

EECS 498 W26 System Design of a Search Engine

Project final reports

Due Tuesday, April 21, 2026, 11:59 pm

Report

Please try for a **short** report that thoughtfully describes your experience and the results of your project. Shoot for about 10 pages + exhibits.

Sizing and functionality should also be reported on the shared Google spreadsheet.

1. Your title page must include your group photo.
2. Identify your group. List your names, one per line, with username, major, and class.
3. *Briefly* summarize how you approached the project.
4. Describe how you distributed your engine across multiple cloud machines and how they communicate.
5. Describe the index build side of your engine, how you constructed a seed list, manage the frontier and deal with politeness, how pages are turned into chunks, any tiering or ordering of chunks, and what data you collect. If you support anchor text, word-breaking, stemming, indexing PDFs, or other special features, explain how you did it.
6. Describe the query serve side of your engine. Focus on how you've approached the problem of getting good results fast.
 - a. Describe how you attacked performance, e.g., by ordering the search (e.g., best chunks first), using lots of machines, setting timers, or other strategies.
 - b. Describe your strategy for ranking. Describe what raw data is available to the ranker when a matching page is found, how your ranker scores the page, and how you tuned it either manually or with an ML approach.
 - c. Describe how you construct and return your result page. If you support snippets, numbered result pages like Google, or other special features, explain how you did it.
7. Describe how you dealt with crashes or complaints.
8. If you had it to do over again, what would you do differently? If you had more time, what would you like to improve?
9. Thoughtful reflections.

10. Include the following exhibits:

- a. A short timeline of major events, e.g., team formed, crawler works, entire engine works on one laptop, engine is distributed, etc. You decide what's major.
- b. A line-count table showing major components, HTML parser, crawler, constraint solver, ranker, etc., down the side and names across the top, with best guess estimates of who did what in LOC as counted by "wc -l". Show individual totals and percentages of the entire engine at the bottom and component totals and percentages in a column to the right. Bottom right corner is the total LOC in your entire engine. Keep it all on the same page, please.

(I understand that sometimes one person might rough out an initial version and then someone else might polish or rewrite it for the final product. But do your best to divide up ownership of the final numbers of lines, perhaps by agreeing on percentages and then multiplying. I'm not asking you to separate it out, so you get to count any lines I gave you in starter code.)

- c. Basic specifications, e.g., number of virtual machines and their configurations, number of documents, number of words, number of unique words, seed list size, size of the frontier, crawl speed, chunk size, index size, query language, query serve time, etc.
- d. Features: A table or list of important features that distinguish your engine.
- e. A *few* important diagrams of your choice explaining the structure of your code or data or system architecture, e.g., source tree, index file format, enddoc contents, message formats between machines, layout in the filesystem. Emphasis on *few*. Choose wisely. Lots of extra stuff that really isn't that interesting or duplicates diagrams from my lectures will not get a better grade.
- f. Screenshots of your engine running a small number of TBD queries (which you should run "cold", without having tuned your engine or previously cached the result for that query.) Your results page must show elapsed (wall clock) time to run the query. Since I may print these pages and for better readability anyway, please do not use "night mode" with black backgrounds. Images must be LARGE and easy to read, with LARGE fonts.