Due Date: No later than start of class on Mon Nov 21, 2011.

```
Document 1 (docID 1):
A sentence is a group of words. Sentences have a subject and a verb.

Document 2 (docID 2):
In a sentence find the verb and then find the subject.

Document 3 (docID 3):
To find the subject ask who or what followed by the verb.
```

Problem 1. (30 points)
Provide v-codes for $28,288$ in hexadecimal.

Problem 2. (30 points)
Using the following stopword list:
`a and in is of the then to`,
for the three documents shown above, show the form of the occurrence lists if (a) Doclist is used, (b) Counts is used, (c) Positions is used.

Problem 3. (35 points)
For Problem (2c) i.e. Positions, what would an implementation of vocabulary and the inverted list table look like ?

Problem 4. (30 points)
You are given an inverted list for a term t that looks like

`(1,1)(1,5)(1,11)(1,20)(2,10)(2,18)(2,30)(3,11)(3,25)(4,90)(4,91)`

Use the discussion of page 24 of Subject 7 to v-byte encode this inverted list by providing the following information. (Note that v-byte encoding is also discussed on 150 of the textbook. However, the example of the textbook incorrectly applies gap encoding to a count field shown here in bold-face, contradicting the algorithm stated there or in the notes: *1,2,1,6,1,***2***, 6, 11, etc.*)
(a) Give the flattened (no parenthesis) form of the to be v-byte encoded list just before the v-byte encoding is applied when the list is flattened but still in decimal notation, and
(b) after the v-byte encoding has been applied and the list given in hexadecimal notation.