# **HW Assignment 0 (A0): Regular Expressions** CS6120: Natural Language Processing Spring 2024

Due Date: Tuesday, January 15, 2024

**(5 points)**

# Regular Expressions - Grouping and string methods - DEV Community

# Scenario

A regular expression (RE) is a sequence of characters that forms a search pattern. RE can be used for string searching and manipulation tasks, such as finding, replacing, or validating text. Regular expressions are powerful tool in many languages for handling text data. They are useful in data cleaning, parsing, and text preprocessing.

# Task

This assignment has two parts to it:

**Part A):** You are given a small csv file with five short stories listed in rows. The file also contains empty columns with header labels. Use RE to extract information for the empty columns.

# **Part B)** Download all 5 volumes of "A system of practical medicine" form Gutenberg Library. Then apply RE search to look for the number of times most common modern health conditions are mentioned in each text. Your objective is to create a df with five rows in it, one for each volume. The df should contain columns for various health conditions and their frequency within each volume. Here are the most frequent health conditions:

1. Heart disease
2. Cancer
3. Stroke
4. Respiratory diseases
5. Alzheimer's disease
6. Diabetes
7. Influenza and Pneumonia
8. Kidney diseases
9. Septicemia
10. Liver disease
11. Hypertension
12. Parkinson's disease
13. Chronic lower respiratory disease
14. Accidents/injuries
15. Osteoporosis
16. Asthma
17. Depression
18. Oral health issues
19. HIV/AIDS
20. Tuberculosis
21. Malaria
22. Dengue fever
23. Hepatitis
24. Epilepsy
25. Multiple sclerosis

# 

# **Expected Output**

Please submit a fully executed Jupyter notebook clearly identifying question number and steps. Make sure to add proper commentary to your solution.