Homework Assignment 2 – Solutions

October 20, 2013

We load the 2 texts. We have to load the text in the file first because ghci wipes all bindings every time a file is loaded.

\texttt{ghci 1> :l textBrown}

\texttt{ghci 2> let text = "Pierre Vinken, 61 years old, will join the board " ++
"as a nonexecutive director Nov. 29. \\
Mr. Vinken " ++
"is chairman of Elsevier N.V., the Dutch publishing group."}

1 Lines, words, checking for words in lines

A. Split this text into lines (on 
\textquote{n}), extract the first sentence, then the second sentence, then print the list of words for the two sentences

\texttt{ghci 3> let ls_text = lines text}

\texttt{ghci 4> ls_text}

["Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. \\
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group."
]

\texttt{ghci 5> ls_text !! 0}

"Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29."

\texttt{ghci 6> ls_text !! 1}

"Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group."

\texttt{ghci 7> let ws_s1 = words (ls_text !! 0)
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 – and print the 2 sets of words
B. Find a way to extract the set of words in both sentences (no duplicates) and print it

```
ghci 20> map words ls_text
[["Pierre","Vinken","","61","years","old","will","join","the","board","as","a","nonexecutive","director","Nov.","29","."],
["Mr.","Vinken","is","chairman","of","Elsevier","N.V.",","the","Dutch","publishing","group","."]]
```

```
ghci 21> let ws_text = concat (map words ls_text)
```

```
ghci 22> ws_text
["Pierre","Vinken","","61","years","old","","will","join","the","board","as","a","nonexecutive","director","Nov.","29","." ,
["Mr.","Vinken","is","chairman","of","Elsevier","N.V.",","the","Dutch","publishing","group","."]]
```

```
ghci 23> nub ws_text
["Pierre","Vinken","","61","years","old","will","join","the","board","as","a","nonexecutive","director","Nov.","29","." ,
["Mr.","is","chairman","of","Elsevier","N.V.",","Dutch","publishing","group"]
```

C. Find a way to count how many comma tokens occur in the text; do the same for the definite article “the”:

```
ghci 24> let { countToken :: (Eq a, Num b) ⇒ a → [a] → b;
  countToken _ [] = 0;
  countToken x (y:ys)
     | x ≡ y = 1 + countToken x ys
     | otherwise = countToken x ys
  }
```
D. Find a way to count the tokens for every word that occurs in the text and print the resulting counts:

\[
ghci 30> \text{let } \{ \text{countItemsInList} :: (\text{Eq } a, \text{Num } b) \Rightarrow [a] \rightarrow [a] \rightarrow [(a,b)];
\]
\[
\text{countItemsInList } [] = [];
\]
\[
\text{countItemsInList } (x:xs) ys =
\]
\[
(x, \text{countToken } x ys) : \text{countItemsInList } xs ys
\]

\[
ghci 31> \text{let } \{ \text{tokenCounts} :: (\text{Eq } a, \text{Num } b) \Rightarrow [a] \rightarrow [(a,b)];
\]
\[
\text{tokenCounts } xs = \text{countItemsInList } (\text{nub } xs) xs
\]

\[
ghci 32> \text{tokenCounts } ws_text
\]
\[
[\text{"Pierre"},1], [\text{"Vinken"},2], [\text{""},3], [\text{"61"},1], [\text{"years"},1], [\text{"old"},1], [\text{"will"},1],
\]
\[
[\text{"join"},1], [\text{"the"},2], [\text{"board"},1], [\text{"as"},1], [\text{"a"},1], [\text{"nonexecutive"},1], [\text{"director"},
\]
\[
[\text{"Nov."},1], [\text{"29"},1], [\text{"."},2], [\text{"Mr."},1], [\text{"is"},1], [\text{"chairman"},1], [\text{"of"},1],
\]
\[
[\text{"Elsevier"},1], [\text{"N.V."},1], [\text{"Dutch"},1], [\text{"publishing"},1], [\text{"group"},1]
\]

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