DS3030: Data Analytics Lab
Assignment 7

Date: Sep 15, 2025

Timing: 2:00 to 4:45 PM Max marks: 18

Instructions

- Submit one .ipynb file containing all answers named as [**student name**]_**assignment**[**number**]**.ipynb**
- Write the **Part number** in **separate text blocks** before the answers.
- Write **justifications/comments** as required.

1 Part I - Text Preprocessing & Feature Extraction

Total Marks: 6

Dataset: Large Movie Review Dataset (Hugging Face)

Dataset Summary: Large Movie Review Dataset. This is a dataset for binary sentiment classification containing substantially more data than older benchmark datasets. It contains a set of 25,000 highly popular movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well.

Use the following code to download it into pandas dataframe:

```
import pandas as pd
splits = {'train': 'plain_text/train-00000-of-00001.parquet',
          'test': 'plain_text/test-00000-of-00001.parquet',
          'unsupervised': 'plain_text/unsupervised-00000-of-00001.parquet'}
df = pd.read_parquet("hf://datasets/stanfordnlp/imdb/" + splits["train"])
```

1. (3 points) **Complete Text Preprocessing Pipeline**

   Using the movie reviews dataset above, perform a comprehensive text preprocessing pipeline:

   Part 0: Load the dataset and sample the first 1000 reviews for easy processing.

   Part A: Basic Preprocessing (1.5 Marks)

   1. **Lowercase conversion** - Convert all text to lowercase
   2. **Contraction expansion** - Expand contractions (I'm → I am, It's → It is, etc.)

3. **Punctuation & special character removal** - Remove all punctuation marks and special characters
4. **Tokenization** - Split text into individual words/tokens
5. **Stopword removal** - Remove English stopwords using NLTK

Part B: Advanced Text Normalization (1.5 Marks)

1. **Stemming** - Apply Porter Stemmer to all tokens and show results
2. **Lemmatization** - Apply WordNet Lemmatizer and compare with stemming results
3. **POS Tagging** - Perform Part-of-Speech tagging on the lemmatized tokens for the first two reviews

2. (3 points) **Feature Extraction & Analysis**
   Using the preprocessed IMDB movie reviews from Question 1:

   Part A: Bag of Words Implementation (1.5 Marks)

   1. Create a complete vocabulary from all preprocessed reviews
   2. Generate the Document-Term Matrix showing frequency counts
   3. Identify the top 5 most frequent words across all reviews

   Part B: N-grams & Feature Analysis (1.5 Marks)

   1. **Bigram Generation** - Extract all bigrams from the preprocessed movie reviews corpus
   2. **Trigram Generation** - Extract all trigrams from the same corpus
   3. **N-gram Frequency Analysis** - Create frequency counts for both bigrams and trigrams, showing the top 5 most frequent bigrams and top 3 most frequent trigrams

2 Part II - Youth Tobacco Awareness/Usage Survey Analysis

Total Marks: 6
Dataset: "GYTS.csv"
Dataset Name: "Global Youth Tobacco Survey (GYTS-4)"
Dataset Description:

- Above survey was conducted in 2019 by the International Institute for Population Sciences (IIPS) under the Ministry of Health and Family Welfare (MoHFW)
- A total of 80,772 students aged 13-15 years who participated in the survey were considered for reporting.
- The objective of the survey was to provide information on tobacco use, cessation, second-hand smoke, access and availability, exposure to anti-tobacco information, awareness, and receptivity to tobacco marketing, knowledge, and attitudes.
- The key features include those for "*State/UT*", "*Area*", "*Usage*" statistics, "*Age*" statistics, "*Exposure*" statistics, "*Awareness*" statistics, "*Source*" statistics and "*School Action*" statistics

More details can be found here.

1. (2 points) **TASK 1: Tobacco Exposure Awareness Chart**
   Prepare a visualization for the various sources of exposure to tobacco smoke in your state using a waffle chart. The output should be similar to Figure 1 (results shown for India here, replace it with your state of origin) :

   Workflow :

   1. Filter the row with columns named (0.5 mark)
      - "*Area*" with text value "*Total*" and
      - "*State/UT*" as the state of your origin, say "*Goa*".
   2. Select the values in the columns with name starting with the text "*Exposure*" from the above row, e.g., *"Exposure To Tobacco Smoke At Home/public Place"*.(0.5 mark)
   3. Use these values to prepare the waffle chart (1 mark)

2. (4 points) **TASK 2: Winsorization Technique For Data Cleaning**

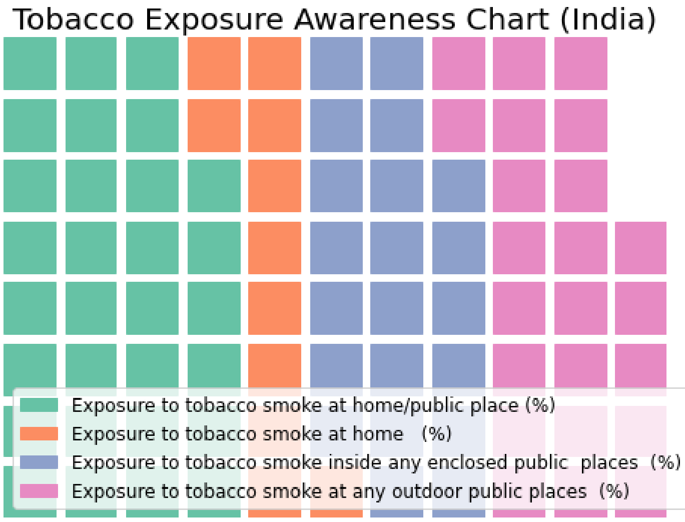

Figure 1: Tobacco Exposure Awareness Waffle Chart (India)

