Lab 1.4
Smarter Searching on the Web

We will begin by reviewing search engines and discussing common strategies for effective web searching. In the remainder of the lab, you will practice applying these strategies to answer a series of questions about the history of computers, and some other things. Although we will focus on web search throughout the lab, this will be an excellent opportunity to learn a little about the pioneering people and events in this field. It seems like computers have been around forever, so you might be surprised at how recently the world took its first steps in computing.

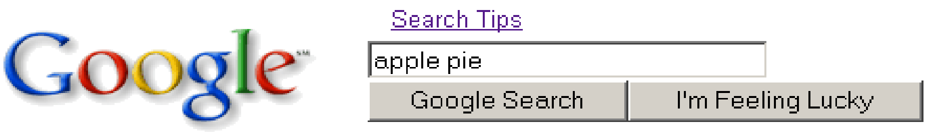
Discussion and Procedure

### Part 1.	Getting Answers with Better Queries

In Chapter 5, we learned the basics of how a *web search engine* works. This knowledge will help you use search engines like Google, the engine we will use in this lab, effectively, quickly finding the answers to your questions. Using a search engine is quite simple: You provide a *query* in the form of one or more *keywords*, and the search engine returns a set of web pages which include the words you specify. The resulting web pages are sometimes called *hits*. Using a search engine effectively by forming good queries, however, requires some careful thought.

What makes a query good? It all depends on the question you want to answer. The typical search engine user's primary goal is usually not pages that contain some set of words. Rather, they are looking for the answer to a more general question not necessarily related to web pages. For example, they might be asking, "What are some good recipes for apple pie?" A good query is one whose resulting web pages are useful for answering the original question. A query that is not as good is one whose keywords match pages that are not relevant to the original question.

In the pie recipe case, a simple and obvious query to try would be the following:

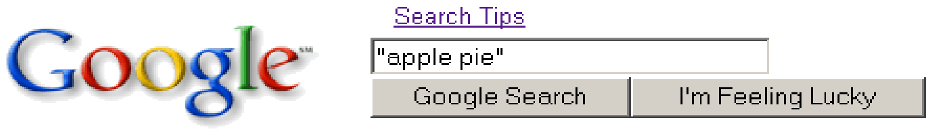


apple pie

However, when we tried this query, it returned about 370,000 pages, including many pages that have nothing to do with the food apple pie or how to make it but happen to have the words "apple" and "pie" on them. For example, only two of the ten first hits actually contained recipes, and the rest were irrelevant. Apple Pie USA (second hit) is a nanny placement agency, New York University has a computer science research project called the Apple Pie Parser (third hit), Warm Apple Pie (fourth hit) is a fan site for the American Pie movies, and there is a company in New York called Apple Pie Web Design (ninth hit).

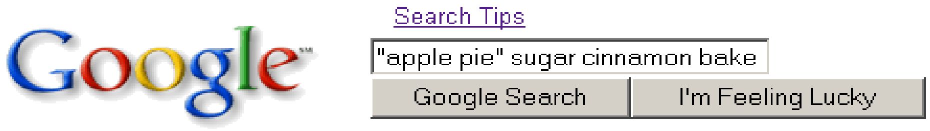
A good query would return pages containing apple pie recipes. In the remainder of this part of the lab, we will discuss several methods for *query refinement*, modifying your query to return a smaller, more relevant set of results. Although we will be providing instructions specific to Google, all major web search engines offer equivalent query features, so you should check their respective help pages to find out how to apply these refinement methods.

Use phrases instead of independent keywords. The above query results in pages that contain the words "apple" and "pie," including pages where the words appear separately. For example, Apple Computer's web site hosts many movie trailers in their QuickTime movie file format, including the trailer for *American Pie 2*, and this trailer's page is included in the 370,000 hits. You can narrow your search to only pages containing the words "apple pie" together by putting the phrase in quotes:


"apple pie"

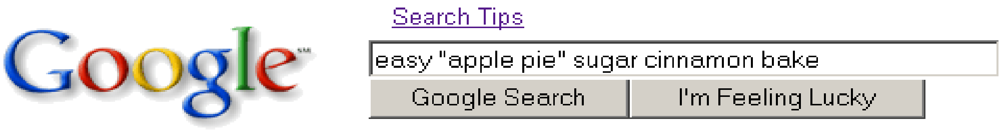
This results in 180,000 hits, which is still a very large number, but much fewer than 370,000.

Add keywords you know will be on relevant pages. The phrase query above is better than the original query, but it can be improved further. For instance, the results still include pages for companies selling apple pies, rather than offering recipes for them. One way to refine your query further is by adding keywords that you are certain will be on any page that is relevant to your original question about apple pie recipes. Sugar and cinnamon are almost always ingredients in an apple pie, and you always bake a pie, so this query is likely to return a smaller set of pages without excluding recipes interesting to you:


"apple pie" sugar cinnamon bake

This query results in about 8,620 pages, most of which actually appear to be apple pie recipes.

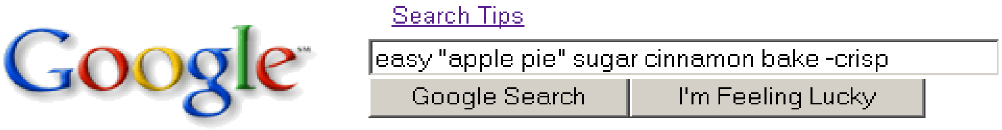
Adding keywords for a more specific answer. You might not be interested in sifting through over eight thousand pages of recipes, so, at this point, you might carefully reconsider your original question and see if you can make it more specific. For instance, you might not be such an expert cook, so to find recipes that are at your level, you might add the keyword, “easy”:



easy "apple pie" sugar cinnamon bake

This query results in 2,750 results, the vast majority of which are apple pie recipes that claim to be easy to prepare. By this point, most users would just try a few of the pages and call the web search done, but we will extend the example further in order to illustrate a few more useful query refinement strategies.

