**Question 1: Text Analysis using Multinomial Distribution and MLE (45 minutes)**

**Background** You are working with a dataset of four paragraphs from different academic fields: physics, economics, biology, and psychology. Your task is to apply basic statistical techniques including Maximum Likelihood Estimation (MLE) to analyze and classify these texts.

**Text Paragraphs** **Paragraph 1 (Physics):** "Quantum mechanics describes the behavior of matter at the atomic scale. The uncertainty principle and wave-particle duality challenge our understanding of reality. Recent experiments with quantum computers have demonstrated potential advantages over classical computing methods. Researchers continue to debate interpretations of quantum theory and its implications for our understanding of the physical world."

**Paragraph 2 (Economics):** "Market equilibrium occurs when supply and demand forces balance, establishing optimal prices. Macroeconomic policies influence inflation, unemployment, and economic growth across entire economies. Behavioral economics has revealed how psychological factors affect financial decision-making. As markets become increasingly global, economists develop models to account for international trade relationships."

**Paragraph 3 (Biology):** "Evolution has shaped the diversity of species through natural selection over millions of years. Cellular processes convert nutrients into energy through complex biochemical pathways. Genetic information flows from DNA to RNA to proteins, with regulatory mechanisms controlling gene expression. Conservation efforts focus on maintaining biodiversity in ecosystems to ensure their long-term stability."

**Paragraph 4 (Psychology):** "Cognitive development progresses through stages as children build more complex mental models. Social psychology studies how individuals are influenced by others in various contexts. The brain demonstrates plasticity, forming new neural connections throughout life during learning experiences. Mental health treatments integrate various therapeutic approaches with consideration of individual differences."

**Instructions**

1. Perform basic text preprocessing:
   * Convert all text to lowercase
   * Remove punctuation and common stop words
   * Create a vocabulary of unique words across all paragraphs
2. Calculate word frequencies:
   * Count the occurrences of each vocabulary word in each paragraph
   * Calculate the relative frequency distribution for each paragraph
3. Apply Maximum Likelihood Estimation:
   * Use the word frequencies to estimate parameters of a probability distribution for each academic field
   * Calculate log-likelihood values for each paragraph based on your model
4. Analyze the following unclassified paragraph: "Neural networks model relationships between variables through layers of connected nodes. Each connection has a weight that is adjusted during training to reduce errors. Learning algorithms optimize these weights based on patterns in the data."
5. For the unclassified paragraph:
   * Calculate the likelihood it belongs to each academic field using your models
   * Determine the most probable classification
   * Identify the key words that influenced your classification result
6. Briefly discuss one limitation of this statistical approach to text classification

**Question 2: Multivariate Normal Distribution Analysis in Medical Data (45 minutes)**

**Background** You are working with a dataset of 1000 patients, where each patient record contains measurements for five medical parameters: temperature (°F), blood pressure (systolic, mmHg), glucose (mg/dL), hemoglobin (Hb, g/dL), and uric acid (mg/dL). The patients are diagnosed as either "Healthy" or "Condition Present."

**Task** Your goal is to analyze this medical data using multivariate normal distribution to understand differences between patient groups.

**Data Description**

* patient\_data.csv contains records for 1000 patients
* Each record includes: temperature, blood\_pressure, glucose, hemoglobin, uric\_acid, and diagnosis
* 600 patients are labeled as "Healthy" and 400 as "Condition Present"

**Instructions**

1. Calculate basic statistics:
   * Compute the mean vector for each diagnosis group
   * Calculate the standard deviation for each parameter in both groups
   * Create a table comparing these statistics between "Healthy" and "Condition Present" groups
2. Analyze parameter relationships:
   * Calculate the correlation matrix for the five medical parameters
   * Identify which parameters have the strongest correlations
   * Create scatter plots of the two most correlated parameters
3. Apply multivariate normal modeling:
   * Fit a multivariate normal distribution to each patient group
   * Compare the mean vectors between groups
   * Calculate the log-likelihood of each model
4. Classification task:
   * For a new patient with measurements: Temperature: 99.0°F, Blood Pressure: 130 mmHg, Glucose: 108 mg/dL, Hemoglobin: 13.5 g/dL, Uric Acid: 6.0 mg/dL
   * Calculate the probability of this patient belonging to each group
   * Determine the most likely diagnosis and explain your reasoning
5. Summarize your findings:
   * Explain which medical parameters show the greatest difference between the two patient groups
   * Discuss how the multivariate approach captures relationships that might be missed by looking at individual parameters