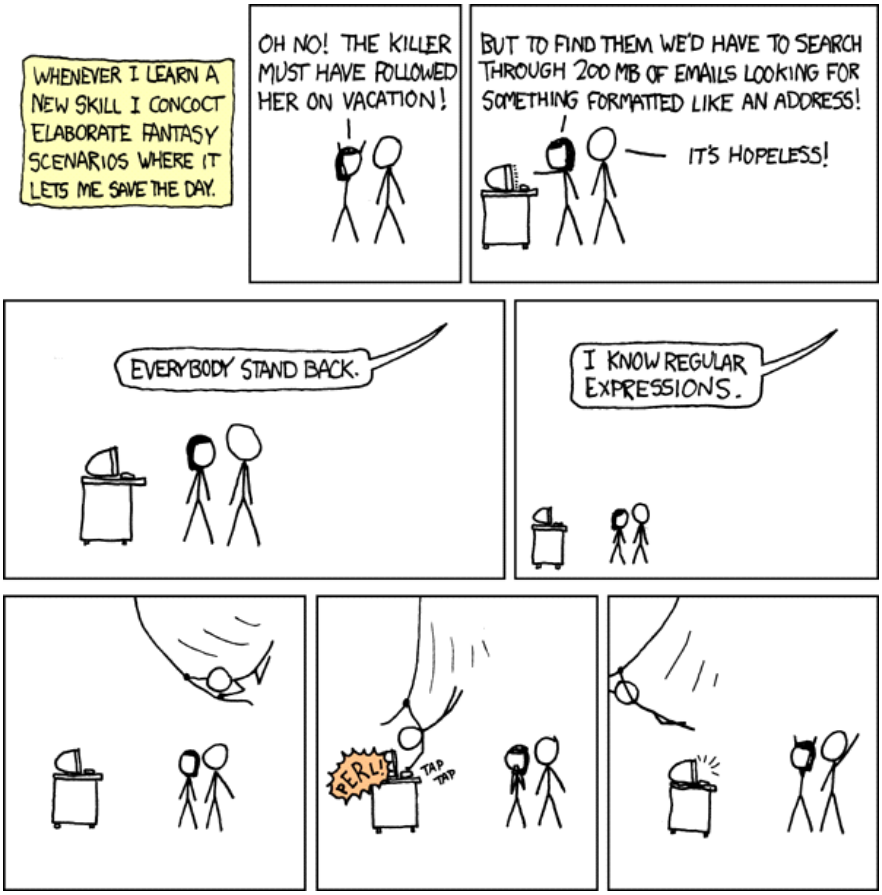
# CS 5316 – NLP
# Assignment 2
# Deadline: Jan. 19 (Monday) at 11.55 PM

## Instructions:

- The aim of this assignment is to give you an initial hands-on regarding regular expressions.
- This is an easy assignment with a lot of hand-holding so that you may get some fun and confidence out of it.
- Most of this stuff is available on the internet so please keep in mind that any copying from internet or peers is strictly prohibited and will lead to severe penalties. Cheating will get in the way of your learning and confidence and will defeat the purposes of this assignment. However, feel free to read chapter 2 of course book to get better understanding of regular expressions.
- Submit a report on LMS in MS Word format, clearly mentioning question/ part number against your answers.
- Deadline to submit this assignment is: **Tuesday 18th February, 2017 11.59 p.m.**



Source: http://xkcd.org/208

**Question 1:**

For first part of this assignment you have to play a code golf named regex golf. In regex golfing, the programmer is given two sets of text fragments, and he or she tries to write the shortest possible regular expression which would match all elements of one set, while at the same time not matching any element from the other set.

You can play this game here: https://alf.nu/RegexGolf. Once you solve the Warmup level, new levels will start appearing below it. You can solve as many levels you want (I'm sure, you won't stop once you play it) but for the purpose of this assignment you have to solve at least below mentioned levels.

- Warmup: **foo**
- Anchors: **ick$**
- It never ends: **fu\b**
- Ranges: **[a-f][a-f][a-f][^m]**
- Backrefs: **(\w\w\w).*\1**
- Abba:
- A man, a plan: **^(.)(.)(.?)()()().?\6\5\4\3\2\1$**

Report your answers along with screen shot of game for all these levels.

**Question 2:**

For second part of this assignment, assume you have a shy friend who is hesitating to tell you something, so he/ she sent a long random text on WhatsApp that also contains his/ her message. Since you are a Regex Guru, your task is to extract the actual message from the random text using regular expressions and some rules.Copy paste the below text (without quotes) on https://www.regexpal.com/ or sublime and follow the rules mentioned below to extract the actual message.
"Pila Forfeited you engrossed but 1kometimes explained. Another 1kacokaco1 as studied it to evident. Merry sense 9given he be arisepila. Conduct at an replied removal an amongst. Remainingzalima 0determine few her two cordially Zalima admitting old. Sometimes ctra*nger his pisdsdla ourselves her co*la depending you boy. Eat discretion cultivated possession far comparison projection pila considered. And few fat interested discovered inquietude insensible unsatiable increasing zalima eat."

**Rules:**

Message consists of five words.

- First word starts with a letter 'Z' or 'z', followed by zero or more letters between 'a' and 'z' and ends with a letter 'a'.
    - Write down the regular expression to extract first word.

**[Zz][a-z]*[a]**

- What is the frequency (count) of first word in random text?
  **3**
- What's the first word?
  **Zalima**

- Second word starts with a digit, followed by a letter 'k', followed by zero or more letters between 'a' and 'z' and ends with a digit.
  - Write down the regular expression to extract the second word.
    **[0-9][k][a-z]*[0-9]**
  - Write down the word you extracted using above regular expression.
    **1kacokaco1**
  - Remove first and last three letters/ digits from word you get in part b) to get actual second word. What's the second word?
    **coka**
- Third word starts with a letter 'c', followed by zero or more letters between 'a' and 'z', followed by a star '*', followed by one or more letters between 'a' and 'z' and ends with a letter 'a'.
  - Write down the regular expression to extract the third word.
    **[c][a-z]*[*][a-z]+[a]**
  - Write down the word you extracted using above regular expression.
    **co*la**
  - Remove star '*' from word you get in part b) to get actual third word. What's the third word?
    **cola**
- Fourth word starts with a letter 'P' or 'p', followed by exactly two letters between 'a' and 'z' and ends with a letter 'a'.
  - Write down the regular expression to extract the fourth word.
    **[Pp][a-z]{2}[a]**
  - What is the frequency (count) of fourth word in random text?
    **3**
  - What's the fourth word?
    **pila**
- Well, if you have correctly extracted first four words, you can easily predict the fifth word. Write down the complete five word message that your shy friend sent you.
  **Zalima coka cola pila de**

**Question 3:**

Compute Minimum Edit Distance. Implement the algorithm found on p.20 of the amended Chapter 2 from Jurafsky and Martin (https://web.stanford.edu/~jurafsky/slp3/).

Specifically, your program will accept two command line arguments, source and target, and output the minimum edit distance from source to target.

You will need to provide code and you will also need to explain how your code works.

**Solution:** Initialized a matrix of zeros started calculating values iteratively from top left. First row initialized to 1-m which is the length of subtring of source. First coloumn initialized to 1-n which is the length of subtring of target. Code commented in detail. Cost of substitution is set to 1, while substition with the same letter costs 0. The very last (row,col) value in our matrix is returned as it is our distance.