Problem set 4

Assigned: April 4
Due: April 18

Please submit the answers to this assignment in a format that supports clickable links (e.g. PDF, HTML, Word).

Problem 1

Consider the following tables in the MRS textbook: Table 5.1 (p. 80); table 5.3 (p. 88); figure 9.8 (p. 176); figure 12.3 (p. 221); table 14.3 (p. 276); table 16.1 (p. 323); and table 16.3(a) (p. 341).

Of course, the text here is PDF rather than HTML, but assume, probably realistically, that it is feasible to extract the table structure.

A. Relate these, as far as possible, to the taxonomy of tables proposed in the paper, Web-scale table census and classification, by E. Crestan and P. Pantel. Note any structural aspects that violate the assumptions in Crestan and Pantel, and any unusual features of the information content.

B. Discuss the difficulties of retrieving these tables for a system with the functionality of the Google WebTables system. In particular, discuss

- B.1 What kinds of queries would be these tables be relevant to? What would be involved in doing that match?
- B.2 What kinds of false positives are a danger here? How, if at all, could those be avoided?

Problem 2

First, two observations:

- You can use Hearst patterns to extract sub- or super-categories (hyponyms or hypernyms) of a particular concept *C* by searching on patterns that have *C*. For instance if you search for the quoted string "animals such as ...", you will collect subcategories of "animal".
- Taxonomies are well-defined and easy to create, either manually or automatically, when the entities are concrete (e.g. physical objects) and the categories are clear-cut. The animal kingdom is a standard example of that. In domains where the entities are abstract and categories are nebulous, constructing taxonomies, either manually or automatically, is much harder.

So I want you to try the following experiment. The object of the experiment is to build a 3-level binary taxonomic hierarchy of subcategories for a given abstract term such as ``event'', ``idea'', ``goal'', ``feeling'', etc. Use the following pseudo-code:

```
R = filtering rule on concept phrases. See below.
T = starting concept;
q = [T]
for (i=1 to 3) {
   for (x in q) {
      newq = [];
      do a Google search on the quoted string "x such as";
      [y,z] = the first two subcategories of x in the results, as indicated
```

```
                by the Hearst pattern, that satisfy rule R;
      mark y and z as subcategories of x;
      add y,z to newq;
     }
   q = newq;
  }
```

(If you really want to, you can implement this, but it will be a lot less work to do it by hand.)

For the filtering rule R, you may choose anything that seems plausible. For example, you should certainly exclude proper nouns. You may require that the terms y and z are single-word nouns or you may allow them to be short noun phrases as in Probase. You may require that y and z are in the plural or you may allow them to be in the singular. You might want to exclude gerunds (verb + "ing") like "playing" or "maintaining" or maybe not. The two constraints are, first, that the rule R must be something that could plausibly be automated does not rely on a human understanding, and, second, that it does not use some pre-existing taxonomy for the domain.

Also, you might want to restrict the Google search in some other way; e.g. to Wikipedia or to the Washington Post. That's also OK.

A. Choose a suitable abstract term to start with and a suitable filtering rule. Carry out the experiment. State the starting term and the rule. Display the taxonomy. List the web pages you used in clickable format.

Dead ends and failures: A term X is a dead end if the search "X such as" gives no new results that satisfy R. A failure occurs when all the leaf nodes are dead end, before getting to depth 3; that is, you don't get to any nodes of depth 3. If the tree has some dead ends, that's not a problem If you run into a dead end somewhere --- that is, there are no results for the search "X such as" that satisfy your rule R --- that's OK. However, if your search results in a failure, then report the search as above; try changing something and do a second search. If you get a failure again, then just report that as well, and you're done.

You may end up with something that is not a tree; it may have a cycle or it may be a DAG. If so, indicate that. However, exclude any self-loops (arcs from a node to itself).

B. Discuss any errors or peculiarities in the final taxonomy and explain how they arise. (If there aren't any, then you don't have to answer this part.)

2-level sample output for part A

Note: This is to illustrate the format. You would not be allowed to use "dog" as the starting point, and you are required to take your tree to depth 3.

Starting point: "animals"

Filtering rule: Plural noun phrases of the form "Noun", "Noun noun" or "Adjective Noun". No proper nouns, no gerunds.

```
dogs---> pit bulls  ---> parolees
      |               |
      |               |-> large heads
      |
      |-> guide dogs ---> labradors
                        |
                        |-> puppy walkers
```

Links

dogs ---> pit bulls
dogs --> guide dogs"
pit bulls --> parolees
pit bulls --> large heads
guide dogs --> labradors
guide dogs --> puppy walkers