The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications, Mai 21 - 25, 2017 - Barcelona, Spain

Tutorial: Data Manipulation and Data Transformation using the 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 in this book appears (hint: use `sort, uniq`).

6. Order the result, starting with the word with the highest frequency. Which word is it?

7. Write all the above steps in one statement (using pipes)

8. Compare the result with the result from the following book: http://www.gutenberg.org/files/2701/2701-0.txt.
   At which positions (rank) appear the first book specific words?

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