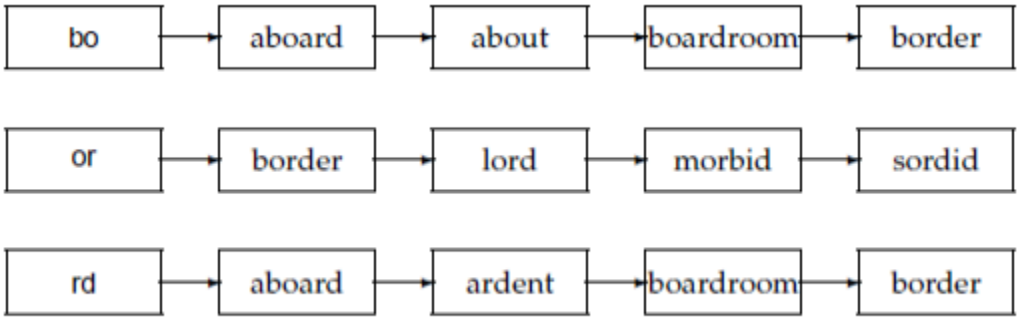
Assignment 2

1. If you wanted to search for s*ng in a permuterm wildcard index, what key(s) would one do the lookup on?

2. For the strings "Saturday" and "Sunday", find the minimum edit distance and one possible sequence of edit operations to transform the first string into the second.

3. Compute the Jaccard coefficients between the query ‘bord’ and each of the terms in that contain the bigram or.



4. For n=15 splits, r=10 segments and j=3 term partitions, how long would distributed index creation take for Reuters-RCV1 in a MapReduce architecture? Base your assumptions about cluster machines on table 4.1.

Table 4.1 Typical system parameters in 2007. The seek time is the time needed to position the disk head in a new position. The transfer time per byte is the rate of transfer from the disk to memory when the head is the right position.

Symbol	Statistic	Value
s	average seek time	$5\ \text{ms} = 5 \times 10^{-3}\ \text{s}$
b	transfer time per byte	$0.02\ \mu\text{s} = 2 \times 10^{-8}\ \text{s}$
	processor’s clock rate	$10^9\ \text{s}^{-1}$
p	lowlevel operation	
	(e.g., compare & swap a word)	$0.01\ \mu\text{s} = 10^{-8}\ \text{s}$
	size of main memory	several GB
	size of disk space	1 TB or more

Collection statistics for Reuters-RCV1. Values are rounded for the computations. The unrounded values are: 806,791 documents, 222 tokens per document, 391,523 (distinct) terms, 6.04 bytes per token with spaces and punctuation, 4.5 bytes per token without spaces and punctuation, 7.5 bytes per term, and 96,969,056 tokens. The numbers in this table correspond to the third line ("case folding")

Symbol	Statistic	Value
N	documents	800,000
L_{ave}	avg. # tokens per document	200
M	terms	400,000
	avg. # bytes per token (incl. spaces/punct.)	6
	avg. # bytes per token (without spaces/punct.)	4.5
	avg. # bytes per term	7.5
T	tokens	100,000,000

5. **a).** Consider a B-tree of order 3. Show the step-by-step process of inserting the keys 25, 50, 75, 15, 10, 60, and 5 into the tree.
b). Given the B-tree from Question 4, show the step-by-step process of deleting the key 50 from the tree. Illustrate the rebalancing steps as necessary.
c). Explain the relationship between the order of a B-tree and its height. How does increasing the order of the B-tree affect its height and overall performance?

6. a. Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3. Compute the tf-idf weights for the term's car, auto, insurance, best, for each document, using the idf values.

b. Can the tf-idf weight of the term in a document exceed 1? Explain your answer.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Table 6.1 Table of term frequency

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Table 6.2 Example of idf values

Here we give idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

7. Discuss the significance of IDF in term weighting. How is IDF calculated, and what is its role in determining the importance of terms? Provide an example where a term with a high IDF value might be more valuable in retrieval.