Group 6 Date: December 14, 2025

# CENG 463 – Introduction to Machine Learning: HW2 Report

---

## Task 1: Bag of Words Analysis

**Methodology**

In this task, we implemented a Bag of Words (BoW) model to analyze the frequency distribution of a specific 9-word vocabulary across three domain-specific documents (Sports, Economy, and Politics). The text data was pre-processed by converting all characters to lowercase and removing punctuation to ensure accurate matching.

**Interpretation of Results**

The resulting histograms demonstrated a clear semantic separation between the documents.

- **Document 1 (Sports):** Showed high frequencies for "team," "victory," and "fan," with zero occurrences of political or economic terms.
- **Document 2 (Economy):** Was dominated by terms like "export," "sector," and "product."
- **Document 3 (Politics):** Exclusively contained the words "law," "party," and "committee."

**Limitations & Observations**

While the BoW model successfully categorized these simple documents, we identified significant limitations during the analysis:

1. **Loss of Context:** The model ignores word order and grammar. For example, "fan" and "fans" are treated as completely different tokens unless stemming/lemmatization is applied.
2. **Sparsity:** With a larger vocabulary, the feature vectors would become extremely sparse, leading to computational inefficiency.
3. **Semantic Gap:** The model cannot understand that "victory" and "win" are synonyms; it only counts exact keyword matches.

---

## Task 2: Text Classification (Spam vs. Normal)

**Methodology**

We built a text classification pipeline involving Chi-Square feature selection, TF-IDF representation, and K-Nearest Neighbors (KNN, $k=3$) classification on a small toy dataset.

**Observations & Insights**

**1. Chi-Square Calculation Discrepancy:**

We observed a numerical difference between our manual calculation ($\chi^2 \approx 5.0$) and the sklearn library output ($\chi^2 = 9.0$) for the word "free".

- **Reason:** Our manual calculation utilized a **binary** approach (presence/absence of the word in a document). In contrast, sklearn.feature_selection.chi2 operates on the **term frequency** matrix. Since the word "free" appeared multiple times within a single spam document, sklearn assigned it a higher statistical weight. Both methods, however, correctly identified "free" as the most discriminative feature for the Spam class.

**2. Data Leakage Defense (TF-IDF):**

In Step 2, we computed the TF-IDF matrix by combining both the Training and Test datasets.

- **Justification:** In a standard machine learning pipeline, this would constitute **data leakage**. However, given the extreme scarcity of the dataset (only 5 training documents), calculating IDF solely on the training set would result in unstable weights or zero-division errors for terms present in the test set but absent in the training set. Merging the corpus was a necessary step to ensure a valid numerical representation for this specific assignment.

**3. Classification Results:**

- **d6 ("dog, cat..."):** Classified as **Normal**. The vector distance was $0.0$ to the animal-related training documents (d3, d4, d5).
- **d7 ("Free, free, smile"):** Classified as **Spam**. Despite containing "smile", the high TF-IDF weight of the word "free" (due to repetition) pulled the document vector spatially closer to the Spam cluster in the Euclidean space.

---

## Task 3: Topic Modeling with BERTopic

**Methodology**

We utilized the BERTopic library to perform unsupervised clustering on a subset of the 20 Newsgroups dataset, specifically targeting the categories: *Sports, Politics,* and *Space*.

**Significant Observation: The "Random State" Issue**

During the model training phase, we encountered an important engineering challenge regarding reproducibility:

- **Initial Attempt:** We initially attempted to fix the random_state of the UMAP dimensionality reduction algorithm to 42 to ensure identical results across runs.
- **Failure Mode:** We observed that this specific seed forced the algorithm into a local optimum where semantically distinct topics (specifically *Space* and *Politics*) were merged into a single, incoherent cluster. The model failed to separate the data effectively.
- **Correction:** We reverted to BERTopic's default dynamic initialization for UMAP. This allowed the algorithm to properly find the manifold structure of the data.

**Final Results & Interpretation**

After the correction, the model successfully identified approximately **21 distinct topics** (excluding outliers).

1. **Granularity:** The model did not just find 3 broad categories but discovered specific sub-topics. For instance, the "Space" category was broken down into meaningful clusters like *Topic 2 (Space Station/Cost)*, *Topic 6 (Oort Cloud)*, and *Topic 5 (Hubble Telescope)*.
2. **Coherence:** The Top-5 words for each topic (e.g., *Topic 0: game, team, baseball*) showed high semantic coherence.
3. **Generalization:** In Step 5, when we fed the model new, synthetic documents (e.g., a sentence about a baseball pitcher), it correctly predicted **Topic 0 (Sports)**. This validates that the underlying SBERT embeddings effectively capture semantic meaning and generalize well beyond the training data.