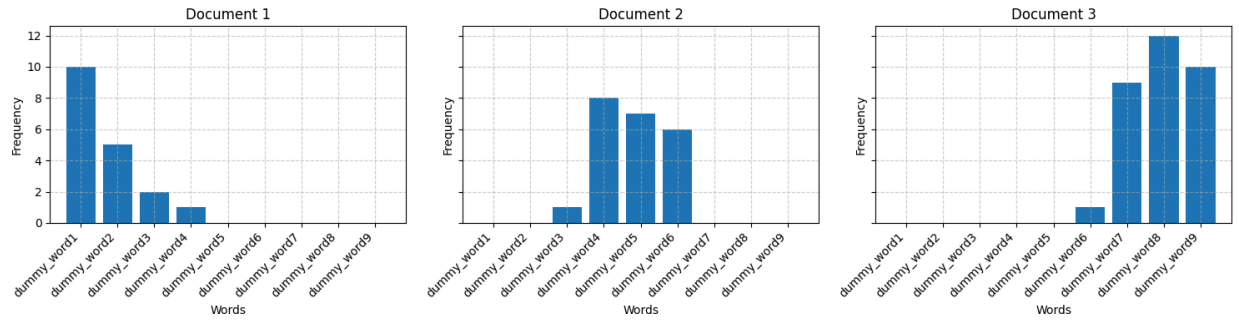
Due Date: 17.12.2025 23:59

CENG 463 – Introduction to Machine Learning HW2

Task 1 (20 pts)

Using the following 9-word vocabulary(*victory, team, fan, export, sector, product, committee, party, law*), create a *1×9 Bag of Words frequency vector (histogram)* for three documents (*sports*, *economy*, *politics*). Count the occurrences of each word and visualize the results using bar charts. Your histogram would be like below.



Task 2 (40 pts)

Data	Document	Words	Class
Training Data	d1	free, free, free, buy, discount, combo, pleasure,	S
	d2	free, free, free, discount, pleasure, smile, smile, smile	S
	d3	cat, mouse	N
	d4	cat, cat, dog, dog, dog, dog	N
	d5	mouse	N
Test Data	d6	dog, cat, mouse, cat	???
	d7	Free, free, smile	???

In this assignment, you will classify short text messages as *Spam* or *Normal* using basic machine learning methods: Chi-Square feature selection, TF-IDF representation, and KNN classification.

Step 1: Feature Selection (15 pts)

Use the *Chi-Square test* to find which words best separate Spam and Normal messages.
Show a short example of how you calculated χ^2 for one word (for example, "free").
List the top 2–3 most discriminative words you found.

Step 2: TF-IDF Representation (15 pts)

Using the selected words, calculate their *TF, IDF,* and *TF×IDF* values for each document. Build a small table showing TF-IDF values for all 10 messages and the 2–3 chosen words. This table will be your *numerical feature matrix* for classification.

Step 3: Classification with KNN (10 pts)

Use *K-Nearest Neighbors (k=3)* to classify the test messages. Compute the *TF-IDF vector* for each test message and measure its distance from all training examples (Euclidean or Cosine distance). Assign each test message to the most common class among its 3 nearest neighbors.

Task 3 (40 pts)

Categorize English news articles from three groups using BERTopic. You can select as many documents as they wish for the analysis.

Step 1: Dataset Creation (4 pts)

- Use `sklearn.datasets.fetch_20newsgroups`. Select only the 3 categories above.

Step 2: Model Creation & Training (16 pts)

- Train a BERTopic (https://bertopic.readthedocs.io/en/latest/) model.
- Report the number of topics.
- Show topic summary (get_topic_info()).

Step 3: Topic Analysis (8 pts)

- Show the top 5 words per topic.
- Assign meaningful names to topics (e.g., “Sports”, “Space”, “Politics”).
- Create a table showing which document belongs to which topic.

Step 4. Visualization (8 pts)

- Bar chart of top words per topic.
- Topic distance map showing topic similarities.

Step 5: New Document Test (4 pts)

- Write 3 new articles (one per category).
- Predict topics using the trained model.

!!!!!IMPORTANT

For each task, write a short discussion in the HW2_Report.pdf file explaining what you did, how you interpreted the results, and any important observations, insights, or limitations you identified during the analysis.

Assignment Rules:

1. In this homework, no cheating is allowed. If any cheating is detected, the homework will be graded as 0, and no further discussion will be entertained.

2. You are expected to submit your homework in groups. Therefore, it will be sufficient if only one member of the group submits the homework.

3. You must **fill in the provided Jupyter Notebook file** named:
   - CENG463_HW2_GROUPID.ipynb

   Follow the steps and instructions **provided in this file carefully.**

4. You must upload a **.zip file** to MS Teams. This file must include:
   - **CENG463_HW2_GROUPID.ipynb**

     Example: CENG463_HW2_G01.ipynb

   - **CENG463_HW2_GROUPID_Report.pdf**

     Example: CENG463_HW2_G01_Report.pdf

5. The .zip file must be named in the following format: **CENG463_HW2_GROUPID.zip**
   Example: CENG463_HW2_G01.zip
6. Please be aware that if you do not follow the assignment rules regarding export format and naming conventions, you will lose points.