**CS 5316–Natural Language Processing**

**Assignment 2**

Deadline: Feb. 15 (Friday) at 11:55 PM

# Question 1: (25 Marks)

1. You are required to implement a regular expression engine that checks for balanced parenthesis in a given text/string. If the parenthesis are balanced and properly ordered too, it should print the valid message accordingly. Note that you only have to use regular expressions. Also note that you may have to use a few loops too. You can use JAVA or PYTHON only. Following are a few examples:
   1. 12((1245))6 ---- ACCEPTED
   2. 12{65}()()( ---- REJECTED
   3. 12({)ppd} ---- REJECTED (note the out of order brackets)
2. Implement an email verification regex (only one regex) that can verify the following emails:
   1. jurafsky@stanford.edu
   2. jurafsky(at)cs.stanford.edu
   3. jurafsky at csli dot stanford dot edu
3. Implement a Mobile number verification Regex (only one regex) that can verify the following:
   1. +1-650-723-0293
   2. (650) 723-0293
   3. (+1): 650-723-0293

# Question 2: (25 Marks)

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](https://www.sublimetext.com/) 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.
* What is the frequency (count) of first word in random text?
* What’s the first word?
* 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.
* Write down the word you extracted using above regular expression.
* Remove first and last three letters/ digits from word you get in part b) to get actual second word. What’s the second word?
* 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.
* Write down the word you extracted using above regular expression.
* Remove star ‘\*’ from word you get in part b) to get actual third word. What’s the third word?
* 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.
* What is the frequency (count) of fourth word in random text?
* What’s the fourth word?

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.

# Question 3: Sentiment Analysis (25 Marks)

Download dataset from <https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences>this link. It has sentences from three domains: i) imdb ii) amazon iii) yelp

Apply preprocessing e.g., removing stop words, lemmatization, stemming. State overall vocabulary size before and after preprocessing.

Create Frequency list of all words for each domain. Plot the top ten positive and negative words frequencies for each domain. Comment of the comparison.

# Question 4: (25 Marks)

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. The cost for intersection, deletion and substitution should be taken from the user.

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