Task [3] Regularities in expressions

Linguists (especially computational linguists) working with large collections of texts, like corpora or websites, often need to know if a particular sequence of letters is present in some text. Such letter sequences are called strings (a string could be a phrase, a word or a part of a word). Finding a word or a phrase in a text is easy if it has only one possible form. But suppose that we want to find out if a text contains the noun "tree" either in a singular or in a plural form. We would have to conduct two separate searches for each form. That would get really tiresome for words with many forms (think about all the different cases a Polish noun can take, like *krzesło, krzesła, krzesłu, krzesłem, etc.*). To deal with such situations in an efficient way a set of conventions, called **Regular Expressions**, was developed to represent patterns matching the strings that we want to find. In addition to standard letters and numbers, Regular Expressions can contain special symbols. The table below lists some of those symbols and explains their functions.

SYMBOL	FUNCTION	EXAMPLE
+	one or more preceding characters	<hello+> matches "hello", "helloo", "hellooo", etc.
*	zero or more preceding characters	<hello*> matches "hell", "hello", "helloo", etc.
?	zero or one preceding character (optional character)	<trees?> matches "tree", "trees"
|	alternative (... or ...)	<a|b> matches "a", "b"

You can also use parentheses to apply the special symbols to strings with more than one character (<(ab)+c> matches strings like "abc", "ababc", "abababc", etc.; <(cat|dog)> matches the string "cat" or the string "dog").

EXERCISE 1

Look at the excerpt from a Wikipedia article on linguistics below. Using the table of special symbols above for reference, write down all strings in the text matching the provided Regular Expressions (RE 1, RE 2), including multiple identical strings. Bold words in the text match RE 0 given as an example.

> *Scientific discipline that analyzes human language as a system for relating sounds (or signs in signed languages) and meaning. Phonetics studies acoustic and articulatory* ***properties*** *of the* ***production*** *and perception of speech sounds and non-speech sounds. The study of language meaning, on the other hand, deals with how languages encode relations between entities,* ***properties****, and other aspects of the world to convey,* ***process****, and assign meaning, as well as to manage and resolve ambiguity.*
>
> [...]
>
> *In the early 20th century, Ferdinand de Saussure distinguished between the notions of* langue *and* parole *in his formulation of structural linguistics. According to him,* parole *is the specific utterance of speech, whereas* langue *refers to an abstract phenomenon that theoretically defines the principles and system of rules that govern a language. This distinction resembles the one made by Noam Chomsky between competence and performance, where competence is individual's ideal knowledge of a language, while performance is the specific way in which it is used.*
>
> [...]
>
> *Although the term "linguist" in the sense of "a student of language" dates from 1641, the term "linguistics" is first attested in 1847. It is now the common academic term in English for the scientific study of language.*

From <https://en.wikipedia.org/wiki/Linguistics>

Example:
RE 0: <pro(pertie|duction|ces)s?>
matching strings:
"properties", "production", "properties", "process"

RE 1: <stud(y|i)e?(nt|s)*>
matching strings:

....

....

RE 2: <l(a|i)ngu(ist)?(ic|age|e)s*>
matching strings:

....

....

EXERCISE 2

Each Regular Expression in the table below is supposed to match all the possible forms of some English verb. However, in each case there is at least one form of a given verb that does not match the RE pattern. Identify the verb, matching and non-matching forms for every Regular Expression. The solution for RE 0 is provided.

#	RE	VERB	MATCHING FORMS	NON-MATCHING FORMS
0	<work(ed|s)?>	WORK	work, worked, works	working
1	<see(n|s|ing)*>			
2	<ma(k|d)e(s|ing)>			
3	<wr(i|o)tt?(e|n)>			
4	<f(i|ou)nds*>			
5	<like(d|ing)>			
6	<read(s|ing)+>			
7	<fl(y|e|o)?w(n|s)*>			
8	<g(i)?av(e|ing)+>			
9	<kn(o|e)w(s|ing|n)>			
10	<match(s|ed|ing)*>			
11	<th(ought|ink)+(s|ed)?>			
12	<move(ing|d|s)*>			
13	<hid+(e|s|ing)?n?>			

Author: mgr Piotr Gulgowski

Task [3] Regularities in expressions

KEY AND SCORE

SCORE

Exercise 1: 2 points for providing all matching strings for each pattern; 4 points in total
Exercise 2: 2 points for correctly identifying all matching and all non-matching forms for each pattern; 26 points in total
Total score: 30 points

ANSWERS TO EXERCISE 1

> *Scientific discipline that analyzes human language as a system for relating sounds (or signs in signed languages) and meaning. Phonetics studies acoustic and articulatory properties of the production and perception of speech sounds and non-speech sounds. The study of language meaning, on the other hand, deals with how languages encode relations between entities, properties, and other aspects of the world to convey, process, and assign meaning, as well as to manage and resolve ambiguity.*
> [...]
> *In the early 20th century, Ferdinand de Saussure distinguished between the notions of langue and parole in his formulation of structural linguistics. According to him, parole is the specific utterance of speech, whereas langue refers to an abstract phenomenon that theoretically defines the principles and system of rules that govern a language. This distinction resembles the one made by Noam Chomsky between competence and performance, where competence is individual's ideal knowledge of a language, while performance is the specific way in which it is used.*
> [...]
> *Although the term "linguist" in the sense of "a student of language" dates from 1641, the term "linguistics" is first attested in 1847. It is now the common academic term in English for the scientific study of language.*

RE 1: <stud(y|i)e?(nt|s)*>
matching strings: "studies", "study" (x2)

RE 2: <l(a|i)ngu(ist)?(ic|age|e)s*>
matching strings: "language" (x6), "languages" (x2), "langue" (x2), "linguistics" (x2)

ANSWERS TO EXERCISE 2

#	RE	VERB	MATCHING FORMS	NON-MATCHING FORMS
0	<work(ed|s)?>	WORK	work, worked, works	working
1	<see(n|s|ing)*>	SEE	see, sees, seen, seeing	saw
2	<ma(k|d)e(s|ing)>	MAKE	makes	make, made, making
3	<wr(i|o)tt?(e|n)>	WRITE	write, wrote	writes, written, writing
4	<f(i|ou)nds*>	FIND	find, finds, found	finding
5	<like(d|ing)>	LIKE	liked	like, likes, liking
6	<read(s|ing)+>	READ	reads, reading	read
7	<fl(y|e|o)?w(n|s)*>	FLY	flew, flown	fly, flies, flying
8	<g(i)?av(e|ing)+>	GIVE	gave	give, gives, given, giving
9	<kn(o|e)w(s|ing|n)>	KNOW	knows, knowing, known	know, knew
10	<match(s|ed|ing)*>	MATCH	match, matched, matching	matches
11	<th(ought|ink)+(s|ed)?>	THINK	think, thinks, thought	thinking
12	<move(ing|d|s)*>	MOVE	move, moves, moved	moving
13	<hid+(e|s|ing)?n?>	HIDE	hide, hid, hidden, hiding	hides