DATE 25 March 2024

Exercise Sheet 1: Text Analysis and Processing on the Command Line

H 1.1 Analysing lists of words

① grep -i "^jacky$" firstnames.jungs firstnames.maedels

② grep -v "[-]" firstnames.jungs firstnames.maedels | wc -l

③ yes, we can use (grep -iE "^[aieou]+$" firstnames.jungs firstnames.maedels) commad

H 1.2 Segmenting texts into word tokens

① sed 's/[.,:;!?]/ /g' ammersee.txt

② sed 's/ /\n/g' ammersee.txt

③ To see how much words in total: 228
sed 's/[.,:;!?()-]/ /g' ammersee.txt | tr -s ' ' '\n' | wc -w

To see total different words (Types): 160
sed 's[.,:;!?()-]/ /g' ammersee.txt | tr '[:upper:]' '[:lower:]' | tr -s ' ' '\n' | sort | uniq | wc -l

④ The ratio is calculated by dividing number of types by number of tokens
• from ammersee.txt we get 160/228 = $\frac{40}{57} \approx 0.17$

DATE 25 March 2029

another wikipedia with 3 paragraph, I get 196 for tokens and 38 for types. So the ratio is $38/196 \approx 0,67$.

From what I can see is, the ratio has tendency around 016-017.

H 1.3 Python and NLTK

① Done installation.

② Tokenization done.

③ you can see in my Github:
https://github.com/nabilfatih/nlp-exercise/blob/main/ue1/h1.3.ipynb