**Workshop Title: Applied Data Science for Public Health Using R**

**Series 3: Natural Language Procesing (NLP)**

**Course Goal:** To equip participants with foundational skills in R to manipulate text data, acquire data from the web, perform natural language processing, and create thematic maps for public health applications.

---

**Day 1: Foundations of R and Analysing Unstructured Text**

This day focuses on building a solid foundation in R and learning to handle and process text data, a common task in public health for analysing reports, surveys, and patient notes.

| Time | Topic | Key Concepts & R Packages | Hands-On Exercise |
|---|---|---|---|
| **9:00 - 10:30** | **R Fundamentals** | Introduction to R & RStudio, projects, scripts, installing packages, basic data structures (vectors, data frames), reading data. | Install relevant packages, import a sample public health dataset. |
| **10:30 - 10:45** | **Break** | | |
| **10:45 - 12:30** | **Data Wrangling Essentials** | Introduction to the "Tidyverse", filtering, selecting, and arranging data, creating new variables. | Clean and prepare the imported patient dataset for analysis. |
| **12:30 - 1:30** | **Lunch Break** | | |
| **1:30 - 3:00** | **String Manipulation** | Working with text, finding and replacing patterns with regular expressions (regex). | Standardize and clean messy text fields, such as addresses or free-text survey responses. |
| **3:00 - 3:15** | **Break** | | |
| **3:15 - 5:00** | **Introduction to Tidy Text Mining** | Principles of tidy text, tokenization, removing stop words, and frequency analysis. | Analyze open-ended survey questions to find the most common words and themes. |

**Day 2: Acquiring Web Data & Natural Language Processing (NLP)**

Day two focuses on gathering data directly from the web and applying NLP techniques to uncover insights from text, such as public sentiment or key topics in health reports.

| Time | Topic | Key Concepts & R Packages | Hands-On Exercise |
|---|---|---|---|
| **9:00 - 10:30** | **Introduction to Web Scraping** | Understanding HTML basics, using CSS selectors to target web page elements. | Inspect a public health organization's website (e.g., WHO, CDC) and identify the HTML structure of a news feed. |
| **10:30 - 10:45** | **Break** | | |
| **10:45 - 12:30** | **Practical Web Scraping** | Scraping text data and tables from static web pages, handling simple pagination. | Scrape a table of disease statistics or a list of public health news headlines from a website. |
| **12:30 - 1:30** | **Lunch Break** | | |
| **1:30 - 3:00** | **Sentiment Analysis** | Understanding sentiment lexicons, calculating sentiment scores for text documents. | Analyze the sentiment of the scraped news headlines to gauge if they are positive, negative, or neutral. |
| **3:00 - 3:15** | **Break** | | |
| **3:15 - 5:00** | **N-gram Analysis & Visualization** | Counting and filtering common word pairs (bigrams) and triplets (trigrams). Visualizing frequencies with bar charts and word clouds | Analyze patient reviews to find the most common phrases (bigrams) describing their experience (e.g., "wait times", "front desk", "helpful staff"). Create a word cloud of the most frequent single terms. |