Problem 1 [10 Points]. Plot rank-frequency curves (using a log-log graph) for words and bigrams in the Wikipedia collection posted under http://students.cs.byu.edu/~cs453ta/projs/wiki.rar. Plot a curve for the combination of the two, i.e., words and bigrams. What are the best values for the parameter $K$ for each curve? Justify your answers for full credit.

Problem 2 [8 Points]. Exercise 4.2

Problem 3. The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs, would you argue, should not be conflated? Explain your answer.

(a) [2 Points]. abandon/abandonment
(b) [2 Points]. absorbency/absorbent
(c) [2 Points]. marketing/markets
(d) [2 Points]. university/universe
(e) [2 Points]. volume/volumes

Problem 4 [6 Points]. Exercise 5.2

Problem 5 [6 Points]. Exercise 5.3

Problem 6 [10 Points]. Construct a $B^+$-tree for the set $\{2, 3, 6, 7, 18, 21, 23, 32, 37, 39\}$ of search-key values. Assume that the tree is initially empty, values are added in the given order, i.e., from left-to-right, and the number of pointers that will fit in one node is four, i.e., $p = 4$. 