Homework Assignment 4

October 20, 2013

1 Word sets again

Consider again the following 2-sentence text (from the Penn Treebank, WSJ section):

```
ghci 1> let text = "Pierre Vinken , 61 years old , will join the board " ++
          "as a nonexecutive director Nov. 29 .\nMr. Vinken " ++
          "is chairman of Elsevier N.V. , the Dutch publishing group ."
```

A. write a function to extract word sets from a text; the function should explicitly test whether a word is in the sentence or not, i.e., you shouldn't use a function that Haskell already provides to do this. Hint: use an **if** − **then** − **else** expression and a fold.

B. generate the word set for the above text using this function

C. now write a function to extract word sets from a text using the data structures and functions in the *Data.Set* module

D. generate the word set for the above text using this function

2 [Optional] Apply all this to the Brown corpus

Download the Brown corpus:

`http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/brown.zip`

Unzip it, and pick a file. Split the text in the file into words (word/tag pairs actually), remove duplicate words, i.e., generate the set of words in the text and also count the number of times each word occurs in the text.