# Notes to the SD Index file

June 23, 2020

**How it is generated**

The index is generated from the TUP electronic edition in ASCII format (SDASCII.TXT) in several steps. The steps are all Unix shell script commands, except for step 7, where the actual indexing is done in a spreadsheet program. The following description is not exhaustive.

Step 0: the header and information is deleted manually, so that the text starts with the PREFACE on [[Vol. 1, Page vii]], and ends with the line "End of Vol. II.". (SDASCII - bewerkt - 0.txt) Please note that the original file does not contain the table of contents.

Steps 1-3: *irrelevant for the process of generating the index file.*

Step 4: All lines consisting of minus signs and indications of footnotes are deleted. All spaces are replaced by paragraph markers. All Volume and Page indication are marked with a "%" sign. (SDASCII - bewerkt - 4.txt)

Step 5: *irrelevant for the process of generating the index file.*

Step 6: All punctuation except the minus sign is removed. The resulting file contains about 688k lines. Each line now contains a word or a volume or page indication.

```
%Vol1
%Pagevii
PREFACE
THE
Author
the
writer
rather
feels
...
```

Step 7: The resulting file is imported into a spreadsheet where every word is linked to a volume and page number. (SDASCII - bewerkt - 7.ods) The file is exported again (as SDASCII - bewerkt - 7.txt).

Step 8: All words are converted to lower case. Lines starting with non-alphanumeric characters and empty lines are deleted. The file is sorted. (SD Index - 8.txt)

Step 9: All unique lines are counted, which means that we count the number of times the word occurs on the page. The resulting file only contains unique lines with words, volume, page and count-on-page numbers. It now contains some 350k lines. (SD Index - 9.txt)

Step 10: The 200 most frequently used words (Gutenberg) in the English language are filtered out. (Except for the word "I".) The file now contains some 250k lines. (SD Index - 10.txt)

Steps 11 and 12: Roman style page numbers are not sorted correctly in step 10.
In the original scanned text, some words contain scanning errors so that they start with digits (0 instead of O). Because lines starting with numbers are eliminated in the process, these words were also deleted. Fortunately this is occurs only in four occasions. The word "Vol" of the last line of volume II is eliminated during the process.

Using again a spreadsheet in step 11 and a few lines of BASIC, the sorting of roman page numbers is corrected. Manually the three other problems are corrected. The list is then sorted again in step 12. (SD Index - 12.txt)

### Remaining problems

All numbers are eliminated from the list, while they may be relevant for the study of the SD.

### Other notes

If the Linux sort command is used with a key, the key should apparently contain also a "to" value (also when only field numbers are used) otherwise numerical sorting gives unexpected results. One environment variable (locale) is said to be important:

```
env LC_ALL=C sort -k1,1d -k2,2d -k3,3g
```

Syntax of a key from "man":

> *KEYDEF is F[.C][OPTS][,F[.C][OPTS]] for start and stop position, where F is a field number and C a character position in the field; both are origin 1, and the stop position defaults to the line's end. If neither -t nor -b is in effect, characters in a field are counted from the beginning of the preceding whitespace. OPTS is one or more single-letter ordering options [bdfgiMhnRrV], which override global ordering options for that key. If no key is given, use the entire line as the key. Use --debug to diagnose incorrect key usage.*

Full documentation is available at: http://www.gnu.org/software/coreutils/sort

### Shell commands in "SD Index.sh" and "SD Index - 12.sh"

```
cat 'SDASCII - bewerkt - 7.txt' | awk '{print tolower($10), $7, $8}' | grep -v -E '^[0123456789 -]' | grep -v -E '^[^[:alpha:]]' | sed '/^%/d' | sed -E s/'"'//g | sed -E s/'[\+!{}]'//g | sed -E s/'--'//g | sed -E s/^\'//g | sed -E s/\'s//g | sed -E s/[\'][[:space:]]/' '/g | env LC_ALL=C sort -k1,1d -k2,2d -k3,3g > 'SD Index - 8.txt'

cat 'SD Index - 8.txt' | uniq -c  | awk '{print $2, $3, $4, $1}' | env LC_ALL=C sort -k1,1d -k2,2d -k3,3g > 'SD Index - 9.txt'

cat 'SD Index - 9.txt' | grep -s -v -E -i -w -f 'Gutenberg word frequency list 200 - 2.txt'  > 'SD Index - 10.txt'

cat 'SD Index - 11.txt' | env LC_ALL=C sort -k1,1d -k2,2d -k6,6g -k5,5g | awk '{print $1, $2, $3, $4}' > 'SD Index - 12.txt'
```