UDA F2021 Assignment 3 (Group work)
Due: 15th November 2021 by 11:59 p.m.

Is a Picture Worth a Thousand Words?

In this assignment, you are an analytics consultant to the National Geographic (Zara). Your objective is to help Zara increase engagement on its Instagram page.

Task A: Scrape Instagram.py to fetch ~200 posts from the Zara Instagram page. This script fetches (i) image URLs, (ii) post caption (the text description of a post), and (iii) # likes.

Task B: Using the image URLs, obtain **image labels** (text) from Google Vision (cloud service)

Task C: Create a column called **binary** (lowercase only) where value =1 (stands for high engagement) or 0 (stands for low engagement) based on whether the number of likes is above or below the median value.

Task D: Run a logistic regression with **binary** as the dependent variable, and the image_labels as independent variables. Before running the regression, replace the column label image_labels with **text**, since the script expects **text** to be the name for the column containing text. What is the accuracy (show the confusion matrix) of this prediction model? The idea is to be able to predict the engagement level for an image.

Accuracy = 1 - # prediction errors / total # cases

What accuracy do you get by using the post_caption words as the independent variables instead of image_labels? As in the first regression, change **post_caption** to **text** before running the logistic regression. Finally, what accuracy do you get by combining (concatenating) the image_labels and post_caption and using them together as independent variables? What can you conclude from your analysis?

Task E: Perform topic modeling (LDA) on the original image_labels. Choose an appropriate number of topics. You may want to start with 4-5 topics, but adjust the number up or down depending on the word distributions you get. **Decide on suitable names for each topic.**

Now sort the data from high to low number of likes (don't use the binary column, use the actual number of likes), and consider the highest and the lowest quartiles of likes. What are the main differences in the **average** topic weights of images across the two quartiles (e.g., greater weight of some topics in the highest versus lowest quartiles)? Show the main results in a table.

Task F: What advice would you give Zara if it wants to increase engagement on its Instagram page based on your findings?

Deliverables: Python notebook with code and answers. Write the names of all team members **inside** the notebook.