

(C) LOLWUT (1/3) [10 points]

Recognizing words – like that the sequence of letters “c-a-t” represents the word “cat” – is the foundation to any technology that works on text... but it’s not always so easy, especially on the internet where writers so often vary their spelling to express emphasis, emotion, surprise, etc. How can you recognize the word “what” when it could appear as “whaaaat” or “whaaa” or “wat” or “waaat” or “whut” or “wut” or...?

One way to recognize many variants at once is to use a *regular expression* (also called a “regex”) – a special sequence of symbols where you can indicate that a letter is optional, that it can occur many times, that one of several letters might occur, etc. For example, the regular expression

`wh?a+t*`

means that the “h” can occur exactly once or not at all, the “a” can occur one or more times, and the “t” can occur any number of times (including zero times). The symbols used in this regex are defined as follows (in this example, *unit* refers to single letters, but as you will see below, a unit can be larger):

- ? The previous unit can occur zero or one times
- * The previous unit can occur zero or more times
- + The previous unit can occur one or more times.

So the regular expression “`wh?a+t*`” would “recognize” the words below (and infinitely many more):

what wha whaaa wat waaaa watt waaaatttt waaaaat

It won’t, however, recognize the word “wht” (because there’s no “a” in this word and it requires at least one “a”), “whut” (there’s no “u” allowed here), “whhhat” (because it only allows zero or one “h” and this word has three), or “waaaah” (because the “h”, if it occurs, must precede the “a”; no reordering is allowed).

There are many more symbols that can appear in regular expressions, but for this problem you only need the above three symbols and parentheses. Parentheses group letters into units that themselves can be operated on by symbols, so that `L(OL)+` would recognize any of the words below (and infinitely many more):

LOL LOLOL LOLOLOL LOLOLOLOL

It won’t, however, recognize the word “L” (because it needs at least one instance of “OL”), or “LOOOOL” (because the entire “OL” part needs to be repeated, not just the “O”), or “LOOLL” (because the “OL” needs to be repeated as a unit).

Answer the questions on the next page in the Answer Sheets.



(C) LOLWUT (2/3)

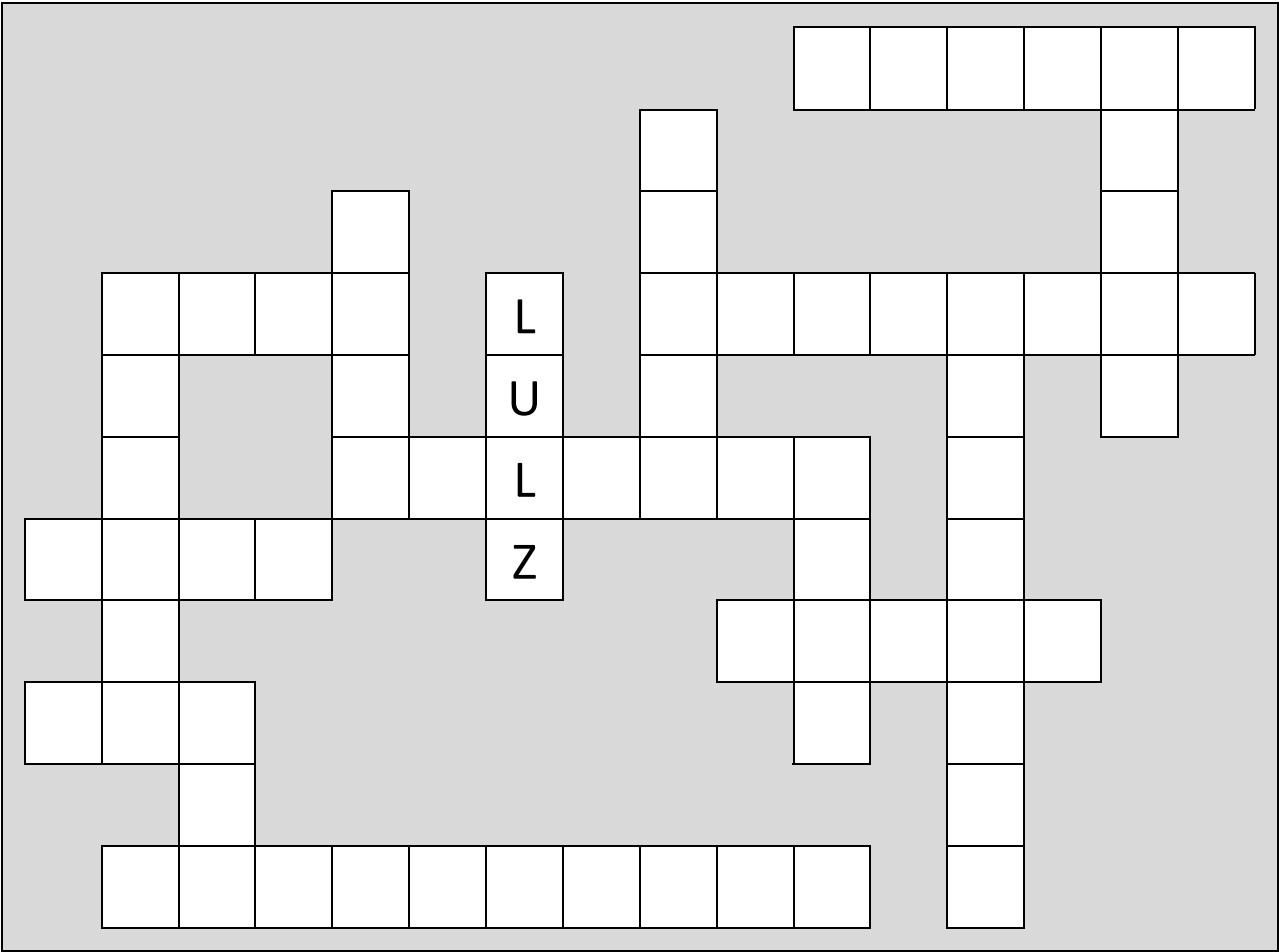
C1. We’ve put together a small crossword puzzle entirely of internet laughter, and clued each entry with a regular expression like “L (OL) +”.

The clues do not appear in any particular order, so you’ll have to work out for yourself where each entry goes, but each clue corresponds to only one entry. To help you get started, we’ve entered one answer into the grid already (there is no clue for this pre-entered answer). Be sure to write your answers in the Answer Sheets; nothing on this page will be graded!

Clues:

L (OL) +	(HO) +	K (EK) *E	(HAR+) +
H (EH) +	ROT?FL	TE (HE+) +	LAW*L
MWA (HA) +	HE (HE) +	LO+L	HAHA*
(AH) +A+	HA+	(JA) +	

Puzzle:



(C) LOLWUT (3/3)

C2. For each of the following regexes, write the shortest word that it could describe.

a. `LOL*LO?(OL)?O+L`

b. `(A?HA)+(LO)?O*`

a.

b.

