CPS 49S Google: The Computer Science Within and its Impact on Society - Spring 2007
Homework 1

- Due date: Tuesday, Jan 30, 2007, in class (2.50 PM). Late submissions will not be accepted (unless there are documented excuses from the dean).
- Submission: In class, or email solution in pdf or plain text to shivnath@cs.duke.edu. Note: emailed submissions received after 2.50 PM on the due date will not be accepted (unless there are documented excuses from the dean).
- Indicate your name on your submission.
- Email questions to shivnath@cs.duke.edu and to asic@cs.duke.edu
- Total points = 100.

Question 1 **Points 10**

Give three ways in which anchor text can be used to improve web search. Give a concrete example for each case.

Question 2 **Points 10**

A *hit* is the occurrence of a word on a page. Give three pieces of information that Google keeps track of for each hit. In each case, give an example to show how this piece of information helps to improve the quality of search results.

Question 3 **Points 5**

The web pages that Google crawls are maintained in a repository. We discussed in class that Google allows users to access this repository. How?

Question 4 **Points 10**

We discussed the terms "structured data" and "unstructured data" in class.

1. What type of data does Google have to deal with?
2. Give an example of structured data and give an example query over this data.
3. Give an example of unstructured data and give an example query over this data.

Question 5 **Points 10**

What does the “hypersearching the web” article say about Yahoo’s approach to search? How does Google’s approach differ from Yahoo’s approach?

Question 6 **Points 10**

In general, will a hub or an authority have the higher PageRank? Justify your answer in 3-5 sentences.

Question 7 **Points 15**

Each subquestion below shows a search query and a web page. For the search query shown, categorize the web page into one of (i) hub for the query, (ii) authority for the query, (iii) both hub and authority for the query, and (iv) neither hub nor authority for the query. Also, argue in 2-4 sentences why you made that choice. (To answer this question, you will have to look at these web pages.)

1. http://en.wikipedia.org/wiki/Data_compression for the search query “compression”
2. http://mail.google.com for the search query “email”.
3. http://dir.yahoo.com/Recreation/Automotive for the search query “automotive”.
4. http://www.cs.duke.edu/courses/spring07/cps049s/ for the search query “google freshman seminar duke”.
5. http://www.goduke.com for the search query “duke football”.

Question 8 **Points 15**

Give three important differences between Google and Clever.

Question 9 **Points 5**

In Chapter 1 of the textbook, what argument does the author give in support of his statement that “Google is a media company”?

Question 10 **Points 10**

Give two examples of how the “Database of Intentions” can be abused by the government.