Exercise 1: Searching with Regular Expressions

Programming Techniques in Computational Linguistics 1
Programming for Linguists

Deadline: 11. October 2023 6 p.m.

Remarks on submission:

- Please submit the exercises in groups of two. If you can't find your assigned partner use the OLAT forum or write us an e-mail.
- File naming convention in general: olatusername1_olatusername2_pcl1_exercise0X.txt/pdf/py/zip. (The 'pcl1' in the filename is replaced by 'pfl' for students in the PfL group!)
- For this first exercise, you can list all the commands that you used in the command window all in the same file (.txt or .pdf).
- Please number the tasks on the submission sheet the same as on the task sheet.

1 Searching the Corpus of the Swiss Alpine Club (SAC)

Using regular expressions and the command line commands which are familiar to you from the lecture, solve the following questions about the SAC Yearbooks of the 1930s. You will find the necessary yearbook files in the OLAT folder 'Materials'.

- You must use the SAC-yearbooks for this task. The files are all called SAC_Jahrbuch_19XX_mul_columns.txt, depending on the year.
- The corpus is formatted in 3 columns (separated by the tabulator symbol) as follows: Word\tPart-of-Speech-tag\tLemma. The lemma is the base form of the word.
- Note that each sentence starts with a tag of the form <s lang="xx" n="1-1"> where xx is a code for the automatically recognised language of the sentence (e.g. de, fr, it, en, rm, or ch-de) and 1-1 is the article number and the sentence number.[1]
- For the word class tags, also called Part-of-Speech or PoS tags, (like VVFIN or NN) please consult the following documents:
  - German (STTS, Stuttgart-Tübingen Tag Set): `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/stts_guide.pdf`
  - French: `www.textberg.ch/ReleaseNotes/French_PoS_tagset_from_Le_Monde.htm`
  - Italian: `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt`
  - English: `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf`

[1] Sentences with lang="ch-de"have been marked as Swiss German dialect. Our language identifier had some problems to distinguish between standard German and written Swiss German dialect. Speakers of Swiss German are invited to check these sentences.

Please provide the commands that you used in the command line for your solutions.

a) How many words (tokens) in all the books start with the letter *m*?
Note: This means you should count both upper case '*M*' and lower case '*m*' together.

b) How often do tokens of German verbs end with 'en'?
Note: Don't forget to include the special characters for German: *üööäß*

c) Out of the two German verbs 'haben' and 'sein' (in all their conjugated forms), which occurs more often in the corpus? (Bonus points if you can do it with a single grep search)

d) How many words in the corpus have the same consonant appearing more than four times? As an example, the German word *Sonnentrunkenheiten* has 6 *n*. Search for unique occurrences in the lemmas.
Hint: You will need to use capturing groups, which are expressions surrounded by parentheses '()', and back-references (for example '\1').

e) Among the words with the same consonant appearing more than four times found above, which PoS tag appears the most, from which language is it?

f) What are the five most common German conjunctions in the corpus?
Hint: Most likely you will need to combine two grep commands. First, your search should look for all German conjunctions (Konjunktionen), and then match only the lemmas to count their occurrences.

g) Which are the three most frequently occurring german verbs whose lemma forms have the letter **'m'** in the second position of the word?
Hint: you only need one grep command. Don't forget to match all verb tags!

h) How many tokens in the corpus contain a dash (-)? What kinds of tokens did you find?

i) **Bonus task: you don't have to answer this to earn the full point, but it will help if you are missing some other task.**
Which German prefix of the following occurs most often with verbs: 'ab-', 'an-', 'auf-', 'aus-'? How can the frequencies be counted at the same time in one command? (Note: These prefixes can be separated from the verb in a German sentence.)

2 Searching a book from Project Gutenberg

Download a book from the collection of Project Gutenberg (www.gutenberg.org) as a text file (preferably German or English). Get a brief overview of the book by answering the following questions with the same tools as in section 1 above. (Please list the commands you have used to answer the questions):

a) How many lines in the book contain direct questions (lines ending with a question mark)? Extract the first five such lines. Do they give insights into the main topics or concerns of the book?

b) Characters sometimes have dialogues pre- or postfixed with their names, like 'Alice: How are you?'. identify if your book contains such a pattern, and extract the ten most frequently mentioned character names. Which character seems to be the most dominant or involved based on this?
Note: If your book does not contain this pattern, propose another way to identify character names in your book.

c) Identify if there are any lines written in all uppercase (this could be indicative of titles, headings, or emphasis). How many such lines are there? Provide the first three such lines you found.

d) Estimate the mood of the book by counting how many times positive words like 'happy', 'joy', 'love', and 'excited' appear compared to negative words like 'sad', 'sorrow', 'hate', and 'angry'. Which group of words appears more frequently?

Reflection/Feedback

a) Summarize your discoveries and your learning progress in two sentences.

b) How long did it take you overall to finish this exercise?