. means any character

Tutorial 1: Regular Expressions and Text Normalization

+ means 1 or more

* means 0 or more

\s is whitespace
\1 refers to first pattern inside the paraenthesis

^ means start of line
$ means end of line

two words grotto and raven can appear in any order + other strings around

Q1. Write regular expressions for the following languages. By "word", we mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth.

1. The set of all alphabetic strings; [a-zA-Z]+
2. The set of all lower case alphabetic strings ending with a letter b; [a-z]*b
3. The set of all strings with two consecutive repeated words (e.g., "Humbert Humbert" and "the the" but not "the bug" or "the big bug"); (\b[a-zA-Z]+\b)\s+\1
4. All strings that start at the beginning of the line with an integer and that end at the end of the line with a word; ^\d+\b.*\b[a-zA-Z]+$
5. All strings that have both the word "grotto" and the word "raven" in them (but not, e.g., words like "grottos" that merely contain the word "grotto"); (.*\bgrotto\b.*\braven\b.)|(.*\braven\b.*\bgrotto\b.*)

HINT: Not all notions are covered in lectures. Your RE may not fully satisfy the specified requirements.

* Dont write $ and ^ anchors if the exam question does not mention "start of line" and "end of line"

Q2. Design prompts to use a GenAI tool (e.g., ChatGPT) to write regular expressions for Q1.

- Test **your own answers** and the **answers provided by GenAI** tool on http://regexr.com/
- You may need to change the textbox to test two cases: (i) the textbox contains one or more matched strings, and (ii) the textbox does not contain any matched string.
- What are the errors (e.g., false positive and false negative) have you observed?

Q3. Select all strings that can be matched by regular expression /E*F+[^Gg]/

A. EFG B. EF C. FFF D. EFFa

[^Gg] can match any other character including eg. one whitespace

Q4. Use the following sentence as an example sentence to illustrate the tokenization results. You may also use a sentence in another language.

Mr. Sherwood said reaction to Sea Containers' proposal has been "very positive." In New York Stock Exchange composite trading yesterday, Sea Containers closed at $62.625, up 62.5 cents.

You may consider different packages or tools for illustration purpose.

https://huggingface.co/docs/transformers/en/main_classes/tokenizer

https://opennlp.apache.org/docs/2.5.5/manual/opennlp.html#tools.tokenizer.introduction

Q5. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 1) of “idea” to “deal”. Show your work.

Q6. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 2) of two sentences “computed the edit distance” to “the edit distance is computed”. Show your work and show the alignment between the two strings. You may use edit distance defined at **word level** instead of character level.