Learning to Program with F#
Exercises
Department of Computer Science
University of Copenhagen

Jon Sporring, Martin Elsman, Torben Mogensen, Christina Lioma

August 24, 2020

0.1 Random Text

0.1.1: The script `readFile.fsx` reads the content of the text file `readFile.fsx`. Convert this script into a function which can read the content of any text file and has the following type:

```
readText : filename:string -> string
```

Add this function to your library.

0.1.2: Add a function to your library which converts a string, such that all letters are converted to lower case, and removes all characters except a...z and space. It should have the following type:

```
convertText : src:string -> string
```

0.1.3: Write a function,

```
histogram : src:string -> int list
```

which counts occurrences of each of the characters `['a'..'z']@[' ']` in a string and returns a list. The first element of the list should be the count of 'a's, second the count of 'b's etc. Use this function to replace the mockup function in your library.

0.1.4: Write a white-box test of the library functions `readText`, `convertText`, and `histogram`, and add these to your test file.

0.1.5: The script `mockup.fsx` contains a number of functions including

```
randomString : hist:int list -> len:int -> string
```

The function `randomString` generates an identically and independently distributed string of a given length, where the characters are distributed according to a given histogram. The script is complete in the sense that it compiles and runs without errors, but its `histogram` function is a mockup function and does not produce the correct histograms.

Create the library file `textAnalysis.fs` and add the functions from `mockup.fsx`

0.1.6: The function `randomString` is not easily tested using white- or blackbox testing since it is a random function. Instead you are to test its output by the histogram of the characters in the string it produces.

Extend your library with the function,

```
diff : h1:int list -> h2:int list -> double
```

which compares two histograms as the average sum of squared differences,

$$\text{diff}(h_1, h_2) = \frac{1}{N} \sum_{i=0}^{N-1} (h_1(i) - h_2(i))^2 \tag{1}$$

where h_1 and h_2 are two histograms of N elements.

Extend your test file with a test of `randomString` as follows:

(a) Convert The Story using `convertText` and calculate its histogram.

(b) Use this to generate a random text using `randomString` with length N, where N is the length of the converted The Story.

(c) Calculate the distance between the histograms of The Story and the random texts using `diff`.

(d) Set the test as passed if the difference is below some threshold.

You must choose a threshold in the above, and you must argue for your choice.

0.1.7: Extend your library with the function

```
cooccurrence : src:string -> int list list
```

which counts occurrences of each pairs of lower-case letter of the English alphabet including space in a string and returns a list of lists (a table). In the returned list, the first element should be a list of the counts of 'a' being the initial character, i.e., how many times "aa", "ab", "ac",...,"az", "a " was observed. The second list should containt the counts of combinations starting with 'b', i.e., how many times "ba", "bb", "bc",... was observed and so on. The function should include overlapping pairs, for example, the input string "abcd" has the pairs "ab", "bc", and "cd".

0.1.8: Extend your test file with a white-box test of `cooccurrence`.

0.1.9: Extend your library with a function

```
markovChain : cooc:int list list -> len:int -> string
```

which generates a random string of length `len`, whose character pairs are distributed according to the cooccurrence histogram `cooc`.

0.1.10: The function `markovChain` is a random function, and you are to test its output by the cooccurrences of the characters in the string it produces.

Extend your library with the function,

```
diff2 : c1:int list list -> c2:int list list -> double
```

which compares two histograms as the average sum of squared differences,

$$\text{diff2}(c_1, c_2) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(c_1(i,j) - c_2(i,j)\right)^2 \tag{2}$$

where c_1 and c_2 are two cooccurrence histograms of N elements such that $c_1(i,j)$ is the number of times character number i is found following character number j.

Extend your test file with a test of `markovChain` as follows:

(a) Convert The Story using `convertText` and calculate its cooccurrence histogram.

(b) Use this to generate a random text using `markovChain` with length N, where N is the length of the converted The Story.

(c) Calculate the distance between the cooccurrence histograms of The Story and the random texts using `diff2`.

(d) Set the test as passed if the difference is below some threshold.

You must choose a threshold in the above, and you must argue for your choice.

0.1.11: The function `randomString` may be considered a zero-order Markov Chain model, since each generated character is independent on any previously generated characters. The function `fstOrderMarkovModel` generates new characters dependent on the previous character and is therefore a first-order Markov Chain model. Consider a similar extension to an n'th-order Markov Chain where the occurrences of n-tupples of characters are stored in `n : int list list ... list`. What possible pit-falls are there with this representation?

