The 19th International Conference on Web Engineering (ICWE 2019),
June 11 - 14, 2019 – Daejeon, Korea

Tutorial: Powerful Data Analysis and Composition with the UNIX Shell
Andreas Schmidt, Steffen Scholz

Exercise II

Tasks:

1. Download the book ’The Adventures of Tom Sawyer‘ from http://www.gutenberg.org/files/74/74-0.txt and store it locally in a file called ’The-Adventures-of-Tom-Sawyer.txt’

2. Count the words and lines in the book ’The-Adventures-of-Tom-Sawyer.txt’

3. What does the following command perform?

   ```
   egrep -o '[A-Za-z]+' The-Adventures-of-Tom-Sawyer.txt
   ```

4. Translate all words of ’The-Adventures-of-Tom-Sawyer.txt’ into lowercase using the command `tr`

5. Count, how often each word (ignore case) in this book appears (hint: use `sort, uniq`).

6. Order the result, starting with the word with the highest frequency. Which five words appear most frequently?

7. If not already done: Write all the above steps in one statement (using pipes)

8. Compare this result with the result from the book Moby Dick: http://www.gutenberg.org/files/2701/2701-0.txt.

   Compare the 20 most frequent words of each book. How many are in common? (hint: use `head, comm, ...`)