# **HW Assignment 1 (A1): Bag of Words**

# CS6120: Natural Language Processing Spring 2024

Due Date: Tuesday, January 29, 2024

**(10 points)**

# An image depicting a computer screen displaying physician notes within an Electronic Health Record (EHR) system. The screen should show a detailed patient chart with various sections such as medical history, medications, lab results, and doctor's notes. The layout should be organized and professional, resembling a typical EHR interface used in hospitals. Emphasize the clarity and precision of the information, with tabs or sections neatly labeled and arranged for easy navigation. The screen should look realistic, giving the impression of an actual EHR system in use by healthcare professionals.The USMLE® Step 2 Clinical Skills Patient Note Corpus - ACL Anthology

# Scenario

Integrating clustering techniques with dimension reduction in unsupervised learning presents a fascinating study area. Dimension reduction, a process that streamlines complex, high-dimensional datasets into a more manageable form, is essential for efficient data analysis and visualization. Techniques like Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation Projection (UMAP) are instrumental in this context. Applying clustering methods such as k-means, c-means, hierarchical clustering, DBSCAN, HDBSCAN, and Expectation-Maximization (EM) to dimensionally reduced datasets offer a comprehensive understanding of how these algorithms can identify patterns and groupings effectively. This approach facilitates a practical application of these algorithms and deepens the knowledge of their collective impact in enhancing data analysis, particularly within unsupervised learning.

# Tasks

In this assignment, you are provided with 40,000 physician notes authored by test-takers of the USMLE. These notes, written for ten standardized patients, offer a unique dataset for analysis. The notes contain a natural ten clusters as the patients are the same for all note writers. The task is a good example of unsupervised learning where the ground truth can be used for post-hoc analysis. Your tasks are as follows:

1. **Data Preprocessing:** 
   1. Begin by preprocessing the physician's notes. This should include:
      1. Case Conversion
      2. Removing Punctuation and Special Characters
      3. Correcting Typos and Spelling. Think of quicker faster ways of doing this.
      4. Standardizing Formats for dates, numbers, and currencies etc.
      5. Handling Contractions: Expanding contractions like "can't" to "cannot".
      6. You may also consider optional steps like stemming and lemmatization.
      7. Apply a stop word list to filter out unnecessary words.
2. **Document-Term Matrix (DTM) Creation:** 
   1. Create a Document-Term Matrix with appropriate weighting, choice of n-grams, and other hyperparameters. This matrix will form the basis for your subsequent analyses.
   2. Describe the DTM from the dataframe (df) perspective, such as size, memory etc.
3. **Machine Learning Implementation:** 
   1. ML Pipeline: This should include normalization of your data to ensure uniformity and outlier analysis to identify and address any anomalies in your dataset.
   2. Dimension Reduction: Apply dimension reduction techniques. The aim here is to reduce the complexity of your data while retaining its essential characteristics. This step is crucial for effective clustering.
   3. Clustering: With your data preprocessed and dimensionally reduced, apply clustering techniques. Choose appropriate clustering algorithms to identify patterns and groupings in your dataset.

# **Expected Output**

Please submit a fully executed jupyter notebook identifying question number and steps. Make sure to add comments to your solution.