

# Empirical Research Methods in CS: FINAL EXAM

Noah Smith and David Smith

December 7, 2005

## 1 Ground Rules

You must turn in your exam electronically (pdf preferred) or in hard-copy (NEB 332, slide under the door if no one is there), *by noon on Saturday, December 17, 2005*. Late submissions will not be accepted.

You are to work alone on this exam. Do not discuss any part of it with anyone except the instructors.

## 2 Background (0 points)

In this course we ended up spending lots of time on three main topics: hypothesis testing, exploratory data analysis, and using statistical models in software. In hypothesis testing, we start with some data gleaned by experiment, and try to demonstrate that something interesting—something significant—is apparent by the data. In EDA, we start out with some data (usually of a more raw form, but not necessarily) and try to find any interesting trends worth considering more carefully. In modeling, we use data as a means to a practical end: software that performs a desired task.

In this exam, you will be focusing on EDA and hypothesis testing, but first you need to acquire data. In a nutshell, this exam asks you to compare three software tools (search engines) and try to discover interesting differences in their output. It might be helpful to imagine that, after taking the exam, you are going to build a program that looks at the ranked list of web pages returned by a search engine, and guesses which engine gave the results. Because you've just recently built a classification system—and because we see little practical use for this program—you won't build it. (Extra credit question: describe a use for such a program.)

## 3 Data-Gathering (25 points)

Come up with about 25 search engine queries. You should make an honest effort to come up with queries that people might submit, so think about the day's news and so forth. Make these queries *difficult*, because (after all) your goal is to track down differences between search engines. Typing in MICROSOFT is going to give lame results.

Submit each of these queries to Google ([www.google.com](http://www.google.com)), Yahoo ([www.yahoo.com](http://www.yahoo.com)), and MSN-Search ([www.msnsearch.com](http://www.msnsearch.com)). We suggest you turn off any personalization options that you have locally set for these engines. For each query/engine pair, consider only the first page of results. You should extract the ranked list of web results, including URLs and the contents of those pages. You may also extract advertisements. You should not extract anything on the page that signifies which search engine was used (e.g., formatting, logos, within-domain links that are stock content on query result pages, etc.). Starting with

a web page in the list, you are free to go wherever you like—you might want to download pages that it links to, for example.

The `wget` tool may be helpful in getting this data. Since `wget` is blocked as a robot by some sites, you can also use `lynx -source [url]`. We've checked that these programs work for the three search engines. *Please* let us know if you run into problems getting data.

**Deliverables:** Just your list of queries and a description of your downloading strategy: just URLs, pages, pages linked to pages, etc.

## 4 EDA (25 points)

What's different about the data returned by the engines? Some ideas to get you started: maybe one engine returns more diverse results. Maybe one engine returns more recently posted/edited web pages. Maybe one engine strongly prefers shorter URLs.

Your goal is to come up with a few interesting qualitative statements about the search engines, and back them up with quantitative measurements. You may ask, "but how do I measure diversity?" This is part of the exercise. Some quantities like time are easy to quantify (if not to measure), and others require some creativity. Speaking informally, a set of  $n$  web pages might be considered diverse if they were very different from each other. How would you measure that?

Your measurements should be objective. It would be nice if you could talk about the *quality* of the results returned by different engines, but it would be artificial and very time-consuming, so we aren't asking for that.

**Deliverables:** A page or two describing what appear to be the real differences in the search engine output, supported by quantitative measurements. Tables and graphs (and other visualizations like we talked about in class) are a plus. Convince us that there's really something going on.

**Caveat:** We're hoping that, as creative people, you will find some interesting things to say about your data. If you try a bunch of ideas and all the engines really seem to be exactly the same, try a different set of queries. Ask yourself this: are these engines all really the same? If you really believe they are, describe what your reasoning.

## 5 Hypothesis Testing (50 points)

Distill your quantitative statements into a testable hypothesis, then carry it out. We expect the following **deliverables**:

1. What's  $H_0$ ?
2. What is the alternative hypothesis?
3. What is the statistic of interest?
4. What statistical test did you use?
5. How probable is the result you observed assuming  $H_0$ ?

Given this information, the devious search engine analyst should be able to replicate your results. But don't worry: unlike the research done by graduate students for the university, your brilliant insights in coursework remain your own.