MSc Practical Data Science - Assignment 3 - Part 1: Image Annotation [5-8/12]

The assignment consists of three parts, which will be announced gradually. This is the first part that focuses on data **annotation**.

Locate your assigned ID and access the dataset provided in Google Drive. Each file in the dataset contains four columns, accompanied by an explanation of what each category represents:

1. Image URL
2. Image
3. Categories
4. Other

Your task is to review each image and select one or more categories that best describe factors hindering or obstructing pedestrians. Base your selections on personal judgment rather than legal criteria, as the images may include both lawful and unlawful (e.g.) parking situations.

Please follow these guidelines:

- Do not modify the file format or structure provided.
- Make selections only from the categories listed, using your personal perspective.
- Write your observations in an online text provided via e-class.

MSc Practical Data Science - Assignment 3 - Part 2: Data Mining [7-12/12]

Introduction

This assignment is designed to assess your understanding and practical skills in data science techniques covered in all the modules of the course. You will be working with a real-world dataset to perform data exploration, preprocessing, feature engineering, and model building. All your work must be done within a notebook.

Dataset

The dataset for this assignment is the "GSV Cities" dataset available on Kaggle. It consists of street-level images from various cities worldwide, along with metadata such as city name and location. This dataset can be loaded directly into code using the Kaggle API.

Tasks

1. **Data Acquisition and Preprocessing:**
    - Download the "GSV Cities" dataset from Kaggle within your notebook.
    - Explore the dataset to understand its structure, features, and potential challenges.
    - Choose randomly images per city to create a dataset that will fit in the memory.
    - Perform necessary data cleaning and preprocessing steps to yield a dataframe that will comprise at least a column of cities and one of images. Document your rationale for each preprocessing step.
2. **Exploratory Data Analysis:**
    - Conduct exploratory data analysis (EDA) to gain insights into the data.
    - Generate descriptive statistics and visualizations to summarize the data's characteristics. Explore relationships between variables and identify patterns or trends.

    - Briefly describe your findings from the EDA.
3. **Unsupervised Learning:**
    - Apply KMeans clustering to group similar images. You can select the K value empirically or based on a hypothesis.
    - Evaluate the clustering results and interpret the clusters by providing a single textual description per cluster.
    - Answer which cities are more similar in terms of cluster assignments (their images are similarly distributed across the clusters).
4. **Supervised Learning (Optional):**
    - Improve your NLP classifier of A2 by creating an ensemble (e.g., combining the predictions of classifiers operating on titles and ones on long documents), by using an LLM to augment imbalanced classes, or by using the long documents to correct corrupted titles.
5. **Conclusion:**
    - Summarize your findings from all tasks.
    - Discuss potential limitations of your analysis and suggest areas for further improvement.

MSc Practical Data Science - Assignment 3 - Part 3: End Application [9-12/12]

Use the images annotated by all students in Part 1 (in Google Drive) to compute the inter-annotator agreement and analyse the task difficulty. Use your trained clustering model from Part 2 to classify each image from Part 1 to a cluster. Based on this clustering assignment, find the three most similar cities to Athens from Part 2 (as in Part 2, step 3, 3rd bullet). Compare the (human) annotations to the (labelled cluster) assignments and report any interesting findings.

Submission

You must submit a single notebook containing all your code, analysis, and visualizations. Ensure that the notebook is well-organized and includes clear explanations of your methods and results.

Grading Rubric

The assignment will be evaluated based on the following criteria:

- Data acquisition and preprocessing (20%)
- Exploratory data analysis (20%)
- Unsupervised learning (10%)
- Supervised learning (10%) - optional
- Overall presentation and clarity (30%) - includes oral presentation on 23/12

Two more tasks will be asked during this assignment (to be announced separately):

- Annotation and task difficulty estimation (10%)
- End application (10%)

Important Notes

- Document your code with comments to explain your steps and logic.
- Use appropriate visualizations to communicate your findings effectively.
- Ensure that your notebook runs without errors and produces reproducible results.

Late submissions will be subject to penalty.