CPS 49S Google: The Computer Science Within and its Impact on Society - Spring 2007
Homework 2

- Due date: Friday, Feb 16, 2007, 11.59 PM. Late submissions will not be accepted (unless there are documented excuses from the dean).
- Submission: In class, via email to shivnath@cs.duke.edu, or via Blackboard's digital dropbox.
- 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 2**

What is googlewhacking?

Question 2 **Points 5**

If we search for "duke football" using the Archie search engine, will the http://www.goduke.com web page be part of the results produced? Why or why not?

Question 3 **Points 5**

When a user of Google says that Google "has good performance", what could be some measures of performance that the user is referring to? List at least two such measures.

Question 4 **Points 5**

One of the problems with the WWW Wanderer was that "it ate up too many processing and bandwidth cycles as it indexed a site's contents." Suggest two techniques that a crawler can use to avoid this problem.

Question 5 **Points 5**

Battelle says that Altavista was the Google of its era. However, it would be fair to say that Altavista was ultimately a failure. In your opinion, what were the three main causes of Altavista's downfall?

Question 6 **Points 5**

Even in its very early days, Altavista was able to "set a thousand crawlers loose at once."

1. What made this feature possible?
2. Why was this feature useful?

Question 7 **Points 5**

Excite was the first search engine that grouped web pages based on their underlying concepts. Give one concrete example that illustrates how such groupings can improve the quality of search results.

Question 8 **Points 5**

Imagine that you have graduated from Duke, and you are now working at Google. Your boss says "Garfield's impact factor is an excellent measure of the impact of a journal". You disagree, and you claim that Garfield's impact factor is flawed. Give two arguments that you will use to support your claim.

Question 9 **Points 5**

In its early days, Google (or back then, BackRub) caused numerous problems for other web servers on the Internet. List four types of problems that Google caused.

Question 10 **Points 5**

In Chapter 4 of the textbook, Steve Hansen recommends that BackRub should do "self-policing". Give one way in which BackRub could have self-policed itself.

Question 11 **Points 5**

What is "search-engine optimization"? How is it related to (search-engine) spamming?

Question 12 **Points 5**

One of the readings says: "any search engine on the Web must address the heterogeneity of HTML documents."

1. What does the term "heterogeneity of HTML documents" mean in this context?

2. Give three ways in which Google deals with this heterogeneity.

Question 13 **Points 5**

June Levy, managing director of Cinahl, says that "manual indexers are able to pick up on the nuances of human language that machines simply cannot do." Would you agree? Justify your answer using one or more concrete examples.

Question 14 **Points 5**

One of our readings suggests that it is hard for an automated indexer to "accurately forge relationships between documents that on the surface are not lexically linked". Justify this statement using a concrete example.

Question 15 **Points 5**

One of our readings says that "Google's web crawler grabs around 100K of text per web page, while Yahoo pulls about 500K." What are the pros and cons of grabbing 100K Vs. 500K? (Here, K stands for KiloBytes.)

Question 16 **Points 5**

What is the Metathesaurus? Is the Metathesaurus useful for a search engine? Justify briefly.

Question 17 **Points 5**

In class we discussed that stemming has both positive effects and negative effects.

1. List two positive effects of stemming.

2. List two negative effects of stemming.

Question 18 **Points 5**

Imagine that you have graduated from Duke, and you are now working at Google. You have been made part of the "stop-word selection team" at Google. This team's job is to

determine which words from the English language should be made stop words for Google's indexer and search-engine. What criteria would you suggest to determine whether a word should be made a stop word or not?

Question 19 **Points 5**

One of the readings states that "there is a trade-off in systems that support contiguous word phrases and proximity measures—higher storage requirements and computational costs." Justify this statement briefly.

Question 20 **Points 8**

Based on the discussion in the "Anatomy of a ... Web Search Engine" paper, explain step-by-step how Google would process the search query "duke freshman seminar" to generate ranked results.