Exercise 3 – Reference Solution

Sprachverarbeitung (VL + Ü)

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

April 18, 2023 (Summer term 2023)

Let's extract a concordance (from poe or any other text)!

- Insert a space before each line end
  - This can be done with `sed -E 's/$/ /g'`. `$` represents the line end, but since sed operates line-wise, it cannot be removed.
- Remove all line breaks
  - This can be done with `tr`: `tr -d '[\n\r\f]'`. `\n` represents the newline character, which is used on Unix operating systems (including Linux and Mac OS). `\r` represents the carriage return character and `\f` the formfeed character. See `https://en.wikipedia.org/wiki/Control_character` for details on them. In most cases, it's sufficient to specify `\n`, but it doesn't hurt to be on the safe side.
  - Alternatively, it can also be done with `sed` like this: `sed -E 's/[\n\r\f]//g'`.
- Unify all space to be a single space
  - Easiest to do with `sed`: `sed -E 's/[[:space:]][[:space:]]+//g'`.
- Feed the output into `grep -o` and inspect the concordance
  - The full pipeline then is `cat poe.txt | sed -E 's/$/ /g' | tr -d '[\n\r\f]' | sed -E 's/[[:space:]][[:space:]]+//g' | grep -E -o '.{20}blood.{20}'`. `.{20}` matches 20 arbitrary characters to the left and right of our search term (blood).
- Our query includes the context in characters. Can you extend it such that we get tokens?
  - The core idea here is to define a token as a sequence of some alphanumeric characters, followed by a space or punctuation symbol. With round parentheses, we can group such a sequence, and look for its repetition. This only works because of our preprocessing, and not in running new texts. Thus, we can use the following `grep` command:

    ```
    grep -o -E '([[:alnum:]]*[[:punct:][:space:]]){5}blood([[:space:][:punct:]]+[[:alnum:]]*){5}'
    ```

Query Ideas

- How does Poe write about men and women, how about cats and dogs?
- How did he use colors, e.g. red and green? What are things that are red, which things are green?
- Poe is a known horror author. Does he use the word “fear” as a noun or verb? In which contexts?