Homework Assignment 2

October 8, 2013

1 Lines, words, checking for words in lines

Consider 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. split this text into lines (on "\n"), extract the first sentence, then the second sentence, then print the list of words for the two sentences

B. check if the word "Vinken" is one of the words in the first sentence and separately, if it's one of the words in the second sentence

C. check if the word "chairman" is **not** one of the words in the first sentence and separately, if it's **not** one of the words in the second sentence

2 Word sets

A. find a function in the module *Data.List* that enables you to extract the set of words in each of the 2 sentences (i.e., remove the duplicates) – and print the 2 sets of words. Hint: take a look at the posted lecture notes, find the one that introduces the *Data.List* module and take a look at the functions discussed there.

B. find a way to extract the set of words in both sentences (no duplicates) and print it

C. find a way to count how many comma tokens occur in the text; do the same for definite article "the". Hint: use pattern matching and guards.

D. find a way to count the tokens for every word that occurs in the text and print the resulting counts. Hint: use the word set and the token counting function applied to each word in the word set.

E. using the same functions, count the word/tag pairs in the following text from the Brown corpus.

```
ghci 2> :l textBrown
```