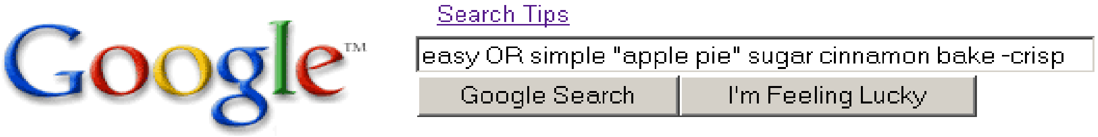
Exclude keywords that you know *do not appear* on relevant pages. Many recipes for apple crisp list pre-made apple pie filling as an ingredient, so the above query returns pages with such recipes. To exclude these pages from your results in Google, you can add the keyword “crisp” with a minus sign before it:



easy "apple pie" sugar cinnamon bake -crisp

This further focuses your result set to 1,750 pages.

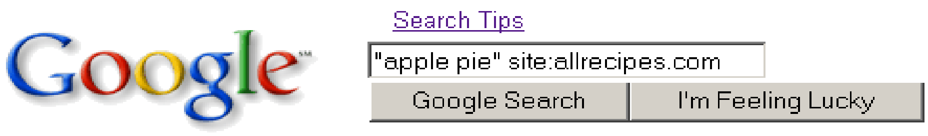
Add keyword alternatives for word variations. “Easy” is only one word that describes recipes suitable for beginning cooks, so you might also want to allow pages with the word “simple” on them to be included in your results. Simply adding this keyword to your query, however, would require that both “easy” *and* “simple” be on the page. Using Google, you can indicate that one or the other (or both) are acceptable by putting “OR” (note capitalization) between the keywords:



easy OR simple "apple pie" sugar cinnamon bake -crisp

Note that in this case, you are likely to increase the number of hits, but this might be useful if you believe you would otherwise exclude relevant pages.

Restrict your search to a particular domain. While going through some of the results from the above queries, you might notice that you are particularly interested in pages from the `allrecipes.com` web sites. You can restrict your Google search to pages on web servers in a particular domain by using “site:*domain*” to your query:



"apple pie" site:allrecipes.com

The above query finds all pages on servers within the `allrecipes.com` domain. (Note that there are no spaces around the colon between “site” and the domain name.)

Part 2. Query Refinement Exercises

Use web searches to find the answers to the questions below. With each answer, write the following:

- the series of queries you used to find the answer, including those which ended up not being so helpful
- an explanation of your search strategy, identifying which of the above query refinement strategies you applied and why
- the URL of the page where you found the answer to the question

In many cases, the page you find to answer one question will answer others. If you find an answer to a question on a page found for a previous question, you don't need to repeat the whole search description; just indicate which previous URL provided the answer. The first one is done for you.

<u>Hint: You can copy and paste the questions into a word document and type in your answers. This will also allow you copy and paste the URLs where you find your answers.</u>

1. IBM produced the first hard disk drive. Find a page on an official IBM web site describing this disk. When was it made, what was it called, and how much storage space did it offer?

   (1) query “IBM” to figure out IBM’s domain name, which is `ibm.com` (Or maybe you just guessed.)
   (2) query “site:ibm.com "hard disk"” for pages on IBM’s web site(s) with phrase “hard disk” in them
   (3) eventually worked to “site:ibm.com "hard disk" first early history” to find out RAMDAC was the name of the system (http://www-1.ibm.com/ibm/history/history/decade_1950.html)
   (4) tried “site:ibm.com "hard disk" RAMDAC” and found the size (5M) at http://domino.research.ibm.com/Comm/bios.nsf/pages/gmr.html

2. How many **gallons** of water are in Lake Mead? (Please read the page carefully; there's a common wrong answer that comes from students not reading what they're seeing.)
   Answer:

   URL(s):

Search Term(s) tried:

3. The type of LAN used in the computer lab is an Ethernet. Who invented Ethernet?
   Answer:

   URL(s):

   Search Term(s) tried:

4. What radio frequencies are used by the WiFi wireless networks? (If you are unsure what a radio frequency looks like, you might search about that first.)
   Answer:

   URL(s):

   Search Term(s) tried:

5. The primary Internet protocols are called TCP/IP. Who are the two co-inventors of TCP/IP?
   Answer:

   URL(s):

   Search Term(s) tried:

6. What was the first computer spreadsheet program called? Who created it, and what sort of computer did it run on?
   Answer:

   URL(s):

   Search Term(s) tried:

7. What wast the first piece of software created by Microsoft? What sort of computer did it run on? (It was not any version of DOS, Windows, or Word.)
   Answer:

   URL(s):

   Search Term(s) tried:

8. PNG is an image format used on the web. What does PNG stand for, and what format was it designed to replace?
   Answer:

   URL(s):

   Search Term(s) tried:

9. What country received the largest share of United States petroleum exports in 2019?
   Answer:

   URL(s):

   Search Term(s) tried:

10. Arlo Guthrie sang the famous popular song, “The City of New Orleans.” Who wrote it?
    Answer:

    URL(s):

    Search Term(s) tried:

11. In the lyrics of that song, what is the northernmost town or city mentioned? What state is it in?
    Answer:

    URL(s):

    Search Term(s) tried: