yes! This is an effective strategy for distributed crawling because otherwise, if we set up both crawling machines with the same starting seed URLs, the chance that both crawlers end up crawling the same pages is large, not only because the starting point is identical, but also because most links on a page lead to pages that are related to it. The likelihood of both crawlers crawling the same pages will decrease if we set them up with different starting seed URLs. The original Google paper stated that Google has a fast distributed crawling system where a single URLserver serves lists of URLs to a number of crawlers (They ran 3). I assume that the URLserver is not serving the same URLs to all 3 crawlers, because if that was the case, there will be no point in adding 2 or more crawlers. I included the bowtie graph below to demonstrate my argument. Let's assume that the starting node for both crawlers is P. It is clear that the node N will not be crawled because there is no path from P to N. On the other hand, if we set two different starting points M and P for the first and second crawler respectively, the first crawler will crawl the node N as well as all other nodes crawled by the second crawler.

Figure 5: Bowtie



References

[1] Python For Beginners. Available from World Wide Web:(http://www.pythonforbeginners.com/).

[2] Cambridge University Press. Available from World Wide Web: (http://nlp.stanford.edu/IR-book/html/htmledition/the-web-graph-1.html).