The Tenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2018), May 20 - 24, 2018 – Nice/France

Tutorial: How to build a Search-Engine with Common Unix-Tools
Andreas Schmidt

Exercise I

.

Download the dataset BBCSport from http://mlg.ucd.ie/files/datasets/bbcsport-fulltext.zip und unzip the content in a directory. In the first exercise, the corpus of documents is to be statistically analysed.

1. Create a list of all words in the news documents in directory football. Each line should contain a single word (use `grep` or `tr`).

2. Transform all the words in lowercase and sort them alphabetically (use `tr, sort`).

3. Write the result in a file called *wordlist-sorted.txt* (outside the *bbcsport* directory).

4. What are the most frequent 50 words? Inspect the file with the command `less`. At which position, does the first topic specific word (topic football) appear? (use `head`, `uniq, sort, less with -N option`)

5. Take the 30 most frequent entries and write them into a file called stopwords-30.txt. The list should look like in Appendix A. (use `cut, sed, head`).

6. Extend step 1-5 to all articles in bbcsport (including tennis, athletics, ..).

7. How many percent of the whole text is covered by the 30 most frequent words? (about 1/3 of the text). (use `grep` with –f option, `wc, sed`)

Appendix A (stopwords-30.txt –ordered by decreasing frequency):

the
to
a
and
in
of
s
for
i
he
on
is
it
was
but
with
that
his
at
have
we
be
has
said
will
as
from
not
after
by

Hint: The solutions for this exercise can be found here:
http://www.smiffy.de/dbkda-2018/IR-exercise-1-solution.pdf