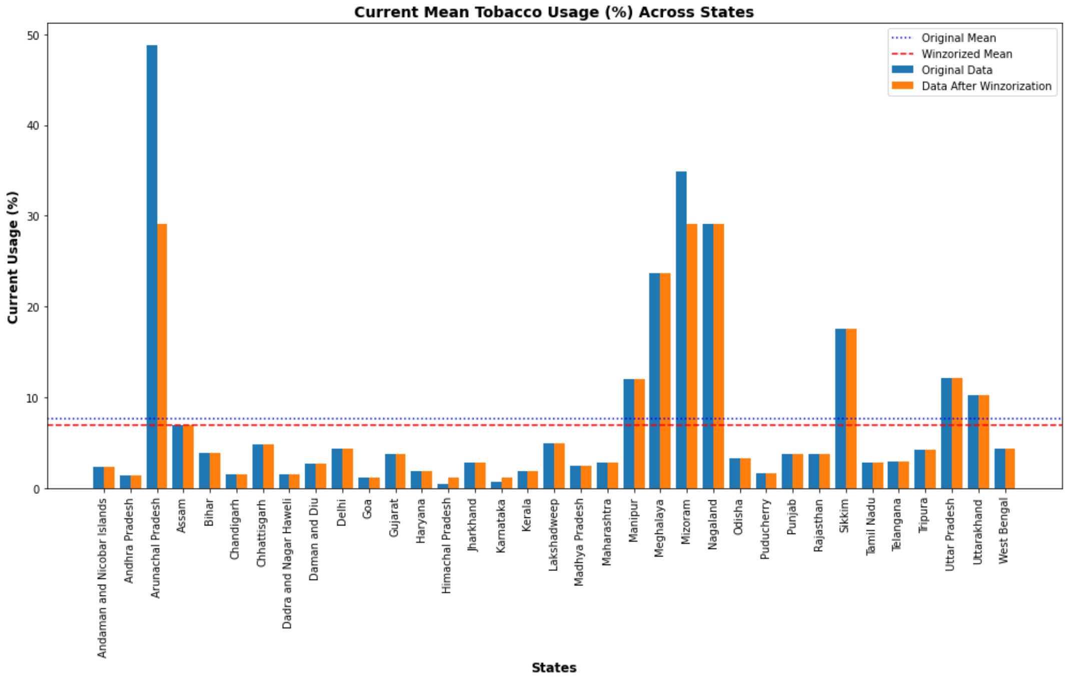

Figure 2: Current Mean Tobacco Usage (%) in each states using Winzorization And Grouped Bar Chart

Prepare a grouped bar chart that displays the original and winsorized data on current tobacco usage in each state. Also, plot the horizontal lines representing the means of the original and winsorized data. The output should be similar to Figure 2.

Workflow :

1. Filter the rows with column "*Area*" as "*Total*" and "*State/UT*", all except "*India*" (0.5 mark)
2. Store as a single array, say y1, the mean of the following column values for each row in the above dataset (0.5 marks)
   - Current tobacco users (%)
   - Current tobacco smokers (%)
   - Current cigarette users (%)
   - Current bidi users (%)
   - Current smokeless tobacco users (%)
3. Apply winsorization on y1 to obtain say, y2 (0.5 mark)
   (**NOTE:** Winsorization limits can be chosen based on your intuition after observing the data values in y1)
4. Compute the arithmetic means of arrays y1 and y2 (0.5 mark)
5. Plot a group chart using y1, y2, and their respective means as horizontal axis lines (1 mark)
   (**NOTE:** Each state has a group of its 2 corresponding values from y1 and y2, respectively)
6. X-axis labels must be the respective state names (0.5 mark)
7. Titles, Labels, and Legends must be visually appealing (0.5 mark)

3 Part III: Spam or ham messages

1. (3 points) Data Cleaning
   You are given a CSV file named `spam.csv`. It contains SMS messages and labels (spam or ham). Perform the following tasks:

   1. Load the CSV file using pandas.
   2. Keep only two columns:
      - **label**
      - **message**
   3. Remove all missing values.(0.25 makes)
   4. Convert all text in the message column to lowercase.(0.25 marks)
   5. Remove punctuation, digits, and English stopwords from the text.(0.5 marks)
   6. Create a new column named **cleaned** that stores the cleaned version of each message.(1 marks)
   7. Save the cleaned dataset as `spam_cleaned.csv`.(1 marks)

2. (3 points) Word Cloud Visualization
   Using the cleaned dataset `spam_cleaned.csv` created in Q1:

   1. Read the CSV file into a dataframe.(0.5 marks)
   2. Select the **top 100 spam** and **top 100 ham** messages (based on the number of words/messages that are the longest).(0.5 marks)
   3. Generate two separate word clouds(1 marks) :
      - One for spam messages(using top 100 by length)
      - One for ham messages(using top 100 by length)
   4. Display both word clouds.(1 marks)