ECE101: Exploring Digital Information Technologies **Fall 2023**
Week 7 Exercises (with solutions in red) **based on L12,L13; tested on Exam II**

Search and Recommendation Engines

These exercises are intended to help you master and remember the material discussed in lectures and explored in labs. In future semesters, we may make some or all of these exercises required, but for now they remain optional. We suggest that you do them as we go over the material, but you may also want to use them to review concepts before the exam.

Rather than using this version directly, we suggest that you use the version without solutions to solve the problems before looking at the answers. Many studies have shown that people often trick themselves into believing that they know how to solve a problem if they are presented with the answer before they try to solve the problem themselves.

1. [L12] Think back to the first part of our course and our discussion of the Domain Name Service (DNS). When your computer needs to find the IP address for whitehouse.gov, it makes a series of requests to the DNS hierarchy, eventually obtaining the desired IP address.

   Many research publications require the authors to select topics and/or keywords from a list published by a related professional organization. For example, an article might discuss “page ranking” or, more generally, “Internet search.” One can then easily turn the relationship around: by filing each publication in a folder corresponding to each of the publication’s keywords, one can allow someone who wants to learn about a topic to easily find all relevant publications.

   Now let’s try to combine those ideas to replace search engines. Rather than crawling the web, each published web page (authors are not required to participate) could register keywords with a service like DNS—let’s call it the Keyword Lookup Service, or KLS. KLS can organize the URLs by adding a given URL to a folder for each of the selected keywords. Then when someone wants to find relevant documents on the web, they can walk the KLS hierarchy to obtain the relevant list of URLs!

   Explain why such an approach is unlikely to work well in practice compared with the approach taken by Internet search engines.

   One drawback is that the approach assumes that web page publishers play nicely. Unless KLS charges on a per-page basis and limits the number of keywords chosen per page, KLS is likely to be inundated with spam pages. For example, thousands of URLs pointing to a single advertisement, each marked with all possible keywords to bring in readers.

   A second drawback is the loss of information about interlinked pages, which form the basis for page ranking. KLS is thus reduced to providing an unordered list of URLs. Again, without some kind of financial incentive, KLS is also unlikely to be able to perform even basic per-user ordering (perhaps on IP address alone, but even that effort requires money for servers to do the work).

2. [L12] Explain the importance of indexing documents when operating a web search engine.

   Web search engines are expected to return results quickly (in less than half a second from the user’s perspective!), so after locating and downloading billions of web pages from the Internet, it is important to organize them in a way that allows fast access when a user provides a keyword or phrase. That organization process is called indexing.

3. [L12] Several companies now offer integrated software platforms that allow your personal information to migrate freely from your desktop to your laptop to your phone and possibly even to a machine that you use at school or at a kiosk in an airport. Providing such conveniences requires that a company collect and retain that personal information on their servers, thus also enabling the company to tailor advertisements and web searches for you. For each of the following sources of information, give an example of how a company's use of a specific type of information collected from that context might benefit you.

   A) web search
   If I have been searching for a particular topic, such as how search engines work, and then later am shopping, I might be quite interested in seeing advertisements for a new book on search engine design and operation.

   B) online purchasing using a stored credit card
   If I have just purchased a plane ticket to Costa Rica, I might interested in having my social network "remind" me that, five years ago, a close friend of mine also visited Costa Rica, by putting one or two of their trip photos in my feed.

   C) email
   If I am exchanging emails with someone about back pain after a recent car accident, I may be interested in having advertisements for injury lawyers show up next time I browse the web, even if I am not actively searching for one.

   D) messaging
   If I have been chatting with my friends about visiting nearby ziplines, I might be interested in news articles about recent zipline failures (I've never heard of such things—just an example!).

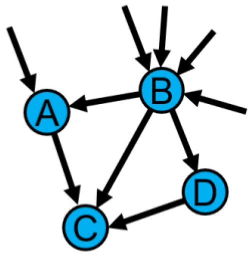
   E) social media
   If I have posted photos of myself with a friend at a restaurant, I may be interested in having similar restaurants (recommendations based on my preferences) pop up when I search for "food near me."

   F) phone calls
   If I call a government agency and am disconnected because they're too busy to talk to me (seriously, it happened four times in one day to me once!), I might like to have my desktop pop up a notice telling me that the government agency's phone line is currently less busy than usual.

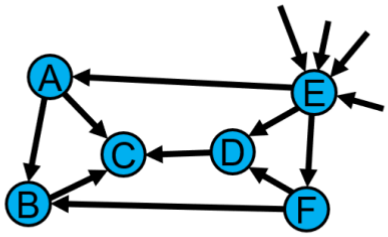
   G) videoconferencing
   If I chat with someone during a seminar, I may appreciate having my social network suggest that I add them as a friend so that we can talk further.

4. [L12] Each of the following graphs comes from a web search—all incoming arcs are shown. For each graph, rank the pages (nodes) in decreasing order of reputation according to page rank. (We don't expect you to do the full computation—just compare the number of incoming arcs. To break ties, use a second level of incoming arcs (how many nodes point to nodes that point to a page).

Graph #1



Graph #2



Graph #1: B (4 incoming arcs), C (3 incoming arcs), A (2 incoming arcs), D (1 incoming arc)
Graph #2: E (4 incoming arcs), C (3 incoming arcs), D (2 incoming arcs from E and F, which in turn have a total of 5 incoming arcs), B (2 incoming arcs from A and F, which in turn have a total of 1 incoming arc), F and A are tied (both have 1 incoming arc from E)

5. [L13] Recommendation systems work best for things that any given person chooses frequently. For example, one person watches many movies, eats food every day, and engages regularly in their hobbies. In contrast, one person may only buy a washing machine every ten or twenty years. Collaborative filtering is thus practically useless for such types of purchases.

   Content-based filtering, however, may play a role. To use the same example, washing machines have a rich feature space, including size, noise, cost, power, style, color, capabilities, and so forth. But a given user typically has no purchase history in the last decade. How can the feature space of washing machines be used to provide guidance on purchasing in a more interactive way? *(Hint: consider first applying clustering from our machine learning discussion in Week 8 to an assortment of products within the feature space.)*

   One approach: cluster all available washing machines into a small number of groups, then show the person a representative sample machine from each group. Once they have chosen one of the samples, repeat the process by forming subclusters from machines in the cluster—these would all be precalculated, of course—and show the person a representative sample machine from each subcluster. Once the number of possible options is small enough, show the person all options similar to the ones in which they've indicated an interest. (To some extent, human salespeople operate in this manner when customers are clueless, as is often the case when buying something as infrequent as a washing machine. Of course, price and commission are also important factors in such cases.)

   A second approach, without using the Week 8 material: expose a reduced version of the feature space to the person in order to allow them to navigate through the many options by selecting subspaces. This approach is used by many large online stores. One can choose, for example size, color, manufacturer, price range, and so forth, as well as certain features. This approach gently introduces the shopper to the wealth of available choices and allows them to indicate their preferences without inundating them with information (please choose your favorite from among these 20,000 washing machines).

6. [L13] For each of the following categories, suggest three possible dimensions for a feature space in which objects from the category could be placed in order to support recommendations.

   A) university Bachelors degree programs
      location, faculty-to-student ratio, cost, variety of programs at the same institution, diversity of student body
   B) picture frames
      size, material, color/finish, glass/plastic/no cover, cost
   C) carbonated beverages
      flavor, volume, cost, natural/artificial, caffeination level