Text Processing Homework Assignment

Read the subsections 2.7.1-2.7.3 from https://web.stanford.edu/~jurafsky/slp3/2.pdf. It describes regular expressions. Explain how they can be used to find ICD-10 codes in a medical texts and what disadvantages in implies.

ICD-10 codes have a structured format:

- 3 to 7 digits
- Digit 1 is alpha
- Digit 2 is numeric
- Digits 3 thru 7 are alpha or numeric

Source: https://www.cco.us/what-are-icd-10-codes/

Therefore, a regex can capture this structure. For example:

r"\b[A-Z][0-9][A-Z0-9](\.[0-9A-Z]{1,4})?\b"

Disadvantage: it may capture strings that look like ICD-10 codes but aren't valid.

Discuss briefly why and when to use a stemmer, when one can have lemmatizer?

Stemming

Pros:
- Very fast, rule-based, no dictionary needed

Cons:
- Less precise, may connect unrelated words

Use Stemming when:
- We need speed over accuracy
- Working with very large datasets

Lemmatization

Pros:
- Handles irregular forms (went → go)
- More accurate

Cons:
- Slower, requires linguistic resources
- More Complex to implement

Use lemmatization when:
- Linguistic precision matters
- Handling irregular words is important