Information Storage and Retrieval Exam 2 Name: _____

1. Which classic retrieval model uses an algebraic rationale and computation to determine returned documents?
2. Which classic retrieval model uses an set theory rationale and computation to determine returned documents?
3. Which classic retrieval model requires learning based on relevance feedback?
4. Which classic retrieval model was used the most in pre-Web IR engines?
5. Which classic retrieval model is used the most in modern (post Web) IR engines?
6. How is a document represented in the vector model?
7. How are document's compared to the query in the vector model?
8. How is term frequency (the tf factor) calculated given a term and a document? (vector model)
9. How is inverse document frequency (idf factor) calculated for a given term? (vector model)
10. How is the weight for a given term and document calculated in the vector model?
11. How are the weights for the query computed in the vector model?
12. What is the initial probability for finding a given term in a relevant document? (probabilistic model)
13. What is the initial probability for finding a given term in a non-relevevant document? (probabilistic model)
14. Explain the basic concept behind latent semantic analysis (and LSI)?
15. What is the vocabulary problem motivating the LSI model?
16. What are polysemy and synonymy and how does each impact retrieval?
17. How does linear discriminant analysis differ from latent semantic analysis?
18. What are the two forms of modifying a query during query reformulation?
19. How should you compute the perfect query to retrieve a given set of documents from a given collection? (vector model)
20. How do the Standard Rochio and Regular Ide variations on query reformulation differ?
21. What is the primary difference between Dec_Hi Ide and other query reformulations?
22. How should you evaluate relevance feedback strategies?
23. Why would you use an automatic query reformulation technique?

24. What is the difference between local analysis and global analysis reformulation techniques?

25. When does the computation occur for local analysis and global analysis reformulation techniques?

26. How do the Association Clustering technique, the Metric Clustering technique, and the Scalar Clustering technique differ?

27. What forms of synonyms can the Scalar Clustering technique recognize that the other two cannot? (why?)

28. How do the Similarity Thesaurus and Statistical Thesaurus differ?

29. How is the Similarity Thesaurus computed? Describe relative to standard vector model's use of term frequency and inverse document frequency?

30. How is the Statistical Thesaurus computed? What determines the creation of document clusters? How are terms selected for a document cluster?

31. How do numbers, hyphens, puntuation, and letter case create issues for lexical analysis?

32. How many bits should a symbol that occurs with probability p be assigned? (in theory)

33. What are the pros and cons of the adaptive statistical model of compression?

34. What are the pros and cons of the static statistical model of compression?

35. What are the pros and cons of the semi-static statistical model of compression?

36. What are the pros and cons of using words as symbols for compression in an IR context?

37. What is the difference between a general Huffman tree and a canonical Huffman tree?

38. Build the canonical Huffman tree for "for each rose, a rose is a rose".

39. How do dictionary methods of compression work?

40. Describe the Ziv-Lempel code.

41. Give the gzip (LZ77) Ziv-Lempel code for "peter_piper_picked"

42. Why are adaptive dictionary methods, like the Ziv-Lempel code, not very useful in an IR context?

43. What are the primary components of an inverted file (inverted index)?

44. What is the purpose of block addressing in inverted files? How does it work?

45. How are inverted files larger than the main memory of the computer generated? What is the time complexity for a collection of size n (in words) on a computer with memory M?