Grading Information

Undergraduate Students

Undergraduate students are required to conduct experiments under several different conditions. The experimental conditions and results are described in a written report. The distribution of points varies somewhat across the different assignments, but generally the breakdown is as follows.

- 88%: Software implementation (performance on hidden test cases)
- 12% Conducting experiments (baselines, parameters, experimental results)

The report template for each assignment provides guidance about what to report and how to report it.

Grades for the software implementation tend to be high because the training cases help people produce correct implementations. Most people receive full credit for the experiments, as long as the design is reasonable, and the results are reproducible; there is no credit if the results are not reproducible.

In most cases, grades for experiments do not depend on the performance of your system or whether an experimental system beats a baseline. Some of your ideas may not work; that is okay, as long as they are reasonable ideas.

Students are welcome to discuss their reports with the TAs during office hours to help understand how to make improvements. We recommend having these discussions early in a homework cycle when everyone is less busy. At the last minute, TAs will prioritize students that need help with software design (more points) over students that need help with reports (fewer points).

Graduate Students

Graduate students are required to conduct experiments and analyze the results to reach conclusions about search engine components and behavior under different conditions. The experiments, analyses, and conclusions are described in a written report. The distribution of points varies somewhat across the different assignments, but generally the breakdown is as follows.

- 50%: Software implementation (performance on hidden test cases)
- 17%: Conducting experiments (baselines, parameters, experimental results)
- 33%: Analysis of results (discussion and analysis of the experimental results)

The discussion and analysis sections are graded along two dimensions: Breadth and depth.

The breadth dimension evaluates the range of issues or behaviors considered in your analysis. Discussions are graded along a continuum that extends from ‘thin’ (weak for a graduate course) to ‘good’ (a typical discussion) to ‘thorough’ (unusually good).

The depth dimension evaluates the quality of the discussion of each issue or behavior. Discussions are graded along a continuum that extends from ‘superficial’ (weak for a graduate course) to ‘good’ (a typical discussion) to ‘insightful’ (unusually good).

‘Thin’ and ‘superficial’ discussions tend to focus on one or two behavior(s) or result(s) that are obvious from glancing at the experiments results (e.g., method A is stronger than method B), or describe how an algorithm works (which is not discussion of the results). They indicate little or no attempt to learn lessons about search engine design or behavior from the experiment.

‘Thorough’ and ‘insightful’ discussions tend to focus on several behaviors or results and to reach conclusions that require a bit of thought about how the observed behavior is related to the design of the search engine, especially the components that are the focus of the assignment. ‘Thorough’ and ‘Insightful’ do not mean ‘long’. They mean ‘careful thought’, not ‘wrote a lot’.

The report template for each assignment contains a few suggestions to help you get started, however you are free to write about any aspect of the experiment that interests you. Don’t obsess over what issues we expect. Focus on how the experimental results inform you about search engine design or behavior. Justify your conclusions. Grading is based on the breadth and depth of your discussion, not whether you picked the right three issues.

The distribution of points varies across the three sections. Grades for the software implementation tend to be high because the training cases help people produce correct implementations. Most people receive full credit for the experiments, as long as the design is reasonable, and the results are reproducible; there is no credit if the results are not reproducible. Discussions are graded on a Gaussian distribution, so most reports land in the ‘Good’/’Good’ range; think of ‘Thorough’/’Insightful’ as extra credit.

In most cases, grades for experiments and discussions do not depend on the performance of your system or whether an experimental system beats a baseline. Some of your ideas may not work; that is okay, as long as they are reasonable ideas.

Students are welcome to discuss their reports with the TAs during office hours to help understand how to make improvements. We recommend having these discussions early in a homework cycle when everyone is less busy. At the last minute, TAs will prioritize students that need help with software design (more points) over students that need help with reports (fewer points).