

Name:

ID Number:

Exam Number:

Grade: 1: ... 2: ... 3: ... 4: ... 5: ... 6: ... 7: ... Total:

SOLVE **ALL the problems** IN THE SPACE PROVIDED

Read the Problems CAREFULLY!

THERE ARE 6 (SIX) PAGES THIS PAGE INCLUDED
Read and Sign the statement below at the end of the exam
Unsigned exams will be marked with grade 0 (zero).

STATEMENT

On my honor, I pledge that I have not violated the provision of the NJIT Student Honor Code.

Sign below at the end of the exam

Signature

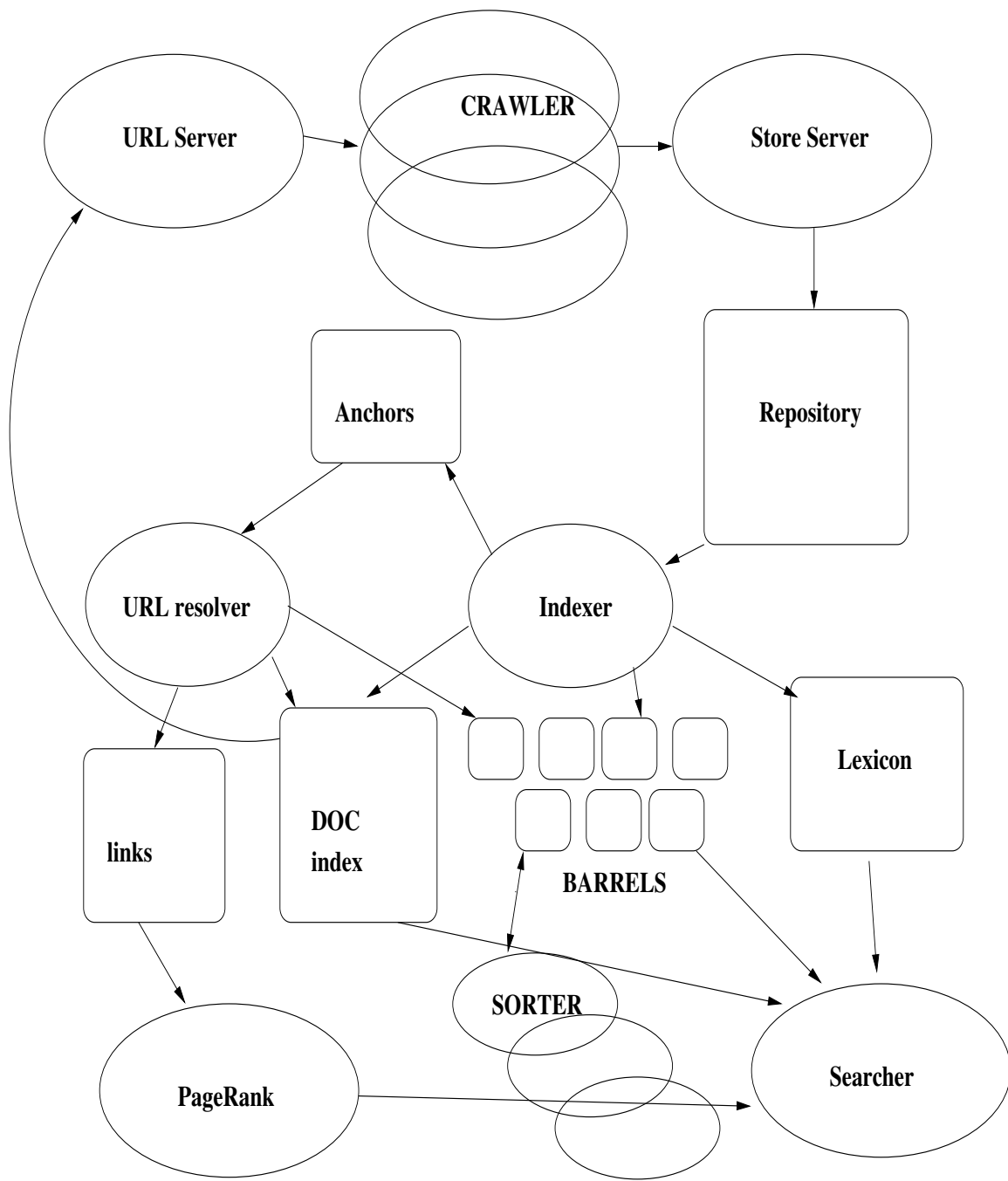


Figure 1: Google System Architecture.

Problem 1. (40 POINTS)

Give short answers to the following questions. (Use the Google System Architecture figure if you need it, for references related to Google.)

- (1) What information does the **URLserver** provide to the Crawler?
- (2) What is the functionality of the **Store Server**?
- (3) Where does Google store retrieved (by the crawler) web-pages?
- (4) Name (do not further discuss) 5 data structures used by Google.
- (5) What is the functionality of **links** in Google?
- (6) What is the input for the **PageRank** computations?
- (7) Where does Google store the forward index?
- (8) Where does Google store the inverted index?
- (9) What do we call a frequently occurring word in a text that is not a keyword?
- (10) What do we call the operation that extracts the root of a word (and thus eliminates) its suffix/ending?

Problem 2. (20 POINTS)

Provide brief answers (one or two sentences maximum) to the following questions.

- (a) What is the difference between a full-text logical view of a document and an index-term logical view?
- (b) Is every word in a document a keyword, and is every keyword an index-term? Explain, or explain the differences of the expressions **word**, **keyword**, **index-term**.

Problem 3. (20 POINTS)

For a specific query involving term t , the corpus (collection of documents) includes the ten documents $d_1, \dots, d_{10}$. The first five $d_1, \dots, d_5$ contain index term t and can thus be relevant, whereas the remaining five $d_6, \dots, d_{10}$ do not contain t at all.

As a response to the query the search engine returned documents

$$d_1, d_3, d_8, d_4, d_7, d_9$$

in that specific order.

- (a) What is the precision of the search query? Explain.
- (b) What is the recall of the search query? Explain.

Problem 4. (12 POINTS)

Which of the AND, OR, NOT are operators in (i) Google, and (ii) Bing?

Problem 5. (8 POINTS)

Give the unary and Elias- γ codes for 5.

Problem 6. (25 POINTS)

You have a machine with 4000MBytes of free memory. We want to use all this memory to store a hash table representing the lexicon of a search engine. The average length of a word w in the lexicon is around 11 characters UNICODE. The hash table will store a pointer p that points to a location in memory where the string word w can be stored; pointer p is 4 bytes long. In addition to p a `wordID` needs to be locatable in addition to w . Organize the hash table so that you extract from it maximum performance. How many words w can the lexicon accommodate? Explain.

Problem 7. (25 POINTS)

Apply the simhash algorithm on the text using b vectors with $b = 4$. (Suggeston: Use first vectors, 0101, 1010, 0011, 1100, 0110 etc before other 4-bit vectors.)

web search index web term web search term

Intentionally left blank. You can use it for scratch paper by tearing it off carefully; do not return it then.

End of Exam 1