Note: The questions listed here are to give you an idea about the kind of questions that would appear on your final. They are not an indicator of the length or format of the Final.

1. Regular expressions

a. Encode zero or more occurrences of 'a' followed by 'b'
b. Write a regular expression for the set of all strings consisting of a single digit followed immediately by a word (an alphabetic string) whose first letter is capitalized.
c. Exercise 2.5 from J&M International Edition
Design an FSA that accepts only valid web addresses. Be sure to accept addresses from the “.com” and “.org” domains, and handle an arbitrary amount of directory nesting. Accept at least “.html”, “.htm” and “.shtml” page types. Consult the official Web standards for other possible extensions to this recognizer.

2. Language Modeling

a. In a bigram (n=2) language model, the probability of the sentence 'I saw the red house' is approximated as?
b. From figure 12.3 in MRS,

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.2</td>
</tr>
<tr>
<td>a</td>
<td>0.1</td>
</tr>
<tr>
<td>frog</td>
<td>0.01</td>
</tr>
<tr>
<td>toad</td>
<td>0.01</td>
</tr>
<tr>
<td>said</td>
<td>0.03</td>
</tr>
<tr>
<td>likes</td>
<td>0.02</td>
</tr>
<tr>
<td>that</td>
<td>0.04</td>
</tr>
<tr>
<td>dog</td>
<td>0.005</td>
</tr>
<tr>
<td>cat</td>
<td>0.003</td>
</tr>
<tr>
<td>monkey</td>
<td>0.001</td>
</tr>
</tbody>
</table>

..........
Calculate the probability for the sentence 'The dog likes that monkey' using both the (unigram) models.

3. Stemming

*Exercise 2.4 from MRS*
For the Porter stemmer rule group shown in (2.1):
  a. What is the purpose of including an identity rule such as SS ->SS?
  b. Applying just this rule group, what will the following words be stemmed to?
circus canaries boss
  c. What rule should be added to correctly stem pony?
  d. The stemming for ponies and pony might seem strange. Does it have a deleterious effect on retrieval? Why or why not?

4. Information Retrieval

*Exercise 6.8 from MRS*
Why is the idf of a term always finite?

*Exercise 6.18 from MRS*
One measure of the similarity of two vectors is the Euclidean distance (or L2 distance) between them. Given a query q and documents d1, d2 . . . we may rank the documents di in order of increasing Euclidean distance from q. Show that if q and the di are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.

*Exercise 6.19 from MRS*
Compute the vector space similarity between the query “digital cameras” and the document “digital cameras and video cameras” by filling out the empty columns in Table 6.1. Assume N = 10,000,000, logarithmic term weighting (wf columns) for query and document, idf weighting for the query only and cosine normalization for the document only. Treat and as a stop word. Enter term counts in the tf columns. What is the final similarity score?
5. PageRank and Link Analysis

Exercise 21.5 from MRS
Write down the transition probability matrix for the example in Figure 21.2.

(Simple PageRank example) from MRS
Walk through the example in the text, solving exercise 21.6 for alpha=0.5.

Exercise 21.10 from MRS
Show that the PageRank of every page is at least alpha/N. What does this imply about the difference in PageRank values (over the various pages) as alpha becomes close to 1?

6. Sentiment Classification

We've seen positive/negative sentiment classification in class. Think of features that can be useful to determine polar/neutral classes. In general is it a easier task or more difficult and reason why.

7. Machine Translation

Consider the following corpus of three English phrases and their Swahili translations:

- long book
  - kitabu mrefu
- long box
  - mrefu sanduku
- big box
  - sanduku kubwa

Provide the initial matrix of translation probabilities and the new matrix of translation probabilities after one round of applying EM. Please show all computations involved.
Figure 6.7 from J&M
Finish the computations in Figure 6.7 (the one whose caption starts "The forward trellis for computing the total observation....", and tell us the final forward value alpha3(end) (also called alpha3(qf))