0.1.12: Extend your library with a function that counts occurrences of each word in a string and returns a list. The counts must be organized as a list of trees using the following `Tree` type:

```
type Tree = Node of char * int * Tree list
```

An illustration of a value of this type is shown in Figure 1 Words are to be represented as the

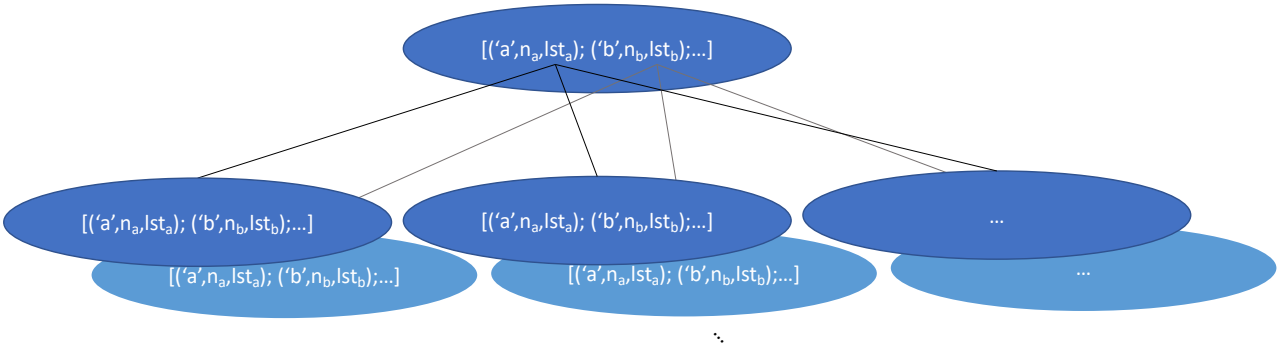

Figure 1: An illustration of a list of values of the type `Tree`.

sequence of characters from the root til a node. The associated integer to each node counts the occurrence of a word ending in that node. Thus, if the count is 0, then no word with that endpoint has occurred. For example, a string with the words "a abc ba" should result in the following tree,

```
[Node ('a', 1, [Node ('b', 0, [Node ('c', 1, [])])]);
 Node ('b', 0, [Node ('a', 1, [])])]
```

Notice, the counts are zero for the combinations "ab" and "b", which are words not observed in the string. The function must have the type:

```
wordHistogram : src:string -> Tree list
```

0.1.13: Write a white-box test of `wordHistogram`, and add it to your test file.

0.1.14: Extend your library with a function

```
randomWords : wHist:Tree list -> nWords:int -> string
```

which generates a string with `nWords` number of words randomly selected to match the word-histogram in `wHist`.

0.1.15: The function `randomWords` is a random function, and you are to test its output by the histogram of the words in the string it produces.

Extend your library with the function,

```
diffw : t1:Tree list -> t2:Tree list -> double
```

which compares two word-histograms as the average sum of squared differences,

$$\text{diffw}(t_1, t_2) = \frac{1}{M} \sum_{i=0}^{M-1} (t_1(i) - t_2(i))^2 \tag{3}$$

where t_1 and t_2 are two word histograms, and M is the total number of different words observed in the two texts.

Extend your test file with a test of `randomWords` as follows:

(a) Convert The Story using `convertText` and calculate its word histogram.

(b) Use this to generate a random text using `randomWords` with M words, where M is the number of words in the converted The Story.

(c) Calculate the distance between the word histograms of The Story and the random texts using `diffw`.

(d) Set the test as passed if the difference is below some threshold.

You must choose a threshold in the above, and you must argue for your choice.

0.1.16: In terms of a Markov Chain, what order is `randomWords`? Suggest an extension of `type Tree` to include first order Markov Chains. Speculate on whether this principle can be used to extend to n'th order mordels, and, in case, speculate on how the storage requirements will grow as n grows.

0.1.17: Write a short report, which

- is no larger than 5 pages;
- includes answers to questions posed;
- contains a brief discussion on how your implementation works, and if there are any possible alternative implementations, and in that case, why you chose the one, you did;
- includes output that demonstrates that your solutions work as intented.