Skills for Hire Atlantic

Data Analytics – Assignment 2

Overview

This assignment is in two parts:

1. Part 1 (20 marks): Analysis of life expectancy
2. Part 2 (20 marks): NLP

You will be graded on the accuracy of your results and on the quality of your plots. Please make sure to clearly indicate the question parts before answering them and to comment on your code. The marks for each problem are indicated between brackets.
You are **required** to write the code in Python. You should write your code in a **notebook** (Google Colab or Jupyter). You are free to use any library you like.

Recommended libraries: pandas, numpy, matplotlib, seaborn, etc.
Note: This assignment will count for **40%** of your Assignments grade.

Due date: **April 14th 2024**

Submission Instructions

Submit through:
https://formesign.com/sm/d7McnHtBg

Submission Requirements

File format: **A compressed ZIP (.zip) file** containing your Python Notebook (.ipynb) file
File name: DA_Group_Assignment2_FirstName_LastName_EmailAddress.ZIP

MAKE SURE TO REPLACE Group with your Group color, FirstName with your first name, LastName with your last name and EmailAddress with your email address.

Hints

Start early. There are many parts to this assignment and it would be very difficult if left to the last minute.
Don't reinvent the wheel. Feel free to use the examples covered in class.
If you get stuck, reach out to your TA for help!
Do not spend hours without asking for help. Good luck!

Question 1 (20)

Dataset: Life_Expectancy_Data.csv

Preparation (5)

1. Load the dataset. (1)
2. Display the first 20 rows. (1)
3. Find the number of null values in the dataset. (1)
4. Impute the missing values with the mean values of the data. You can use SimpleImputer from sklearn.impute. (1)
   Bonus: Instead of imputing the missing values with the mean value of the whole column, impute it with the mean value of the column that corresponds to the country. (1 bonus point)
5. Find the count, mean, standard deviation, quartiles and extrema for the numeric columns. (1)

Visualization (15)

6. Find the correlation between the numeric columns and display your findings on a heatmap. (2)
7. Plot a histogram of the life expectancy. (2)
8. Compare the life expectancy in developed countries to that in developing countries using violin plots next to each other. (2)
9. On the same line plot, display the life expectancy from 2000 to 2015 for Canada, the United Kingdom and the United States of America. (2)
10. Compare the average infant deaths over the years against the average life expectancy over the years using a scatter plot for the following countries: Belgium, Brazil, Cameroon, Canada, China, France, Ghana, India, the United Kingdom and the United States of America. What can you conclude? (2)
11. In the year 2012
    a. Compare the life expectancy with schooling using a scatter plot. (2)
    b. What is the Pearson correlation? (1)
    c. Draw the best regression line on the same plot as (a). (1)
    d. What can you conclude? (1)

Question 2 (20)

Dataset: text.csv

Preprocessing (2)

1. Load the dataset. (1)
2. Display the first 5 rows. (1)

Cleaning (8) - You might want to create a new column that will store the cleaned text. After each part (3-6), display the first 5 rows to check if you cleaned the data correctly.

3. Convert the text in the *review* column to lowercase.(1)
4. Remove stopwords. (2)
5. Remove punctuation signs. (2)
6. Apply lemmatization to every word in the cleaned column. (2)
7. Remove rows that contain missing values. (1)

TF-IDF (10)

8. Create 2 new dataframes – one that contains reviews with positive sentiment (where the label = "*pos"*) and one with negative sentiment (where the label = "*neg"*). Note that we will be working with the cleaned text. (4)
9. Calculate the TF-IDF for positive cleaned reviews and for negative cleaned reviews. (4)
10. What are the 10 most important words in each dataset? (2)