Exercise 2 – Reference Solution

Sprachverarbeitung (VL + Ü)

Nils Reiter, nils.reiter@uni-koeln.de

April 11, 2023 (Summer term 2023)

Our project Gutenberg dump contains two editions of Doyles' "The Valley of Fear". We want to study how they differ (if they differ).

- Find out their id numbers.
  - Executing `grep -i "Valley of Fear"/resources/gutenberg/GUTINDEX*` gives us an excerpt of all the index files, including the id numbers 3289 and 3776 for a Jun 2002 and Feb 2003 edition.
- Extract their word frequencies.
  - Word frequencies for the text 3289 can be extracted with `cat /resources/gutenberg/3/2/8/3289/3289.txt | tr '[[:punct:]]' ' ' | tr '[[:space:]]' '\n' | sort | uniq -c | sort -n`. The other text needs a different path. For case indifference (i.e., a and A are treated as the same letter), add `-i` to `uniq`.
- Inspect and compare them (manually). Do you think it's the same text?
  - Probably :)