Assignment: Text Analysis Using Bag of Words and Distance Metrics

Objective:

The goal of this assignment is to analyze a set of documents from different domains—politics, science, and sports—using the Bag of Words (BoW) model. Students will then compute various distance metrics to assess the similarity between documents.

Assignment Instructions:

1. Data Preparation:

 - You are provided with a dataset consisting of 9 documents: 3 from the domain of politics, 3 from science, and 3 from sports.

 - Each document is in plain text format. Ensure all documents are properly loaded and accessible for processing.

2. Bag of Words Model:

 - Implement the Bag of Words (BoW) model to convert the text documents into a numerical format.

 - Create a term-document matrix where each row represents a document and each column represents a term from the vocabulary. The values in the matrix should be the term frequencies of the respective terms in each document.

3. Distance Metrics:

 - Calculate the following distance metrics between all pairs of documents:

 1. Jaccard Distance:

 - Compute the Jaccard distance between pairs of documents. The Jaccard distance is defined as \(1 - Jaccard\ similarity\), where Jaccard similarity is the size of the intersection divided by the size of the union of the term sets of the documents.

 2. Euclidean Distance:

 - Calculate the Euclidean distance between pairs of documents using their BoW vectors.

 3. Cosine Similarity (and Distance):

 - Compute the Cosine similarity between document pairs. Convert this similarity measure to a distance metric by subtracting the cosine similarity from 1.

 4. Kullback-Leibler (K-L) Divergence:

 - Calculate the K-L divergence for pairs of documents. Since K-L divergence requires probability distributions, normalize the term frequencies to obtain these distributions before computing the divergence.

4. Analysis:

 - Create a distance matrix for each of the four metrics mentioned above.

- Visualize the distance matrices using heatmaps to help interpret the results.

- Discuss any patterns or insights you can derive from these matrices. For example, do documents within the same domain tend to be more similar to each other compared to documents from different domains?

5. Report:

- Submit a report that includes the following:

  1. Methodology:

     - Describe the Bag of Words model implementation and the approach used for calculating each distance metric.

  2. Distance Matrices:

     - Include the distance matrices for Jaccard, Euclidean, Cosine, and K-L metrics.

  3. Visualizations:

     - Provide heatmaps or other visual representations of the distance matrices.

  4. Analysis and Interpretation:

     - Discuss the similarities and differences observed in the distance matrices. Highlight any significant findings related to the document domains.

  5. Code and Documentation:

     - Provide your code along with comments explaining each part of the process. Ensure that your code is well-organized and easily understandable.

Submission Guidelines:

- Submit your report as a PDF document.

- Ensure that all files are properly named and organized.

Evaluation Criteria:

- Accuracy and correctness of the Bag of Words model implementation.

- Correctness of distance calculations and adherence to the distance metric definitions.

- Clarity and quality of visualizations.

- Depth of analysis and interpretation of results.

- Organization and readability of the report and code documentation.

Assignment Question 2:

You are provided with six grayscale images labeled as follows:

- **OR**: The original image.
- **GT**: The ground truth image.
- **Algo1** to **Algo5**: Five images generated by different algorithms.

Tasks:

1. **Calculate the Difference Images:**
    - For each image I(where I represents GT, Algo1, Algo2, Algo3, Algo4, and Algo5), compute the absolute difference image D(I,OR) between the original image (OR) and the image I. That is, calculate D(I,OR) = ∣I−OR∣ for each of the six images.
2. **Jaccard Distance Calculation:**
    - Consider the difference image D(GT,OR) obtained from the ground truth (GT) and the original image (OR) as the reference difference image.
    - For each of the difference images D(Algo1,OR), D(Algo2,OR), D(Algo3,OR), D(Algo4,OR), and (Algo5,OR):
        - Calculate the Jaccard Distance between the reference difference image D(GT,OR) and each difference image D(Algoi,OR) (where i=1i = 1i=1 to 555).
3. **Interpret the Results:**
    - Discuss what the Jaccard Distances indicate about the similarity or dissimilarity between the algorithms' outputs and the ground truth.

Deadline: 19/09/2024