

Functionality comparison	Points available	Team 1	Team 2	etc.
Level 1: Parse HTML files				
Parses HTML documents to extract words and links.	5			
Level 2: Build a crawler				
Number of pages crawled. linear: 500M = 20	20			
Number of machines involved in crawling. 25 = 5	5			
Retrieves documents by HTTP and HTTPS.	5			
Parses the HTML to find links words to be indexed.	5			
Manages a frontier of URLs to be explored.	5			
URLs on the frontier are prioritized.	5			
Seed list size and how it was created. 1000 = 20 max.	20			
Considers distance from the seed list when crawling.	5			
Polite. Obeys robots.txt and does not DOS anyone.	10			
Deals with redirects and common errors.	5			
Parallelized by multithreading.	5			
Able to identify and suppress loops.	5			
Crawler is restartable.	5			
Has automatic crash recovery.	5			
Level 3: Reverse word index				
Number of documents indexed. 500M = 20	20			
Source or parsed HTML forms of indexed pages is retained.	5			
Number of machines the index is spread across. 25 = 5	5			
Accumulates useful statistics.	5			
Accumulates useful information about each document.	5			
Provides a dictionary to find the posting lists for words.	5			
Word locations are numbered start of the document and a location is skipped between documents.	5			
Locations are encoded as deltas in a posting list.	5			
Variable length encoding is used for deltas.	10			
Uses a seek table to allow fast seeking of ISRs.	10			
Supports word posting lists.	5			
Supports endoc posting lists with useful attributes.	5			
Uses decoration.	5			
Words in the body, title, and URL are indexed on separate posting lists.	5			
Words in anchor text are also indexed onto a separate posting list.	10			
Provides an index stream reader class that can seek.	5			
Level 4: Create a user interface				
Provides a simple command line interface.	5			
Provides a simple HTTP server as a wrapper UI.	10			
Report clickable URL for result pages.	5			
Reports title for result pages.	5			

Reports a snippet with each result.	5
Number of machines searching in parallel. 25 = 5	5
Fast query response. < 1 s = 20, < 5 s = 15. < 10 s = 10	20

Level 5: Constraint solver and query parser

Supports AND, OR, phrase, and container ISRs.	10
Supports compiling a query into a structure of ISRs.	10
Query parser supports AND, OR, NOT, phrases and parentheses.	10
Able to compile a query and find matching pages.	10

Level 6: Build a ranker

Ranker considers static page attributes	10
Matches in the title are given more weight than in the body text of a page.	10
Matches in the URL are considered in ranking.	10
Matches in the URL not already indexed are found with string search during ranking.	10
Matches in anchor text are considered in ranking.	10
Ranker uses heuristics or other methods that consider where the words appear, proximity, or ordering, e.g. exact phrase in the title.. Describe briefly.	20
Demonstrate ability to produce a useful 10 best search results.	30

Level 7: Distribute the engine

Crawling, indexing, and query processing are distributed across a small network of laptops or cloud machines.	10
Crawls from multiple machines without overlapping. Describe.	5

Level 7: Advanced features

Ranker has been tuned by machine = 10 or by hand = 5	10
Ranking parameters can be changed without recompiling or restarting the engine.	5
Ranker uses a neural net. Jackpot. Describe.	10
Provides snippets to go with reported matches.	5
Supports stemming and stop words.	5
Supports PageRank or similar.	5
Supports Gzip in crawling.	5
Indexes PDFs or other filetypes.	5
Identifies and suppresses spam, porn, non-English content. Describe.	5
Uses shingles or other technique to suppress duplicates.	5
Long-running queries terminated. (Describe.)	5
Does automatic recrawl.	5

Total

500

Class max

Class average

Class min

Index comparisons

Team 1

Team 2

etc.

You may enter comments into a cell and reformat it if you like.

Seed list size in URLs

Number of crawler threads per machine

Number of crawler machines

Estimated overall crawl speed (documents/hour)

Number of documents crawled

Number of documents indexed

Total bytes of source documents

Number of tokens in the index

Number of unique tokens

Chunk size

Number of chunks

Total size of the index and other files

Number of index machines running a query.

Estimated average query time (ms)